Lecture notes for STAT 547S: Topics in Symmetry in Statistics and Machine Learning (draft; in progress)

Ben Bloem-Reddy

February 16, 2023

Contents

List of Exercises

Notation and background assumptions

First, a few background assumptions. I assume the existence of a probability space $(\Omega, \mathcal{H}, P)$ that is rich enough to support all of the random variables, etc., that we introduce. I also assume that we're always working on standard Borel spaces, so that regular conditional probability distributions exist and we can represent conditional distributions by Markov probability kernels. If we need to relax this latter assumption, I will say so explicitly. Since we'll be working on standard Borel spaces, if I don't specify a σ-algebra for a set then it is safe to assume that it's the Borel σ-algebra, denoted generically for a set S by $\mathcal{B}(S)$.

A few words on notation. I will use P generically to denote a probability measure or distribution; what it is the measure or distribution of should be clear from its argument. E will denote expectation, with subscripts indicating what the expectation is with respect to if necessary. For example, for a random variable X that takes values in $\mathbb{R}$, we might write $P(X \in A)$, $A \in \mathcal{B}(\mathbb{R}))$, for the probability that X is in the set A. $E[h(X)] = \int_{\mathbb{R}} h(x)P(dx)$ is the expectation of $h(X)$. If a probability measure has a density (assumed to be with respect to Lebesgue measure if we're working on a subset of $\mathbb{R}^n$) then the subscript will indicate what the density is. For example, f_X is the density of the distribution of X, so that $E[h(X)] = \int_{\mathbb{R}} h(x) f_X(x)\,\mathrm{d}x$.

A family of probability measures $\mathcal{P} = \{P_\theta \colon \theta \in \Omega\}$ is called a **parametric family** if $\Omega \subseteq \mathbb{R}^m$, for some finite m.

Conditioning may be represented in a few different ways. For example, the conditional probability of X given Θ may be $P(X \in A \mid \Theta)$ or represented by a Markov probability kernel as $K_{X|\Theta}(\Theta, A)$ (or, for integration, $K_{X|\Theta}(\Theta, dx)$). The latter will allow us to express things like conditioning on specific values of $\Theta = \theta$ as $K_{X|\Theta}(\theta, A)$, for which we may also write $P(X \in A \mid \theta)$. Conditional expectations will be written as, e.g., $E_{X|\Theta}[h(X)]$ for the conditional expectation of $h(X)$ given Θ.

1 Equivariant statistical analysis for a location parameter

Supplemental reading: Schervish [Sch95, Ch. 6.1.1], Robert [Rob07, Ch. 9.1, 9.2]

We'll start with an extended example which, although simple, contains most of the important aspects of more general problems. We'll use it to build some intuition before abstracting. I'll generally follow Schervish [Sch95, Ch. 6.1] in this section.

1.1 Location families

Let $X := (X_1, X_2, \ldots, X_n)$ be a vector of exchangeable random variables, each $X_i \in \mathbb{R}$ conditionally independent and identically distributed (IID) given a random parameter $\Theta \in \mathbb{R}$. In our extended example, Θ will be a **location parameter**.

Definition 1.1. Let X be a random vector in $\mathbb{R}^n$ and Θ a random scalar. Let $\mathbf{1}$ denote the appropriate-length vector of all 1's. If the conditional distribution of $X - \Theta\mathbf{1}$ given $\Theta = \theta$ is the same for all θ, then Θ is a **location parameter** for (the distribution of) X.

Theorem 1.2. *Let Θ be a location parameter for X, and denote the conditional distribution of X given Θ by $K_{X|\Theta}(\theta, dx)$. If $K_{X|\Theta}(\theta, dx)$ has a density with respect to Lebesgue measure for each $\theta \in \mathbb{R}$, denoted $k_{X|\Theta}(x \mid \theta)$, then $k_{X|\Theta}(x \mid \theta) = q(x - \theta\mathbf{1})$ for some density function q.*

See Schervish [Sch95, Thm. 6.2] for the proof.

Example 1.1 Normal with unknown mean as a location family. Suppose that $X_i \mid \Theta = \theta \sim \mathcal{N}(\theta, 1)$. Then

$$k_{X|\Theta}(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x - \theta\mathbf{1})^\intercal (x - \theta\mathbf{1})} .$$

A family of conditional densities $\{k_{X|\Theta}(x \mid \theta) \colon \theta \in \mathbb{R}\}$ is a **location family** if it has the form stated in Theorem 1.2. A frequentist approach to this problem [see, for example, LC98, Ch. 3] starts with a location family. We'll see that when we adopt certain symmetry-based principles, there ends up not being a meaningful difference between treating Θ as random (Bayesian) or deterministic (frequentist) in this problem.

Takeaway 1.1. The structure that drives all of our subsequent analysis is already present. Consider two vectors of data: x and $x' = x + c\mathbf{1}$. Then $k_{X|\Theta}(x \mid \theta) = k_{X|\Theta}(x' \mid \theta + c)$. That is, if we shift our data by a fixed amount c, then the conditional density (or likelihood) remains invariant if we also shift the parameter by the same amount. In some sense, there is redundant structure in the problem, and we'd like to take advantage of that. The redundant structure is encoded by the correspondence that occurs when we simultaneously shift data and parameter. The correspondence of change between two (or more) things is known broadly as **equivariance**.

Plot density functions.

Pushforward measures and kernels

To state this precisely, I'll introduce some notation that will be handy for generalizing things later on. For a measurable function $h \colon \mathbb{R}^n \to \mathbb{R}^n$ and a measure P on $\mathbb{R}^n$, let $h_{\#}P$ be the pushforward,

or image measure, of P under h. Recall that the pushforward measure is defined by

$$(h_{\#}P)(A) = P(h^{-1}A) \;, \quad A \in \mathcal{B}(\mathbb{R}^n) \;, \tag{1.1}$$

and that integration with a pushforward can be carried out via the identity,

$$\int_{\mathbb{R}^n} f(x)(h_{\#}P)(dx) = \int_{\mathbb{R}^n} f(h(x))P(dx) \;, \tag{1.2}$$

for any integrable function $f\colon \mathbb{R}^n \to \mathbb{R}$. We use similar notation for the pushforward of a Markov kernel, as

$$(h_{\#}K_{X|\Theta})(\theta, A) = K_{X|\Theta}(\theta, h^{-1}(A)) \;, \quad \int_{\mathbb{R}^n} f(x)(h_{\#}K_{X|\Theta})(\theta, dx) = \int_{\mathbb{R}^n} f(h(x))K_{X|\Theta}(\theta, dx) \;. \tag{1.3}$$

Suitably modified versions of these definitions and identities hold for general measurable spaces; in particular, h may have different range and domain. I will drop the parentheses unless they are needed for readability, i.e., $(h_{\#}P)$ will be denoted by $h_{\#}P$.

