An Analysis of the Effect of Invariance on Generalization in Neural Networks

Clare Lyle [1] **Marta Kwiatkowksa** [1] **Yarin Gal** [1]

Abstract

Invariance is often cited as a desirable property of machine learning systems, claimed to improve model accuracy and reduce overfitting. Empirically, invariant models often generalize better than their non-invariant counterparts. But is it possible to show that invariant models provably do so? In this paper we explore the effect of invariance on model generalization. We find strong Bayesian and frequentist motivations for enforcing invariance which leverage recent results connecting PAC-Bayes generalization bounds and the marginal likelihood. We make use of these results to perform model selection on neural networks.

1. Introduction

Real-world data often exhibits invariance properties that naive neural network architectures fail to capture. For example, a convolutional neural network trained on upright handwritten digits will see a dramatic drop in accuracy on rotations of its training set, and most image recognition systems fall prey to adversarial examples, even when trained on noisy inputs (Szegedy et al., 2013; Carlini & Wagner, 2017). Invariance is thus cited as a desirable property of neural networks, and one that is notoriously difficult to implement architecturally.

The success of architectures like convolutional neural networks (LeCun et al., 1995), DeepSets (Zaheer et al., 2017), and group-equivariant convolutional layers (Cohen & Welling, 2016), is often attributed to their invariance properties. However, there has been relatively little work examining the effect of invariance on learning algorithms. Early work on symmetries in neural networks (Shawe-Taylor, 1993) suggests that invariance can improve generalization bounds, but as most of these bounds are vacuous to begin with these results do not yield practical guidance for model selection.

This then begs the question: *why should we prefer invariant models*? Intuitively, there are many possible reasons: perhaps invariant architectures simplify the input space in some sense, thus reducing sample complexity. Perhaps they allow us to simplify parameter space, making optimization more efficient and reducing the chance of overfitting. Perhaps the space of invariant functions is better-behaved in some sense than non-invariant ones, even holding the number of parameters and input space fixed.

In this paper we endeavour to develop a better theoretical understanding of invariance in learning algorithms. Our findings are summarized as follows.

1. We show that enforcing invariance to a set of transformations exhibited in the data increases the Bayesian model evidence, and reduces the model's loss function whenever the loss is convex with respect to the predictions.
2. We leverage existing work to show that incorporating invariance into model structure improves a PAC-Bayes generalization bound.
3. We verify our claims empirically, and show that an approximation of the marginal likelihood can be used to select for invariant neural networks.

2. Background

We first introduce notation that will recur throughout the paper.

Definition 1 (Reynolds operator)**.** *Let $\mathcal{F}$ be a class of functions with input space $\mathcal{X}$. Let $\mathcal{G}$ be a finite group acting on $\mathcal{X}$. Then the operator $R : \mathcal{F} \to \mathcal{F}$ will yield a function invariant to $\mathcal{G}$ as follows*

$$R(f)(x) = \left(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} f(gx)\right)$$

We will often use the notation $\bar{f} = R(f)$.

An analogous operator can be considered for distributions, where, given a probability distribution η on a set of functions $\mathcal{F}$, we define $R_{\mathcal{D}}(\eta)$ over the invariant functions

[1]University of Oxford. Correspondence to: Clare Lyle <clare.lyle@cs.ox.ac.uk>.

ICML 2019 Workshop on Understanding and Improving Generalization in Deep Learning, Long Beach, California, 2019.

$R(\mathcal{F})$ by

$$R_{\mathcal{D}}(\eta)(\bar{f}) := \sum_{f \in R^{-1}(\bar{f})} \eta(f)$$

and will use the notation $\bar{\eta} = R(\eta)$.

This operator corresponds to a practical method that we will refer to as *feature averaging*, whereby the outputs of some layer of a neural network are averaged over a set of transformations of the input. In practice this is generally done at the final prediction layer of the network, and its use goes back as far as 2012, when state-of-the-art results on ImageNet (Krizhevsky et al., 2017) averaged predictions over random crops of the input image at test time.

