Group averaging and statistical marginalization: Is there a connection?

Hogg & Clark for Villar

August 2025

In machine-learning contexts, there is a known trick—known but not (to my knowledge) widely used—that a non-equivariant machine-learning method can be made equivariant (and made more accurate) by *group averaging*. That is, if a trained function $f(x\,;W)$, with weights W found by training on training data, was trained to predict labels y, and if the true relationship between x and y is equivariant to a group G, then the group-averaged function $\bar{f}_G(x\,;W)$ will (in many cases) outperform $f(x\,;W)$ in predictive accuracy. The group averaging looks like

$$\bar{f}_G(x\,;W) = \frac{1}{|G|}\,\sum_{g\in G} g_y^{-1}\cdot f(g_x\cdot x\,;W)\;, \tag{1}$$

where the outer $g_y^{-1}\cdot$ operator is the action of the (inverse) group operator g^{-1} acting on the output y-space, and the inner $g_x\cdot$ is the action of the group operator g acting on the input space. In the case of group-invariant functions (which are of most importance here), the outer $g_y^{-1}\cdot$ is just the identity.

The conditions under which the group-averaged function $\bar{f}_G(x\,;W)$ provably makes better predictions for the labels y are...SOMETHING. But in practice it often seems to help.

In inference contexts, *marginalization* is used to remove nuisance parameters that are not of interest to the final inferences. If a likelihood function $p(y\,|\,\theta,\alpha)$ depends on parameters of interest θ and nuisance parameters α, and if the investigator has, as part of their model (or, if Bayesian, part of their beliefs about the world) a prior pdf $p(\alpha)$ for the nuisance parameters α, then the likelihood can be marginalized such that it only depends on θ:

$$p(y\,|\,\theta) = \int p(y\,|\,\theta,\alpha)\,p(\alpha)\,\mathrm{d}\alpha\;, \tag{2}$$

where the integral (implicitly) goes over the whole domain of the nuisances α, and (implicitly) all pdfs are properly normalized. This marginalization can be done by frequentists if the distribution over α is part of their model. This marginalization *must* be done by Bayesians if they want to make their best possible probabilistic predictions for θ. In the end, Bayesian predictions are made by multiplying the likelihood by priors and renormalizing:

$$p(\theta\,|\,y) = \frac{1}{Z}\,p(\theta)\int p(y\,|\,\theta,\alpha)\,p(\alpha)\,\mathrm{d}\alpha\;, \tag{3}$$

where $p(\theta\,|\,y)$ is a pdf representing posterior beliefs about the parameters of interest θ (given the data y), $p(\theta)$ is a pdf representing prior beliefs, and Z is a normaliza-

tion constant to make the LHS a pdf. There are Cox theorems (HOGG CITE) that prove that Bayesian reasoners win bets against all other reasoners, in contexts in which the Bayesians bet against the other reasoners, so there are conditions in which this marginalization will provably lead to better predictions than any other operation. Also in practice this kind of marginalization is useful and works well.

Now let's get more specific. The world generates a true label $\tilde{y}$ from features x in such a way that it is *invariant* to some latent angle or set of latent angles α. This ...

1 Can we run this in practice?

Well... Running the marginalization is quite expensive. This is because we usually can't integrate it directly, so we usually MCMC it:

$$\int p(y|\theta,\alpha)p(\alpha)d\alpha = \mathbb{E}_{\alpha\sim p(\alpha)}[p(y|\theta,\alpha)] \tag{4}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} p(y|\theta,\alpha_i) \tag{5}$$

So if we were doing this inference on an incredibly large dataset (for example, all of the Gaia data), you're out of luck (in that particular example, you'll burn 20% of Flatiron's yearly compute budget). So what if there was some measure $\nu(\alpha)$ (hopefully discrete / a smaller support) which approximated $p(\alpha)$ up to the k'th moment? This is the study of **k-designs**.

1.1 k-Designs

Consider a function f_k which is a polynomial to the power k. We say probability distribution ν forms a k-design with respect to the target distribution μ if

$$\mathbb{E}_{x\sim\nu}[f_k(x)] = \mathbb{E}_{x\sim\mu}[f_k(x)] \tag{6}$$

Example (Spherical Designs) [1]: Let $p_k : \mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$ be a polynomial in d variables, with all terms homogenous in degree at most k. A set $X = \{x : x \in \mathcal{S}(\mathbb{R}^d)\}$ is a spherical k-design if

$$\frac{1}{|X|}\sum_{x\in X} p_k(x) = \int_{\mathcal{S}(\mathbb{R}^d)} p_k(u)d\mu(u) \tag{7}$$

holds for all possible p_t, where $d\mu$ is the uniform, normalized sphereical measure. A spherical k-design is also a k'-design for all $k' < k$.

1.2 Finding k-Designs

Dries: you can find them via optimal control (it's not easy but that is an algorithmic way of finding them).

References

[1] Charles J. Colbourn and Jeffrey H. Dinitz, editors. *Handbook of Combinatorial Designs*. Chapman and Hall/CRC, 2 edition, 2006.