Let $g_c\colon \mathbb{R}^n \to \mathbb{R}^n$ be a family of (measurable) functions indexed by $c \in \mathbb{R}$ such that $g_c(x) = x + c\mathbf{1}$. Note that $g_c^{-1} = g_{-c}$. Abusing notation, let g_c also denote a function $\mathbb{R} \to \mathbb{R}$ such that $g_c(\theta) = \theta + c$.

Proposition 1.3. *Let Θ be a location parameter for X. Then for each $c \in \mathbb{R}$,*

$$K_{X|\Theta}(g_c(\theta), \bullet) = g_{c\#}K_{X|\Theta}(\theta, \bullet) \;, \quad \theta\text{-a.e.} \tag{1.4}$$

If $K_{X|\Theta}$ has density $k_{X|\Theta}(x \mid \theta)$, θ-a.e., then

$$k_{X|\Theta}(x \mid \theta) = k_{X|\Theta}(g_c(x) \mid g_c(\theta)) \;, \quad (x, \theta)\text{-a.e.} \tag{1.5}$$

An "in words" version of this can be found in Schervish [Sch95, Prop. 6.3]: "The conditional distribution of $X + c\mathbf{1}$ given $\Theta = \theta$ is the same as the conditional distribution of X given $\Theta = \theta + c$." Note that (1.5) implies the sometimes more useful identity

$$k_{X|\Theta}(g_c(x) \mid \theta) = k_{X|\Theta}(x \mid g_c^{-1}(\theta)) = k_{X|\Theta}(x \mid g_{-c}(\theta)) \;. \tag{1.6}$$

Exercise 1 (Shift-equivariance in location families —— **)**:**

Prove Proposition 1.3.

Hint: For (1.4), Definition 1.1 implies that for any positive measurable function $h\colon \mathbb{R}^n \to \mathbb{R}$ and each $b, \theta \in \mathbb{R}$,

$$\int K_{X|\Theta}(\theta, dx)h(x - \theta\mathbf{1}) = \int K_{X|\Theta}(\theta + b, dx)h(x - (\theta + b)\mathbf{1}) \;.$$

Example 1.2 Continuation of Example 1.1. Let's verify (1.5) for the normal density from Example 1.1. In that case,

$$\begin{aligned} k_{X|\Theta}(g_c(x) \mid g_c(\theta)) &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x+c\mathbf{1}-(\theta+c)\mathbf{1})^\intercal (x+c\mathbf{1}-(\theta+c)\mathbf{1})} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\theta\mathbf{1})^\intercal (x-\theta\mathbf{1})} = k_{X|\Theta}(x \mid \theta) \ . \end{aligned}$$

Exercise 2 (Alternate characterization of equivariance. —— *)**:**

Verify (1.6).

1.2 Equivariant estimation in a location family

Consider the problem of estimating a location parameter Θ from a sample $X \sim P(\bullet \mid \Theta)$. We'll call our estimator $\rho\colon \mathbb{R}^n \to \mathbb{R}$. A loss function $L\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called **location invariant** if $L(\theta, c) = \ell(\theta - c)$ for all $\theta, c \in \mathbb{R}$, where $\ell\colon \mathbb{R} \to \mathbb{R}$ is some function. We can see that the invariance here is when we apply the same shift to each argument of L:

$$L(\theta + b, c + b) = \ell(\theta + b - (c + b)) = \ell(\theta - c) = L(\theta, c) \ , \quad b, c, \theta \in \mathbb{R} \ . \tag{1.7}$$

Since it is a loss function, we will usually require $\ell(t)$ to be increasing in $|t|$.

Definition 1.4. An estimator ρ is **location equivariant** if it commutes with shifts of the data. That is, if

$$\rho(x + b\mathbf{1}) = \rho(x) + b \ , \quad b \in \mathbb{R} \ , \ x \in \mathbb{R}^n \ . \tag{1.8}$$

Some examples are x_i and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Location equivariant estimators have some special properties.

Proposition 1.5. *If ρ is location equivariant and Θ is a location parameter, then the conditional distribution of $\rho(X) - \Theta$ given $\Theta = \theta$ is the same for (almost) all $\theta \in \mathbb{R}$.*

Exercise 3 (Conditional invariance in location equivariant estimators —— **)**:**

Prove Proposition 1.5.

Hint: Express the conditional pushforward of $K_{X|\Theta}$ under $\rho(X) - \Theta$ and use Exercise 1.

Exercise 4 (Conditional invariance in location equivariant estimators for the normal mean —— *)**:**

Verify Proposition 1.5 for the normal mean problem from Example 1.1 for the estimator $\rho(x) = \bar{x}$.

The **risk function** of an estimator ρ is the expected conditional loss, given $\Theta = \theta$,

$$R(\theta, \rho) := \int_{\mathbb{R}^n} L(\theta, \rho(x)) K_{X|\Theta}(\theta, dx) \ . \tag{1.9}$$

Corollary 1.6. *The risk of a location equivariant estimator ρ is constant if the loss function is location invariant:*

$$R(\theta, \rho) = R(\theta + b, \rho) \ , \quad b \in \mathbb{R}, \ \theta \in \mathbb{R} \ . \tag{1.10}$$

Proof. Assume ρ is location equivariant. Fix arbitrary $\theta, b \in \mathbb{R}$. Then

$$\begin{aligned}
R(\theta + b, \rho) &= \int L(\theta + b, \rho(x)) K_{X|\Theta}(\theta + b, dx) && \text{(definition of conditional risk)} \\
&= \int L(\theta + b, \rho(x)) g_{b\#} K_{X|\Theta}(\theta, dx) && \text{(Proposition 1.3)} \\
&= \int L(\theta + b, \rho(x + b)) K_{X|\Theta}(\theta, dx) && \text{(definition of pushforward)} \\
&= \int L(\theta + b, \rho(x) + b) K_{X|\Theta}(\theta, dx) && \text{(equivariance of } \rho\text{)} \\
&= \int L(\theta, \rho(x)) K_{X|\Theta}(\theta, dx) && \text{(invariance of } L\text{)} \\
&= R(\theta, \rho) \ .
\end{aligned}$$

□

Takeaway 1.2. If we use an invariant loss, an equivariant estimator has the same risk for all values of the location parameter. This is an important property that will (partially) generalize to many other problems. Moreover, if we restrict our estimator to be equivariant, then we might answer a number of generally difficult questions through relatively simple means. For example, "What is the minimum-risk equivariant estimator?" or "What is the estimator that has the best worst-case risk across all values of θ?" We only need to observe data from a single parameter to make strong statements about the entire parametric family.