Feature averaging can be likened to data augmentation, though the two are not quite identical. In the linear setting, feature averaging can be viewed as a first order approximation of data augmentation (Dao et al., 2018). Leveraging invariances using this averaging approach has seen success in Gaussian Processes (GPs), where van der Wilk et al. (van der Wilk et al., 2018) find that by optimizing the marginal likelihood of the GP, they're able to recover nontrivial sets of transformations over which to average the model's kernel to improve its performance.

Beyond applying transformations to input and averaging the model's output, other approaches construct equivariant neural network layers so that intermediate representations are equivariant to a desired group. Recent work by Cohen et al. (Cohen & Welling, 2016; Cohen et al., 2018) extends the convolutional filter to more exotic groups than translations in $\mathbb{Z}^2$, allowing networks to exhibit equivariance to these groups. Other work proposes a more general framework for equivariance through parameter sharing (Ravanbakhsh et al., 2017). While this is a prolific field of study, it is unclear whether the networks proposed benefit from equivariance specifically, or from the specific filters used. These methods reduce the number of parameters necessary to obtain a given training set performance, and anecdotally reducing the number of parameters in a model tends to improve generalization error.

The connection between generalization and invariance is drawn by Shawe-Taylor (Shawe-Taylor, 1993) in multilayer perceptrons. These generalization bounds depend on the VC dimension of the neural network, and thus are generally vacuous on modern architectures. Nonetheless, invariance is broadly considered to be a desirable property by deep learning practitioners. More theoretical properties of invariant models are explored in (Bloem-Reddy & Whye Teh, 2019). Further work on the relationship between the marginal likelihood and PAC Bayes bounds (Germain et al., 2016) is suggestive of a connection between invariance and generalization, although it does not make this point explicitly. We will formalise these connections in our study of invariance.

3. Theoretical properties of invariant models

We first observe a few properties of the operator R.

Lemma 1. *Given a neural network architecture f, let f_θ, where $\theta \in \mathbb{R}^d$ denote the function defined by architecture f and parameters $\theta \in \mathbb{R}^d$. Then there exists a neural network f such that $\mathcal{F} := \{f_\theta : \theta \in \mathbb{R}^d\} \neq \bar{\mathcal{F}} := \{\bar{f}_\theta : \theta \in \mathbb{R}^d\}$. That is, the class of functions we can compute using f is distinct from the class of functions we can compute by averaging inputs over the action of a group.*

The proof of this statement and those that follow can be found in the supplementary material. This result is significant as previous comparisons between feature averaging and data augmentation in the kernel methods setting had suggested that the two approaches were closely linked, being equivalent to first order, while our observation shows the limitations of this observation in the deep learning setting. Indeed, $\mathcal{F}$ and $\bar{\mathcal{F}}$ are computing fundamentally different classes of functions, and so optimal parameters for one may be far removed from the optimal parameters for the other.

Although these two classes of functions are different, and based on empirical evidence presented in section 4 the feature averaging approach appears to perform better, we would like to demonstrate a more principled motivation for performing feature averaging. Do feature averaged models have nicer theoretical properties? The answer to this question is affirmative, as we can show that, provided the function we wish to compute satisfies the invariance given by $\mathcal{G}$, averaging out our predictions increases the Bayesian model evidence of our class of models.

3.1. The Bayesian perspective

Invariances can have powerful effects on Bayesian approaches to generalization bounds and model selection. We will use M to denote a model (e.g. a neural network architecture), and θ to denote a set of parameters. We first consider the regression setting, where our model implicitly defines a probability distribution over $\mathcal{Y}$ given by some output $M_\theta(x)$.

