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 ...

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).

After looking at the equation above for sometime, you might ask, "What if I knew the 'right' set of $\{\alpha_i\}_i$, that me closest to the 'best' answer?" You can think of the discrete set $\{\alpha_i\}_i$ as a measure corresponding to a discrete distribution. And the measurement of best can be what recovers the k'th moments of a sufficient statistic. After some googling, you realize this is the study of...

k-Designs Let's start with a heuristic definition, and then talk about real examples. 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

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

So a couple comments:

- A natural interpretation is the k-design preserves the k'th moment.
- Say μ is a continuous distribution, ν need not be continuous, it can be discrete! This means instead of sampling, we can evaluate the function on a finite set of points– very computationally quick!
- ν need not be unique.

You can see finding k-designs saves us time & has an implicit bound on how we diverge in our sufficient statistic. In this statistics context, it seems that we'll only need a 2-design.

Example (Spherical Designs [1, 2]): 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 spherical measure. A spherical k-design is also a k'-design for all $k' < k$. These designs are very pretty

because the elements of their event space are vertices the certain platonic solids (for example: the regular $(k+1)$-gon in $\mathcal{S}^2$ is a spherical k-design).

Example (Toric Design [3]): A toric k-design is a measure space $(X \subset T^n, \Sigma, \nu)$ such that

$$\int_X \exp\left(i\sum_{j=1}^n \alpha_j\phi_j\right) d\nu(\phi) = \int_{T^n} \exp\left(i\sum_{j=1}^n \alpha_j\phi_j\right) d\mu_n(\phi) \tag{8}$$

for all $\alpha \in \mathbb{Z}^n$ satisfying $\sum_{j=1}^n |\alpha_j| \leq t$.

Finding k-Designs

- Dries: you can find them via optimal control (it's not easy but that is an algorithmic way of finding them).
- Clark: This reminds me of rejection sampling (like construct a set $A = \{\alpha_i\}_i$, and ask which element $\alpha \in A$ should I s.t. the error of your sufficient statistics changes the least, do this over a training dataset and test set.).

1 Problem Statement

1.1 Radial Velocity Data in Binary System

Consider a binary system. We can make N_{obs} observations of the radial velocity $\mathcal{D} = \{v_t\}_{t=1}^{N_{obs}}$ (this is the planet's velocity going towards/away from us). The theoretical velocity $v(t;\theta)$ is non-linear in parameters $\theta = \{P, e, \omega, \phi_0, s\}$ (corresponding to the period of the orbit, the eccentricity, pericenter phase & argument, and jitter of the uncertainties).

If we assume the uncertainties in measurement are distributed according to a Gaussian, we get the likelihood:

$$\log p(\mathcal{D}|\theta) = -\frac{1}{2}\sum_{i=1}^{N_{obs}} \frac{(v_t - v(t_i;\theta))^2}{\sigma_i^2} + \text{constant} \tag{9}$$

In this problem we are interested in the Bayesian question, what is posterior $p(\theta|\mathcal{D})$? The state of the art method for sampling this posterior relies on rejection sampling [4]. This method cost exponentially more compute in the number of parameters $|\theta|$, so if we were to get rid of any parameters, that'd be super useful!

Here's where the k-design / group averaging comes in... The ω, ϕ_0 are of interest to us because (1) it's standard [4] to give them the prior Uniform$(0, 2\pi)$ and (2) they live on $\mathbb{T}^1 \times \mathbb{T}^1$ (2-torus). Meaning, if you marginalize these parameters out, you're actually doing group averaging, meaning we should apply our thinking about k-designs!

Here is the new proposed procedure:

1. Construct a marginalized likelihood by integrating out ω, ϕ_0

$$p(\mathcal{D}|P,e,s) = \int p(\mathcal{D}|P,e,\omega,\phi_0,s)p(\omega)p(\phi_0)\ d\omega\ d\phi_0 \tag{10}$$

$$= \int_{\mathbb{T}^1\times\mathbb{T}^1} p(\mathcal{D}|P,e,\omega,\phi_0,s)d\mu_H(\omega,\phi_0) \qquad \text{Equivalent to group average} \tag{11}$$

$$= \mathbb{E}_{(\omega,\phi_0)\sim\mu_H}[p(\mathcal{D}|P,e,\omega,\phi_0,s)] \tag{12}$$

$$= \mathbb{E}_{(\omega,\phi_0)\sim\nu}[p(\mathcal{D}|P,e,\omega,\phi_0,s)] \qquad \nu \text{ is the "appropriate" } k \text{ design} \tag{13}$$

$$= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} p(\mathcal{D}|P,e,\omega,\phi_0,s) \tag{14}$$

2. Sample the resulting posterior

$$p(P,e,s|\mathcal{D}) \propto p(\mathcal{D}|P,e,s)p(P)p(e)p(s) \tag{15}$$

   using the rejection sampling procedure proposed in [4].

The research questions are:

1. Traditional toric designs talk about $\mathbb{E}_\phi[e^{i\sum_j \alpha_j\phi_j}]$. In our problem, our minimal problem is a Gaussian $\mathbb{E}_\phi[e^{-(v_t-v(t;\phi))^2}]$ and the mean is nonlinear in ϕ. So we can't rely on results from traditional toric deigns.... So do k-designs even exists for our setup problem?

2. After proving existence (or if we even want to do that), we need to find measures which work on different examples of radial velocity data. So how can we find ν? (maybe optimal control theory, or something else)

References

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

[2] Philippe Delsarte, Jean-Marie Goethals, and Johan J. Seidel. Spherical codes and designs. *Geometriae Dedicata*, 6:363–388, 1977.

[3] Joseph T. Iosue, T. C. Mooney, Adam Ehrenberg, and Alexey V. Gorshkov. Projective toric designs, quantum state designs, and mutually unbiased bases. *Quantum*, 8:1546, December 2024.

[4] Adrian M. Price-Whelan, David W. Hogg, Daniel Foreman-Mackey, and Hans-Walter Rix. The joker: A custom monte carlo sampler for binary-star and exoplanet radial velocity data. *The Astrophysical Journal*, 837(1):20, February 2017.