Some properties of location equivariant and invariant functions

Without further motivation (for now), we'll require ρ to be location equivariant, and the loss to be location invariant (e.g., $L(\theta, c) = (\theta - c)^2$). To make further, specific progress on the problem, we need to consider the class of equivariant estimators; we need to understand that class in a way that allows us to construct specific estimators and that—hopefully—allows us to optimize over the class with respect to the risk function.

A special role is played here by the statistic $Y := (X_1 - X_n, X_2 - X_n, \ldots, X_{n-1} - X_n)$. Note that any of the observations X_i could play the role of X_n here, since they're exchangeable. We'll see that the choice doesn't affect our choice of equivariant estimator in the end.

Lemma 1.7. *A function $u \colon \mathbb{R}^n \to \mathcal{U}$ is location invariant if and only if it is a function of x through $y = (x_1 - x_n, \ldots, x_{n-1} - x_n)$. That is, $u(x) = v(y(x))$, for some function $v \colon \mathbb{R}^{n-1} \to \mathcal{U}$.*

Proof. Clearly, if u depends on x only through y, then $u(x + b\mathbf{1}) = u(x)$, since $y(x + b\mathbf{1}) = y(x)$. Conversely, suppose that u is location invariant. Let $b = -x_n$. Then $u(x - x_n\mathbf{1}) = u((y, 0)) = u(x)$, where the second equality follows from the invariance of u. □

Lemma 1.8. *If u is location invariant then the conditional distribution of $u(X)$ given $\Theta = \theta$ is the same for (almost) all $\theta \in \mathbb{R}$; it does not depend on θ. Therefore, $u(X)$ is an* ***ancillary statistic*** *for Θ, $u(X) \perp\!\!\!\perp \Theta$ (or, being frequentist, the distribution of $u(X)$ does not depend on θ).*

Exercise 5 (Location invariant statistics are ancillary —— ***)**:**

Prove Lemma 1.8.

Lemma 1.9. *Suppose that ρ_0 is location equivariant. Then ρ_1 is location equivariant if and only if there is some location invariant function such that $\rho_1 = \rho_0 + u$.*

Proof. It is easy to check that $\rho_0 + u$ is location equivariant when u is location invariant. Conversely, if ρ_0 and ρ_1 are both location equivariant then $u = \rho_1 - \rho_0$ is location invariant. □

Takeaway 1.3. Lemmas 1.7 and 1.9 tell us that once we have *any* equivariant estimator with finite risk, say ρ_0, finding the lowest-risk equivariant estimator amounts to minimizing the (constant) risk of $\rho_0(x) + u(y(x))$ over functions u. This is a big simplification.

Minimum risk equivariant estimator under L_2 loss

Let's look at a specific version of this. Suppose that use L_2 loss, $L(\theta, c) = (\theta - c)^2$. Then the minimum risk equivariant (MRE) estimator is the so-called Pitman estimator for location families. Recall that $Y := (X_1 - X_n, X_2 - X_n, \ldots, X_{n-1} - X_n)$. Let $E_0[\rho_0(X)]$ denote the conditional expectation of $\rho_0(X)$ given $\Theta = 0$.

Theorem 1.10. *Suppose that $L(\theta, c) = (\theta - c)^2$ and that ρ_0 is any location equivariant estimator with finite risk. Then the MRE estimator is $\rho^*(X) = \rho_0(X) - E_0[\rho_0(X) \mid Y]$.*

Proof. Let ρ_0 be any location equivariant estimator with finite risk. By Lemmas 1.7 and 1.9, any other location equivariant estimator is of the form $\rho(X) = \rho_0(X) - v(Y(X))$, for some function v. The risk of ρ is

$$\begin{aligned} R(\theta, \rho) = R(0, \rho) &= E_0[(\rho_0(X) - v(Y))^2] \\ &= E_0[E_0[(\rho_0(X) - v(Y))^2 \mid Y]] \,. \end{aligned}$$

This is minimized by minimizing the inner conditional expectation uniformly in Y, achieved by setting $v(Y) = E_0[\rho_0(X) \mid Y]$. □

Exercise 6 (Pitman's estimator for the normal mean —— **)**:**

Show that Pitman's estimator for the normal mean (Example 1.1) is $\rho^*(X) = \bar{X}$.

Hint: Theorem 1.13 below makes this easy.

Theorem 1.10 can be generalized to other losses, though we may not be able to write the resulting estimators down in closed form.

Proposition 1.11. *If there exists a function $v^*(y)$ that minimizes*

$$E_0[\ell(\rho_0(X) - v(Y)) \mid Y] \,,$$

then the MRE estimator under the loss ℓ is

$$\rho^*(X) := \rho_0(X) - v^*(Y(X)) \,.$$

The proof is pretty much exactly as in Theorem 1.10 and can be found in [Rob07, Lemma 9.2.2]. Note that the proposition requires the existence of a minimizing function, which is not necessarily guaranteed; we usually need ℓ to satisfy some further conditions. See, for example, Schervish [Sch95, Lemma 6.15], for when ℓ is strictly convex and non-monotonic.

1.3 A Bayesian connection

Pitman's estimator is typically expressed in a different (more explicit) form, which allows us to make a connection to Bayesian inference for the problem. To do so, we need a bit more background from Bayesian decision theory.

A Bayesian approach to the problem from the previous section would be to specify a prior density on Θ, denoted f_Θ and to obtain the posterior density as

$$k_{\Theta|X}(\theta \mid x) = \frac{k_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{f_X(x)} \,.$$

The **posterior risk** of a location estimator ρ is

$$r(\rho; x) = \int_{\mathbb{R}} L(\theta, \rho(x)) k_{\Theta|X}(\theta \mid x) \, \mathrm{d}\theta \,. \tag{1.11}$$

A **formal Bayes rule** is an estimator ρ_0 such that $r(\rho_0; x) < \infty$ for all $x \in \mathbb{R}^n$ and $r(\rho_0; x) \leq r(\rho; x)$ for all x and all estimators ρ.

"Formal Bayes rule" is not so common these days; "Bayes optimal estimator" is a more modern term that generally means the same thing. The following is a standard result; see e.g., [Sch95, Ch. 3].