Theorem 1. *Let M be a model such that*

$$P(y|\theta, x, M) \propto \exp(-\ell(M_\theta(x), y))$$

where ℓ is a convex loss function. Let G acting on $\mathcal{X}$ be such that $y(gx) = y(x) \forall g \in G$. Let $\bar{M}_\theta(x) = \frac{1}{|G|} \sum_{g \in G} M_\theta(gx)$. Then

$$P(Y|X, M) \leq P(Y|X, \bar{M}).$$

In the classification setting, where $\mathcal{Y}$ is a discrete set of labels, we can show a similar result.

Theorem 2. *Let* $P(Y|\bar{M}_\theta, X) = \frac{1}{|G|}\sum_{g\in G} P(y|M_\theta, gx)$ *then*

$$P(Y|\bar{M}_\theta, X) \geq P(Y|M_\theta, X)$$

It's then trivial to show that the marginal likelihood of invariant models obtained by the Reynolds operator will always be lower bounded by the marginal likelihood of the non-invariant model from which they are derived. Because increasing the marginal likelihood of a model has been shown to improve a PAC-Bayes generalization bound whose loss function is the negative log likelihood (Germain et al., 2016), we obtain immediately from the previous result that invariant models should additionally improve this PAC-Bayes generalization bound. The results that follow explore whether this will still be true in more general settings.

Theorem 3 ((Catoni, 2007))**.** *Given a distribution* $\mathcal{D}$ *over* $\mathcal{X}\times\mathcal{Y}$ *a hypothesis set* $\mathcal{F}$*, a loss function* $\ell' : \mathcal{F}\times\mathcal{X}\times\mathcal{Y}\to[0,1]$*, a prior distribution* π *over* $\mathcal{F}$*, a real number* $\delta\in(0,1]$*, and a real number* $\beta > 0$*, with probability at least* $1-\delta$ *over the choice of* $(X,Y)\sim\mathcal{D}^n$*, letting*

$$u(\rho) = -\beta\mathbb{E}_{f\sim\rho}\hat{\mathcal{L}}^{\ell'}_{X,Y}(f) - \frac{1}{n}(KL(\rho||\pi) + \ln\frac{1}{\delta}),$$

we have

$$\forall\hat{\rho}\text{ on }\mathcal{F} : \mathbb{E}_{f\sim\hat{\rho}}\mathcal{L}^{\ell'}_D(f) \leq \frac{1}{1-e^{-\beta}}[1-e^{u(\hat{\rho})}]$$

Lemma 2. *Let* $\mathcal{F}$ *be a set of functions mapping spaces* $\mathcal{X}$ *to* $\mathcal{Y}$ *and let* G *be a group acting on* $\mathcal{X}$*. Let* ρ *and* π *be two distributions on* $\mathcal{F}$ *such that* $\rho\ll\pi$*. We then have*

$$KL(\rho||\pi) \geq KL(\bar{\rho}||\bar{\pi})$$

with equality when ρ *is such that* $\frac{\rho(f)}{\pi(f)}$ *constant over the orbits*

Theorem 4. *Let* u *be defined as in Theorem 3. Then we have that for any distribution* ρ*,*

$$u(\rho) \leq u(\bar{\rho})$$

And if π *is invariant and* ρ *assigns nontrivial density to only one element from the orbit of* G*, with the map* R *inducing equivalence classes of size* M*, then*

$$u(\rho) \leq u(\bar{\rho}) - \frac{1}{n}\log M.$$

And thus the PAC-Bayes bound of Catoni is reduced by enforcing the invariance.

3.2. The frequentist perspective

Observation 1. *If* ℓ *is a convex loss function and* $\bar{f}$ *corresponds to a symmetry exhibited in dataset* D*, then by Jensen's inequality*

$$\sum_{x\sim D}\ell(f(x),y) \geq \sum_{x\sim D}\ell(\bar{f}(x),y).$$

This need not hold if ℓ is not convex. For example, if the model averages its predicted logits before feeding this average through a softmax layer, we may not observe an improvement in the cross-entropy loss.

We further observe that enforcing invariance via feature averaging changes the training dynamics of stochastic gradient descent. Invariant models see less variance in their gradients, and so may exhibit different trajectories in weight space during optimization. While it is easy to show that the variance of the gradients differs, we do not prove that this difference leads to differences in performance. Empirical analysis of this behaviour follows in section 4.

4. Empirical investigation

4.1. Feature averaging vs data augmentation

While we have proven that invariance improves many measures of worst-case generalization properties, the types of bounds described in the previous sections are often vacuous. Indeed, given the overparametrization of most architectures used in practice, one would expect that the primary benefit of feature averaging over data augmentation is simply to reduce the variance of the model outputs over regions of the input space over which we know a priori the model should be invariant.

We first empirically validate that models which perform feature averaging do indeed see an improvement in performance over models which train on an augmented dataset. We evaluate models on the data set FashionMNIST augmented with the discrete rotation group of 4 elements (i.e. rotations of multiples of 90 degrees). We use a fixed convolutional neural network architecture consisting of three convolutional layers followed by two dense layers, from which we train a non-invariant model and a model which averages its predictions over the four rotations of the input image. Both models are trained on the augmented dataset. We treat the classification problem as a regression problem both to make the task more difficult for the networks and to take advantage of the convexity of the mean squared error loss.

We also note that feature averaging when employed only at test time cannot be used to make up for training on non-augmented data. For example, training a convolutional network on upright MNIST and then averaging its predictions

Log Hessian	FA-CNN	CNN
FashionMNIST	-20.6	-19.4
MNIST	-52.2	-34.9
(Lower better)		

Figure 1. Model complexity term $\log\det\nabla\nabla C(\theta)$ for CNN and invariant (FA-CNN) models.

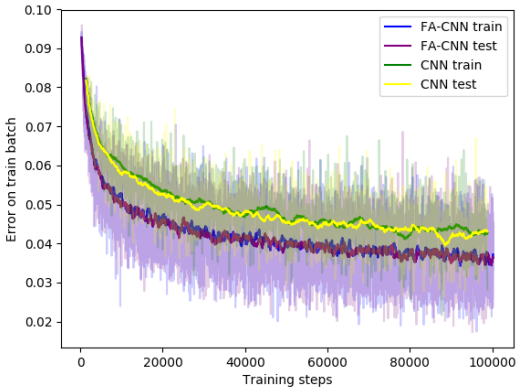

Figure 2. Test and train error for CNN and invariant (FA-CNN) models trained on FashionMNIST.

over rotations of the input does not yield an accurate classifier. However, when the training dataset is biased in some way over the orbits of the inputs (for example, including mostly upright images in a dataset when the desired model should be rotation invariant), feature averaging is observed to improve performance.

4.2. Bayesian model selection

In the previous section, we assumed a known invariance that could be incorporated into model structure. But often it is not known precisely what invariances a dataset may exhibit. We investigate here whether it is practically possible to perform Bayesian model selection on neural networks to discover what invariances are exhibited in a dataset. Recall that selecting the model with the maximum marginal likelihood is equivalent to selecting a model with the minimum PAC-Bayes generalization bound.

The marginal likelihood of a deep neural network is highly intractable, so we leverage a Laplace approximation under the assumption that the parameters obtained by SGD correspond to the maximum-likelihood solution, proposed by (Germain et al., 2016). This requires computing the Hessian of the neural network, which is in general also intractable, but can be approximated by methods such as those used in (Jastrzebski et al., 2018). In particular, we note that the approximation we will use only requires computing top eigenvalues of the Hessian provided the model has found a sufficiently flat minimum.