Proposition 1.12. *If the posterior variance of Θ given $X = x$ is finite then the formal Bayes rule for L_2 loss is $\rho(x) = E[\Theta \mid X = x]$.*

Theorem 1.13. *Pitman's estimator (as in Theorem 1.10) can be written as $E[\Theta \mid X]$ when a "uniform" prior is used for Θ. That is, the MRE estimator for L_2-loss is the formal Bayes rule with respect to using Lebesgue measure as the prior, if it has finite risk.*

Proof. Recall that for a location family, $k_{X|\Theta}(x \mid \theta) = q(x - \theta\mathbf{1})$. Transform X to (Y, X_n), and observe that the Jacobian-determinant of the transformation is 1. Therefore (recalling Theorem 1.2),

$$k_{Y,X_n|\Theta}(y, x_n \mid \theta) = q(y + (x_n - \theta)\mathbf{1}, x_n - \theta) \,.$$

The marginal density of Y is, with $u := x_n - \theta$,

$$k_{Y|\Theta}(y \mid \theta) = \int q(y + (x_n - \theta)\mathbf{1}, x_n - \theta)\,\mathrm{d}x_n = \int q(y + u\mathbf{1}, u)\,\mathrm{d}u\ .$$

Lemma 1.8 tells us that $k_{Y|\Theta}(y \mid \theta)$ does not depend on θ (because Y is ancillary).

Now let $\rho_0(x) = x_n$ in Theorem 1.10. We'll need to calculate the conditional expectation of X_n given Y at $\theta = 0$. We can write the conditional density of X_n given Y, Θ at $\theta = 0$ as

$$k_{X_n|Y,\Theta}(x_n \mid y, 0) = \frac{k_{Y,X_n|\Theta}(x_n \mid y, 0)}{k_{Y|\Theta}(y \mid 0)} = \frac{q(y + x_n\mathbf{1}, x_n)}{\int q(y + u\mathbf{1}, u)\,\mathrm{d}u}\ .$$

Then

$$v(y) = E_0[\rho_0(X) \mid Y = y] = \frac{\int uq(y + u\mathbf{1}, u)\,\mathrm{d}u}{\int q(y + u\mathbf{1}, u)\,\mathrm{d}u}\ .$$

Changing variables from u to $z = x_n - u$ so that $u = x_n - z$ and $y_i + u = x_i - z$, we find that

$$\begin{aligned}\rho(x) = x_n - v(y(x)) &= \frac{\int (x_n - u)q(y + u\mathbf{1}, u)\,\mathrm{d}u}{\int q(y + u\mathbf{1}, u)\,\mathrm{d}u} \\ &= \frac{\int zq(x - z\mathbf{1})\,\mathrm{d}z}{\int q(x - z\mathbf{1})\,\mathrm{d}z} = \frac{\int \theta k_{X|\Theta}(x \mid \theta)\,\mathrm{d}\theta}{\int k_{X|\Theta}(x \mid \theta)\,\mathrm{d}\theta} \\ &= E[\Theta \mid X = x]\ .\end{aligned}$$

The last equality follows if and only if the Lebesgue measure is used as the "prior". □

The use of a non-probability measure as prior (typically called an *improper prior*) is not what we're used to in Bayesian statistics. Strict Bayesians may take issue with it. I'll defer engaging with this deeply for now, but it's worth pointing out a few things. The Lebesgue prior $f_\Theta(\theta) = 1$ is compatible with the invariance structure present in the statistical model, and can be considered "non-informative", though that word should be treated with caution; see Robert [Rob07, Ch. 3.5] for an extended discussion on non-informative prior selection. As we will see, the properties of invariance-compatibility and non-informativeness generalize to other problems, and demonstrate a close connection between MRE estimators and certain Bayes estimators.

1.4 Minimaxity of location MRE estimators

It turns out that for location parameter problems, the Pitman (MRE) estimator is **minimax** under L_2-loss. That is, it attains the lower bound of the maximal risks,

$$\inf_\rho \sup_\theta R(\rho, \theta)\ .$$

Intuitively, since any location equivariant ρ_e estimator has constant risk, its worst-case risk is the same as its best-case, and does not depend on θ: $R(\rho_e, 0)$. Conversely, let ρ be any estimator (not necessarily equivariant) and consider

$$\tilde{\rho}(x) := \int_{\mathbb{R}} (\rho(x + s\mathbf{1}) - s)\,\mathrm{d}s\ , \quad x \in \mathbb{R}^n\ .$$

If ρ is location equivariant, then $\tilde{\rho}(x) = \rho(x)$. If it is not equivariant, then *if the integral converges*,

$$\tilde{\rho}(x + b\mathbf{1}) = \int_{\mathbb{R}} (\rho(x + (b+s)\mathbf{1}) - s)\, \mathrm{d}s \tag{1.12}$$

$$= \int_{\mathbb{R}} (\rho(x + u\mathbf{1}) - (u - b))\, \mathrm{d}u = \tilde{\rho}(x) + b\,, \tag{1.13}$$

and $\tilde{\rho}$ is location equivariant. This type of averaging procedure is quite general, but without further conditions *there's no guarantee that the integral converges.*

Moreover, for the sake of argument, *imagine* that the average in (1.12) was with respect to an invariant *probability measure* $\lambda(ds)$. Then if ℓ is convex, we could apply Jensen's inequality and Fubini's theorem to get

$$\begin{aligned}
R(\tilde{\rho}, \theta) &= \int_{\mathbb{R}} \ell\left(\int_{\mathbb{R}} (\rho(x + s\mathbf{1}) - s)\lambda(ds) - \theta\right) K_{X|\Theta}(\theta, dx) \\
&\leq \int_{\mathbb{R}} \int_{\mathbb{R}} \ell(\rho(x + s\mathbf{1}) - (s + \theta)) K_{X|\Theta}(\theta, dx)\lambda(ds) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \ell(\rho(x) - (s + \theta)) K_{X|\Theta}(s + \theta, dx)\lambda(ds) \\
&= \int_{\mathbb{R}} R(\rho, s + \theta)\lambda(ds) \\
&= \int_{\mathbb{R}} R(\rho, \theta)\lambda(d\theta) \leq \sup_{\theta} R(\rho, \theta)\,.
\end{aligned}$$

The upshot of this is that if we had a minimax estimator ρ then we could use it to generate a location equivariant estimator with the same risk function; it would also be minimax.

This argument works in some other (nice) cases, but falls apart here and in general because, except in a certain subclass of symmetry-related problems, *there is no invariant probability measure.* In general, we will have an invariant measure, but not one that can be made to integrate to 1. We can partially address this with improper priors, though that raises additional technical issues. We'll turn to the general framework and encounter this issue again.

It turns out that the Pitman estimator is, in fact, minimax for L_2-loss. The proof proceeds via a different argument than the (false) one outlined above, and can be found in Lehmann and Casella [LC98, Thm. 5.3.5].