To evaluate the marginal likelihood, we perform a Laplace approximation as described by (), using a Gaussian prior of variance $\sigma^2 = 0.1$. The Laplace approximation of the marginal likelihood given that SGD has converged to parameters θ is thus:

$$P(y|x, M) \approx \exp\left\{ -\left(C(\theta) + \frac{1}{2}\ln(C''(\theta)/\sigma^2)\right)\right\}$$

where $C(\theta)$ is the loss function of the network and $C''(\omega)$ is its Hessian. From the observation $\ln(C''(\theta)/\sigma^2)$ can be expressed as

$$\ln(C''(\theta)/\sigma^2) = \sum_{\lambda_i} \ln(\lambda_i/\sigma^2)$$

we present a similar heuristic (Smith & Le, 2017) and consider only $\lambda_i > \sigma^2$. This heuristic can be seen as an approximation tool when the model assumes some noise in the output labels, as the corresponding noise term means that eigenvalues close to zero in noiseless Hessian will be close to this noise term in the corresponding noisy model, and hence will not contribute much to the log sum.

Using this approximation, we go back to the FashionMNIST experiments from the previous section and evaluate the hessian of each model, confirming that the invariant models have higher marginal likelihood than their non-invariant counterparts (since the loss function for the invariant model has already been shown to be lower than the non-invariant model, we need only consider the log hessian term in the marginal likelihood approximation). Though not conclusive – the difference in the log hessian values for the non-invariant and invariant models is small for the relatively flat minima arrived at after training on fashionMNIST for 100k steps – this provides promising evidence for the utility of the marginal likelihood in model selection in neural networks.

5. Conclusions and future work

This work has explored the effect of invariance on model performance and generalization. We've found principled

motivations for incorporating invariance into models in terms of both Bayesian model selection and model accuracy. Further, we have presented the marginal likelihood as a method of evaluating hypothesized invariances, and demonstrated its effectiveness in a simple model selection task. Future work may explore the generalization properties of different approaches to enforcing invariance and equivariance, particularly in earlier layers of the neural network.

References

Bloem-Reddy, B. and Whye Teh, Y. Probabilistic symmetry and invariant neural networks. *arXiv e-prints*, January 2019.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Catoni, O. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *Inst. of Mathematical Statistic*, 2007.

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

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. *International conference on learning representations*, 2018.

Dao, T., Gu, A., Ratner, A. J., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. *arXiv preprint arXiv:1803.06084*, 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

Jastrzebski, S., Kenton, Z., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. On the relation between the sharpest directions of dnn loss and the sgd step length. 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6): 84–90, 2017. doi: 10.1145/3065386. URL http://doi.acm.org/10.1145/3065386.

LeCun, Y., Bengio, Y., et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Equivariance through parameter-sharing. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2892–2901. JMLR. org, 2017.

Shawe-Taylor, J. Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, 4(5):816–826, 1993.

Smith, S. L. and Le, Q. V. A Bayesian Perspective on Generalization and Stochastic Gradient Descent. *ICLR 2018*, October 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

van der Wilk, M., Bauer, M., John, S., and Hensman, J. Learning Invariances using the Marginal Likelihood. *arXiv e-prints*, August 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in neural information processing systems*, pp. 3391–3401, 2017.

A. Proofs of results

Proof of lemma 1

Given a neural network f, let f_θ, where $\theta \in \mathbb{R}^d$ denote the function defined by architecture f and parameters θ. Let G be a finite group acting on the input space $\mathcal{X}$. Denote by $\bar{f}$ the function $\bar{f}(x) = \frac{1}{|G|}\sum_{g\in G} f(gx)$. We claim that there exists a neural network f such that $\mathcal{F} := \{f_\theta : \theta \in \mathbb{R}^d\} \neq \bar{\mathcal{F}} := \{\bar{f}_\theta : \theta \in \mathbb{R}^d\}$. That is, the class of functions we can compute using f is distinct from the class of functions we can compute by averaging inputs over the action of a group.