1.5 Additional exercises

Exercise 7 (Half-normal MRE —— ***)**:**

[Sch95], Ch. 6, Problem 4 (p. 389).

2 Crash course: Group theory meets measure theory

Supplemental reading: Schervish [Sch95, Ch. 6.2], Eaton [Eat89, Ch. 1-2], Kallenberg [Kal17, Ch. 7.1]

2.1 A crash course in basic group theory

A **group** is a nonempty set, $\mathbf{G}$, equipped with a binary operation, $\circ$, such that the following conditions hold:

i) **Closure under $\circ$:** $g_1, g_2 \in \mathbf{G}$ implies $g_1 \circ g_2 \in \mathbf{G}$.

ii) **Associativity:** $(g_1 \circ g_2) \circ g_3 = g_1 \circ (g_2 \circ g_3)$, for $g_1, g_2, g_3 \in \mathbf{G}$.

iii) **Identity:** There exists an element $\mathrm{e} \in \mathbf{G}$ such that $\mathrm{e} \circ g = g \circ \mathrm{e} = g$ for $g \in \mathbf{G}$.

iv) **Inverse:** For each $g \in \mathbf{G}$, there exists a unique element $g^{-1} \in \mathbf{G}$ such that $g \circ g^{-1} = g^{-1} \circ g = \mathrm{e}$.

For ease of notation, the group operation is often written as multiplication, with $\circ$ suppressed, i.e., $g_1 g_2$ is written instead of $g_1 \circ g_2$. Technically, a group consists of the set and the operation, and would be denoted by $(\mathbf{G}, \circ)$, but when the operation is clear from context it is often written as $\mathbf{G}$.

Example 2.1 Examples of groups. Here are some examples of groups. You can check that they satisfy the defining group properties as an easy exercise.

- $(\mathbb{R}^n, +)$, the n-dimensional real numbers with addition as the group operation. Here, $\mathrm{e} = 0 \cdot \mathbf{1}$, and each $g \in \mathbb{R}^n$ has inverse $-g$. This group is **commutative** (also called **Abelian**) because $g_1 + g_2 = g_2 + g_1$.
- $(\mathbb{R} \setminus \{0\}, \times)$, the real numbers except zero, with multiplication as the group operation. The identity is 1 and the inverse of g is $1/g$.
- $\mathrm{GL}(n, \mathbb{R})$, or $\mathrm{GL}(n)$ for short, the **general linear group**, consisting of all $n \times n$ invertible (i.e., non-singular) matrices with matrix multiplication as the group operation. The identity is the identity matrix $\mathbf{I}_n$, and $g \in \mathrm{GL}(n, \mathbb{R})$ is a $n \times n$ $\mathbb{R}$-valued matrix with group inverse equal to the matrix inverse g^{-1}.
- $\mathrm{O}(n)$, the group of real $n \times n$ orthogonal matrices, with matrix multiplication as the group operation. (Recall that a matrix g is orthogonal if $g^\top g = \mathbf{I}_n$.) Note that $\mathrm{O}(n) \subsetneq \mathrm{GL}(n)$ with the same group operation, and that $\mathrm{O}(n)$ is closed in $\mathrm{GL}(n)$; thus $\mathrm{O}(n)$ is a **subgroup**[a] of $\mathrm{GL}(n)$.
- $\mathrm{SO}(n)$, the subgroup of $\mathrm{O}(n)$ of orthogonal matrices with determinant equal to $+1$. This is the **special orthogonal group**.

[a] Proper definition: $\mathbf{H} \subset \mathbf{G}$ is a subgroup of $\mathbf{G}$ if $\mathbf{H}$ is a group with the group operation inherited from $\mathbf{G}$.

Group morphisms

Let $\mathbf{G}$ and $\mathbf{H}$ be two groups with group operations $\circ$, $*$, respectively. A **group homomorphism** is a function $\phi\colon \mathbf{G} \to \mathbf{H}$ such that

$$\phi(g_1 \circ g_2) = \phi(g_1) * \phi(g_2)\ , \quad g_1, g_2 \in \mathbf{G}\ . \tag{2.1}$$

Thus, $\phi(e_{\mathbf{G}}) = e_{\mathbf{H}}$ and $\phi(g^{-1}) = \phi(g)^{-1}$ for all $g \in \mathbf{G}$. In short, a group homomorphism preserves the group structure.

Exercise 8 (Group homomorphism properties —— *)**:**

Show that (2.1) implies that $\phi(e_{\mathbf{G}}) = e_{\mathbf{H}}$ and $\phi(g^{-1}) = \phi(g)^{-1}$ for all $g \in \mathbf{G}$.

A group homomorphism that is bijective is a **group isomorphism**; two groups between which there is a group isomorphism are called isomorphic, and for most practical purposes can be considered identical.

Example 2.2 Group morphisms. Here are some easy examples of -morphisms.

- Fix $c \in \mathbb{R}$, and take the group $(\mathbb{R}, +)$. Then $M_c(g) := cg$ is a homomorphism to $(\mathbb{R}, +)$ because $M_c(g_1 + g_2) = M_c(g_1) + M_c(g_2)$.
- Take the group as $(\mathbb{R}_{\neq 0}, \times)$, and let $s(g)$ return the sign of g. Then s is a group homomorphism into $(\{-1, +1\}, \times)$.
- The exponential function is a group isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}_+, \times)$.
- Take any two groups, $\mathbf{G}$ and $\mathbf{H}$. The *trivial homomorphism* is $\phi(g) = e_{\mathbf{H}}$ for all $g \in \mathbf{G}$.
- A linear representation of a group $\mathbf{G}$ is a group homomorphism from $\mathbf{G}$ to $\mathrm{GL}(n, \mathbb{R})$. This represents a group element as a real $n \times n$ matrix that acts naturally on vectors in $\mathbb{R}^n$.

Group actions

One of the primary uses of a group $\mathbf{G}$ is to index a set of one-to-one and onto transformations of a set $\mathbf{X}$ to itself; in this case, the elements of $\mathbf{G}$ are regarded as functions on $\mathbf{X}$, and the group operation is function composition. Thus, a group action of $\mathbf{G}$ on a set $\mathbf{X}$ is a group homomorphism from $\mathbf{G}$ into $\mathrm{Aut}(\mathbf{X})$, the set of automorphisms of $\mathbf{X}$. While correct, the definition is a bit abstract; here is a more usable definition.

A function $\varphi\colon \mathbf{G} \times \mathbf{X} \to \mathbf{X}$ is a (left) **group action** if it satisfies:

i) $\varphi(e, x) = x$, all $x \in \mathbf{X}$.

ii) $\varphi(g_1 g_2, x) = \varphi(g_1, \varphi(g_2, x))$, all $g_1, g_2 \in \mathbf{G}$ and $x \in \mathbf{X}$.

We can hook this into the abstract definition as follows. For each $g \in \mathbf{G}$, define $T_g\colon \mathbf{X} \to \mathbf{X}$ by $T_g(x) = \varphi(g, x)$. Then T_e is the identity function and $T_{g_1}(T_{g_1}(x)) = T_{g_1 g_2}(x)$. It is easy to verify that $T_{g^{-1}} = T_g^{-1}$, and that each T_g is bijective. Thus $\{T_g\colon g \in \mathbf{G}\}$ is a subgroup of $\mathrm{Aut}(\mathbf{X})$ with function composition as the group operation.

Using this notation for φ and T can get burdensome, so it is common to write that the action of $\mathbf{G}$ on $\mathbf{X}$ is via the map $(g, x) \mapsto gx$, with $ex = x$ and $(g_1 g_2)x = g_1(g_2 x)$. I will follow this convention and most of the time there's no issue, but I urge you to keep in mind the "group action is a group of transformations that is a subgroup of the automorphisms" perspective. In particular, the gx notation can easily hide the fact that the same group can be made to have more than one action on a set. (For example, two non-equivalent linear representations of the same group.)

A group action induces an equivalence relation $\sim$ between elements of $\mathbf{X}$: $x_1 \sim x_2$ if and only if there is some $g \in \mathbf{G}$ such that $gx_1 = x_2$. This partitions $\mathbf{X}$ into disjoint sets called **orbits**. The

orbit of x is

$$\mathbf{G}x := \{gx \colon g \in \mathbf{G}\} \,, \tag{2.2}$$

and contains exactly the points that are equivalent to x. An action is **transitive** on $\mathbf{X}$ if there is only one orbit, in which case $\mathbf{X}$ is sometimes called a **homogeneous space** of $\mathbf{G}$, or $\mathbf{G}$-space.

The **stabilizer subgroup** (or **isotropy subgroup**) of a point x is

$$\mathbf{G}_x := \{g \in \mathbf{G} \colon gx = x\} \,. \tag{2.3}$$

An action on $\mathbf{X}$ is **free** if $gx = x$ for *some* $x \in \mathbf{X}$ implies that $g = \mathrm{e}$. In other words, $\mathbf{G}_x = \{\mathrm{e}\}$ for each $x \in \mathbf{X}$. A weaker property is a **faithful** action: $gx = x$ for *all* $x \in \mathbf{X}$ implies that $g = \mathrm{e}$.

If $\mathbf{G}$ acts freely and transitively on $\mathbf{X}$ then $\mathbf{X}$ is sometimes called a **principal homogeneous space** of $\mathbf{G}$. If we choose a point $x_0 \in \mathbf{X}$, then $\mathbf{G}$ is essentially in one-to-one correspondence with $\mathbf{X}$. We will in this case say that the action is **exactly transitive**.

Example 2.3 Group actions. Here are some examples of group actions.

- Consider $\mathbf{G} = \mathrm{GL}(n)$ and $\mathbf{X} = \mathbb{R}^n$. Define the group action by $\varphi(g, x) = gx$, which is matrix-vector multiplication. It is an easy exercise to check that this defines a valid group action.
- Now consider $\mathrm{SO}(n)$ acting on $\mathbb{R}^n$. Since $\mathrm{SO}(n)$ is a subgroup of $\mathrm{GL}(n)$, we can just restrict the action of $\mathrm{GL}(n)$, so that again we have $\varphi(g, x) = gx$ defined by matrix-vector multiplication. This is just a rotation of x, which gives $\mathrm{SO}(n)$ its other common name, the **rotation group**. The orbit of x is the set of all $x' \in \mathbf{X}$ with $||x'|| = ||x||$. The action is not transitive on $\mathbb{R}^n$ because there is no $g \in \mathrm{SO}(n)$ such that $gx = y$, for any $||y|| \neq ||x||$.
- Now consider $\mathrm{SO}(3)$ acting on $\mathbb{S}(2)$, the unit sphere embedded in $\mathbb{R}^3$, again as gx. The action is transitive but not free; the north and and south poles are fixed by all rotations about the "z-axis".
- Consider $\mathrm{SO}(2)$ acting on $\mathbb{S}(1)$, again as gx, i.e., rotations of the unit circle. This action is transitive and free.

2.2 Integration on groups and spaces on which they act

So far, we haven't said anything about measurability, etc., but we'll need that in order to talk about measures, integrals, and everything else we need to develop generalizations of the results in Section 1.

For those interested in the technical bits, our blanket assumptions on any group $\mathbf{G}$ that we encounter are:

i) The set $\mathbf{G}$ is a topological space whose topology has a countable base (also known as second-countable[1]); it is also locally compact (every point in $\mathbf{G}$ has a compact neighborhood) and Hausdorff.[2] For short, we say that such a group is lcscH (locally compact, second-countable

[1] This is also called a **completely separable space**. The definition is that the space has a countable collection of open sets, $\mathcal{U}$, such that *any* open subset of the space can be obtained as the union of some subcollection of sets in $\mathcal{U}$.

[2] A topological space $\mathbf{X}$ is a **Hausdorff space** if every pair of distinct points have neighborhoods (sets containing an open set that contain the point) that are disjoint.

Hausdorff).[3]

ii) The functions $(g_1, g_2) \to g_1 g_2$ (composition) and $g \to g^{-1}$ (inversion) are continuous.

iii) Generating a σ-algebra from the topology, we have $(\mathbf{G}, \mathcal{B}(\mathbf{G}))$ is a standard Borel space; the continuity of the group composition and inversion makes them measurable.

Haar measure

I will state a number of results here without proof. A good reference for proofs and details is Folland [Fol16, Ch. 2.2]. Eaton [Eat89], Wijsman [Wij90], and Schervish [Sch95] each cover parts of the theory of the Haar integral, too.