Proof. Let $\phi_a(x)$ be a network defined by a single ReLu unit with parameter $a = (a_1, a_2)$, i.e. $\phi_a(x) = \max\{0, a_1 x + a_2\}$. Then the class of functions $\mathcal{F}$ will be asymmetric except for the constant functions. Let $G = \mathbb{Z}_2$, with the action $gx = -x$ for the nontrivial element in G, and so $\bar{f}(x) = f(x) + f(-x)$ will be symmetric. Since no non-constant functions in $\mathcal{F}$ are symmetric, and no function in $\bar{\mathcal{F}}$ is asymmetric, the two classes are not equal, and nor is one a subset of the other. □

In the case where $\phi_a(x)$ is defined by two ReLU units, $\mathcal{F}$ would be capable of representing nontrivial symmetric functions. However, $\bar{\mathcal{F}}$ includes symmetric functions which have > 2 non-differentiable points, whereas $\mathcal{F}$ does not. So even when the original function class is capable of representing nontrivial functions respecting the desired symmetry, performing averaging over the group action can still yield a richer class of functions.

Proof of theorem 1

Suppose that we have a model such that

$$P(y|\theta, x, M) \propto \exp(-\ell(M_\theta(x), y))$$

, where ℓ is a convex loss function. Suppose that $\bar{M}_\theta = \sum_{g\in G} M_\theta(gx)/|G|$. Then $P(y|x, M) \leq P(y|x, \bar{M})$.

Proof. Simple application of Jensen's inequality.

$$\begin{aligned}
P(Y|X, M) &= \int_\theta P(Y|X, \theta, M)d\theta \\
&= \int_\theta \Pi_{x\in X} \exp(-\ell(M_\theta(x), y))d\theta \\
&= \int_\theta \exp(-\sum_{x\in X} \ell(M_\theta(x), y)d\theta \\
&= \int_\theta \exp(-\sum_{x\in X/G}\sum_g \ell(M_\theta(gx), y))d\theta \\
&\leq \int_\theta \exp(-\sum_{x\in X/G}\sum_g \ell(\bar{M}_\theta(gx), y))d\theta \\
&= \int_\theta \exp(-\sum_{x\in X} \ell(\bar{M}_\theta(x), y))d\theta \\
&= \int_\theta \Pi_{x\in X} \exp(-\ell(\bar{M}_\theta(x), y))d\theta \\
&= P(Y|X, \bar{M})
\end{aligned}$$

□

Proof of theorem 2

Let $P(y|\bar{M}_\theta, x) = \sum_{g\in G} P(y|M_\theta, gx)/|G|$ then

$$P(Y|\bar{M}_\theta, X) \geq P(Y|M_\theta, X)$$

Proof.

$$\begin{aligned}
\int \Pi p(y|M_\theta(x))d\theta &= \int \Pi_{X/G}\Pi_{g\in G} p(y|M_\theta(gx))d\theta \\
&= \int \Pi_{X/G} \sqrt[|G|]{\Pi_{g\in G} p(Y|M, \theta, X))^{|G|}}d\theta \\
&\leq \int \Pi_{X/G}(\frac{1}{G}\sum p(y|M, \theta, x))^{|G|}d\theta \\
&= \int \Pi_{X/G} p(y|\bar{M}, \theta, x)^{|G|}d\theta \\
&= \int \Pi_X p(y|\bar{M}, \theta, x)d\theta
\end{aligned}$$

□

Proof of lemma 2

Let $\mathcal{F}$ be a set of functions mapping spaces $\mathcal{X}$ to $\mathcal{Y}$ and let G be a group acting on $\mathcal{X}$. Let ρ and π be two distributions on $\mathcal{F}$ such that $\rho \ll \pi$. We then have

$$KL(\rho||\pi) \geq KL(\bar{\rho}||\bar{\pi}).$$

If ρ assigns nontrivial weight to a single element in each class and π is invariant over the equivalence classes given by R and these equivalence classes are of size M, then the difference in the KL divergences can be computed exactly:

$$KL(\rho||\pi) = KL(\bar{\rho}||\bar{\pi}) + \log(M)$$

Proof.

$$\begin{aligned}
KL(\rho||\pi) &= \int_f \rho(f)\log(\frac{\rho(f)}{\pi(f)}) \\
&= \int_{\bar{f}\in\mathcal{F}/R} \sum_{f\in R^{-1}(\bar{f})} \rho(f)\log(\rho(f)/\pi(f)) \\
&= \int \bar{\rho}(\bar{f}) \sum \rho(f)/\bar{\rho}(\bar{f})\log(\rho(f)/\pi(f))
\end{aligned}$$

We can use the calculus of variations to show that the optimal ρ' for the above equation satisfying $\sum_{f\in R^{-1}(\bar{f})} \rho'(f) = \bar{\rho}(\bar{f})$ is given by

$$\rho'(f) := \frac{\bar{\rho}(\bar{f})}{\bar{\pi}(\bar{f})}\pi(f)$$

Further, $KL(\rho||\pi) \geq KL(\rho'||\pi)$ for all $\rho \neq \rho'$, so it remains to show that $KL(\rho'||\pi) \geq KL(\bar{\rho}||\bar{\pi})$, which is easy as:

$$\begin{aligned}
KL(\rho'||\pi) &= \int_f \rho'(f)\log(\frac{\rho'(f)}{\pi(f)})df \\
&= \int_f \rho'(f)\log(\frac{\bar{\rho}(\bar{f})}{\bar{\pi}(\bar{f})})df \\
&= \int_{\bar{f}} \log(\frac{\bar{\rho}(\bar{f})}{\bar{\pi}(\bar{f})} \sum_{f\in R^{-1}(\bar{f})} \rho'(f)df \\
&= KL(\bar{\rho}||\bar{\pi}))
\end{aligned}$$

Finally in the setting where π is invariant and ρ assigns nontrivial weight to a single element, we have:

$$\begin{aligned}
KL(\rho||\pi) &= \int_{\bar{f}} \sum_{R^{-1}(\bar{f})} \rho(f)\log\frac{\rho(f)}{\pi(f)}d\bar{f} \\
&= \int \rho(f)\log\frac{\bar{\rho}(\bar{f})}{\bar{\pi}(\bar{f})/M}df \\
&= \int \bar{\rho}(\bar{f})(\log\frac{\bar{\rho}}{\bar{\pi}} + \log(M))d\bar{f} \\
&= KL(\bar{\rho}||\bar{\pi}) + \log(M)
\end{aligned}$$

□

A note on the assumptions of theorem 4

While the assumptions of theorem 4 may appear arbitrary, they are motivated by practical approaches to generalization bounds in neural networks. Many PAC-Bayes bounds construct a Gibbs classifier from a deterministically trained neural network by setting the weights to follow a normal distribution with mean μ arrived at during training and some variance σ^2. Meanwhile, the prior on the weights is typically set to be iid gaussian centered at zero. Because one can obtain equivalence classes under certain group actions on the inputs by permuting the weights of the neural network in a particular way, we get that the prior weight (which is permutation invariant) assigned to any two networks which are in the same equivalence class under the operator R will be equal. Meanwhile, the majority of the mass of the posterior will be centered about the values obtained during training, which are unlikely to exhibit permutation invariance, and so only one element of the equivalence class is likely to be assigned nontrivial density under the posterior ρ.

Proof of theorem 4

By the observation of the previous theorem, the result follows immediately as

$$e^{u(\rho)} = e^{u(\bar{\rho}) - \frac{1}{n}\log(M)}$$

and since e^x is monotone, we have that

$$\frac{1}{1-e^{-\beta}}(1 - e^{u(\rho)}) \geq \frac{1}{1-e^{-\beta}}(1 - e^{u(\bar{\rho})})$$