For a set $B \in \mathcal{B}(\mathbf{G})$ and $g \in \mathbf{G}$, we write

$$gB := \{gh \colon h \in B\} \quad \text{and} \quad Bg := \{hg \colon h \in B\} \, .$$

We also write $B^{-1} := \{g^{-1} \colon g \in B\}$. A measure μ on a group $\mathbf{G}$ is *left-invariant* if

$$\mu(gB) = \mu(B) \, , \quad g \in \mathbf{G}, \ B \in \mathcal{B}(\mathbf{G}) \, . \tag{2.4}$$

A measure ν on $\mathbf{G}$ is *right-invariant* if

$$\nu(Bg) = \nu(B) \, , \quad g \in \mathbf{G}, \ B \in \mathcal{B}(\mathbf{G}) \, . \tag{2.5}$$

A left-invariant Radon measure[4] $\lambda_\ell \neq 0$ on $\mathbf{G}$ is called a **left Haar measure** on $\mathbf{G}$. A right-invariant Radon measure $\lambda_r \neq 0$ on $\mathbf{G}$ is called a **right Haar measure** on $\mathbf{G}$. If $f \colon \mathbf{G} \to \mathbb{R}$ is an integrable function then

$$\int_{\mathbf{G}} f(gh) \lambda_\ell(dh) = \int_{\mathbf{G}} f(h) \lambda_\ell(dh) \, , \quad g \in \mathbf{G} \, , \tag{2.6}$$

and

$$\int_{\mathbf{G}} f(hg) \lambda_r(dh) = \int_{\mathbf{G}} f(h) \lambda_r(dh) \, , \quad g \in \mathbf{G} \, . \tag{2.7}$$

Moreover, if one has a left Haar measure λ_ℓ, one can get a right Haar measure by defining $\lambda_r(B) = \lambda_\ell(B^{-1})$. Then it is straightforward to see that λ_r so defined is right-invariant:

$$\lambda_r(Bg) = \lambda_\ell(g^{-1} B^{-1}) = \lambda_\ell(B^{-1}) = \lambda_r(B) \, , \quad g \in \mathbf{G} \, , \ B \in \mathcal{B}(\mathbf{G}) \, . \tag{2.8}$$

When $\mathbf{G}$ is lcscH, Haar measures are known to exist and are unique up to normalization. I'll state this as a theorem for posterity; see, e.g., Kallenberg [Kal02] Ch. 2 for a proof of this.

[3] We won't ever need to appeal to these directly; they're just sufficient conditions for the existence/regularity of the measures, integrals, etc., that we will use. I'm including the assumption here for those who want to know these things.

[4] A **Radon measure** is a measure on the Borel sets that is finite on all compact sets, outer regular on all Borel sets, and inner regular on all open sets. I won't define these terms here; see, e.g., [AB06] if you're interested. The main thing for our purposes is that a Radon measure is finite on compact sets.

Theorem 2.1. *On every lcscH group* $\mathbf{G}$*, there exists a left-invariant Radon measure* $\lambda \neq 0$*, and it is unique up to normalization. If* $\mathbf{G}$ *is compact then* λ *is also right-invariant.*

That is, if $\lambda_\ell, \lambda'_\ell$ are two left Haar measures on $\mathbf{G}$, then there is some positive constant c such that $\lambda_\ell = c\lambda'_\ell$. Similarly for right Haar measures.

It turns out that if we have *any* left Haar measure, there's a nice way to obtain any other left Haar measure. It's surprisingly simple: just shift the left Haar measure that we do have from the right.

In particular: suppose we have a left Haar measure λ_ℓ, then fix $g \in \mathbf{G}$ and define $\lambda'_\ell(B) := \lambda_\ell(Bg)$. That is, λ'_ℓ is just the pushforward of λ_ℓ under the group operation on the right. Observe that for any integrable f, the integral

$$\int_{\mathbf{G}} f(h_0 h g^{-1})\lambda_\ell(dh) = \int_{\mathbf{G}} f(hg^{-1})\lambda_\ell(dh) = \int_{\mathbf{G}} f(h)\lambda'_\ell(dh)\ , \quad h_0 \in \mathbf{G}\ ,$$

is left-invariant. Since this is true for any integrable f, this determines that λ'_ℓ is another left Haar measure; the uniqueness up to normalization implies that

$$\int_{\mathbf{G}} f(hg^{-1})\lambda_\ell(dh) = c\int_{\mathbf{G}} f(h)\lambda_\ell(dh)\ , \tag{2.9}$$

where $c \in \mathbb{R}_+$ is the constant of proportionality between the two left Haar measures.

The group element g was arbitrary, and we could repeat the argument with some $g' \neq g$, which would lead to a different left Haar measure and a different constant of proportionality; thus the c in (2.9) is really the evaluation at g of a function $\Delta\colon \mathbf{G} \to \mathbb{R}_+$. The function Δ is called the (right-hand) **modular function**[5] of $\mathbf{G}$. It is straightforward to show that Δ is a group homomorphism from $\mathbf{G}$ to $(\mathbb{R}_+, \times)$ (it is also continuous, and therefore measurable, though that's harder to show; see Prop. 2.24 in [Fol16]). In particular, it satisfies

$$\Delta(g_1 g_2) = \Delta(g_1)\Delta(g_2)\ , \quad \Delta(\mathrm{e}) = 1\ , \quad \Delta(g^{-1}) = 1/\Delta(g)\ . \tag{2.10}$$

Moreover, Δ is "universal" in that it does not depend on the original choice of Haar measure λ_ℓ.

For any integrable f,

$$\int_{\mathbf{G}} f(h)\Delta(h^{-1})\lambda_\ell(dh) = \int_{\mathbf{G}} f(h)\lambda_r(dh) = \int_{\mathbf{G}} f(h^{-1})\lambda_\ell(dh)\ . \tag{2.11}$$

Note that some authors use g instead of g^{-1} in (2.9), which would define "different" modular function $\tilde{\Delta}$ related to Δ by $\tilde{\Delta}(g) = \Delta(g^{-1})$. So there's no difference except for notation.

A group $\mathbf{G}$ is called **unimodular** if $\Delta(g) = 1$ for all $g \in \mathbf{G}$. Two important classes of unimodular groups are:

- Compact groups. Since Δ is a group homomorphism from $\mathbf{G}$ into $(\mathbb{R}_+, \times)$, if $\mathbf{G}$ is compact then the image of Δ must be a compact subgroup of $(\mathbb{R}_+, \times)$. There is exactly one such subgroup: $\{1\}$. In this case, $\lambda_\ell(\mathbf{G}) < +\infty$ and therefore there is a *unique* normalized Haar measure $\lambda = \lambda_\ell/\lambda_\ell(\mathbf{G}) = \lambda_r/\lambda_r(\mathbf{G})$ that is both left- and right-invariant. For compact groups, it is typical to refer to the unique normalized Haar measure simply as "the Haar measure." Observe that since $\lambda(\mathbf{G}) = 1$, it is a probability measure on $\mathbf{G}$ that acts like the uniform distribution.

[5] Some authors call this the modulus of $\mathbf{G}$.

- Finite discrete groups. Similar reasoning as above.
- Commutative groups. If $\mathbf{G}$ is commutative, then $gB = Bg$, and therefore a left Haar measure is also right-invariant: $\lambda_\ell(gB) = \lambda_\ell(Bg) = \lambda_\ell(B)$. As with compact groups, this is the unique Haar measure and implies that $\Delta = 1$.

Takeaway 2.1. For a lcscH group $\mathbf{G}$, we can (for our purposes) summarize this section with the following useful relationships.

- Modular function:

$$\Delta(g_1 g_2) = \Delta(g_1)\Delta(g_2)\ , \quad \Delta(\mathrm{e}) = 1\ , \quad \Delta(g^{-1}) = 1/\Delta(g)\ .$$

- Left Haar measure:

$$\lambda_\ell(gB) = \lambda_\ell(B) \quad \text{and} \quad \lambda_\ell(Bg) = \Delta(g)\lambda_\ell(B)\ , \tag{2.12}$$

and

$$\int_{\mathbf{G}} f(gh)\lambda_\ell(dh) = \int_{\mathbf{G}} f(h)\lambda_\ell(dh) \quad \text{and} \quad \int_{\mathbf{G}} f(hg)\lambda_\ell(dh) = \Delta(g^{-1})\int_{\mathbf{G}} f(h)\lambda_\ell(dh)\ . \tag{2.13}$$

- Right Haar measure:

$$\lambda_r(Bg) = \lambda_r(B) \quad \text{and} \quad \lambda_r(gB) = \Delta(g^{-1})\lambda_r(B)\ , \tag{2.14}$$

and

$$\int_{\mathbf{G}} f(hg)\lambda_r(dh) = \int_{\mathbf{G}} f(h)\lambda_r(dh) \quad \text{and} \quad \int_{\mathbf{G}} f(gh)\lambda_r(dh) = \Delta(g)\int_{\mathbf{G}} f(h)\lambda_r(dh)\ . \tag{2.15}$$

- Left and right Haar measure:

$$\lambda_\ell(B) = \lambda_r(B^{-1})\ , \tag{2.16}$$

and

$$\lambda_r(dg) = \Delta(g)^{-1}\lambda_\ell(dg)\ . \tag{2.17}$$

Example 2.4 Haar measures. Here are some examples of Haar measures.

- Take $\mathbf{G} = (\mathbb{R}, +)$, which is commutative. Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is left- and right-invariant, and it is σ-finite, so it is the unique Haar measure on $(\mathbb{R}, +)$, and $\Delta = 1$. Similarly for $\mathbb{R}^n$.
- Let $\mathbf{G}$ be any finite (discrete) group. Then counting measure is left- and right-invariant, and it can be normalized by $|\mathbf{G}|$.
- Let $\mathbf{G}$ be an infinite countable (discrete) group. Again, counting measure is left- and right-invariant, but $\lambda(\mathbf{G}) = +\infty$ so it can't be normalized into a probability measure.

Eaton [Eat89, Ch. 1.3] has some examples of groups that are not unimodular.

Invariant measures induced by a group action

In the next section, we will be interested in building statistical models from spaces on which a group acts, and we'll need to induce measures that behave like Haar measure. To that end, let $\mathbf{X}$ be a space on which $\mathbf{G}$ acts measurably and transitively. Fix some $x_0 \in \mathbf{X}$, and define $\beta\colon \mathbf{G} \to \mathbf{X}$ by

$$\beta(g) = gx_0 \ .$$

This maps $\mathbf{G}$ onto $\mathbf{X}$ because it is assumed that the action is transitive. However, it might not be one-to-one; this will be the case when the action of $\mathbf{G}$ on $\mathbf{X}$ is not free, i.e., there are non-trivial stabilizer subgroups.

Recall that $\mathbf{G}_{x_0} = \{g \in \mathbf{G}\colon gx_0 = x_0\}$ is the stabilizer subgroup of x_0. If it is not trivial (i.e., contains more than just e), then if $x = hx_0$, then also $x = hgx_0$ for any $g \in \mathbf{G}_{x_0}$. In particular, the inverse image of $\{x_0\}$ is $\beta^{-1}(\{x_0\}) = \mathbf{G}_{x_0}$. Moreover, it is easy to show that if $x = hx_0$ then $\mathbf{G}_x = h\mathbf{G}_{x_0}h^{-1}$. So when the action is transitive (which we're assuming in this section) then all of the stabilizer subgroups are equivalent in this sense; if one is measure theoretically problematic, then they all are.

Exercise 9 (Stabilizers are conjugate —— *)**:**

> Let $\mathbf{G}_x$ be the stabilizer subgroup of $x \in \mathbf{X}$. Fix some $y \neq x$ with $y = hx$ for some $h \in \mathbf{G}$. Show that $\mathbf{G}_y = h\mathbf{G}_x h^{-1}$. Two subgroups that satisfy such a relationship are called **conjugate subgroups**.

Back to the task at hand. We'd like to come up with an invariant measure λ_ℓ^x on $\mathbf{X}$, in the sense that

$$\lambda_\ell^x(gB) = \lambda_\ell^x(B) \ , \quad g \in \mathbf{G} \ , \ B \in \mathcal{B}(\mathbf{X}) \ . \tag{2.18}$$

(Actually, this isn't quite what we want but we can get what we want from there.)

Intuitively, we would like to define a measure on $\mathbf{X}$ as the image of λ_ℓ under β, i.e., $\lambda_\ell^x(B) := \lambda_\ell(\beta^{-1}(B))$. We could run into trouble if, for example, $\lambda_\ell(\mathbf{G}_{x_0}) = \infty$. One way to avoid this is to require the action to be **proper**: the pre-image of every compact subset of $\mathbf{X}$ is a compact subset of $\mathbf{G}$. More precisely, the map $(g, x) \to (gx, x)$ must be proper. This is a sufficient condition for the existence of the induced invariant measures that we'll need.

Theorem 2.2. *Let $\mathbf{G}$ be a lcscH group that acts transitively and properly on $\mathbf{X}$. Fix a reference point $x_0 \in \mathbf{X}$ and define $\beta\colon \mathbf{G} \to \mathbf{X}$ by $\beta(g) = gx_0$. Then $\lambda_\ell^x := \beta_\# \lambda_\ell$ is a left-invariant Radon measure on $\mathbf{X}$. $\lambda_r^x = \beta_\# \lambda_r$ that is a "right-invariant" Radon measure on $\mathbf{X}$, in the sense that*

$$\lambda_r^x(gB) = \Delta(g)^{-1}\lambda_r^x(B) \ , \quad g \in \mathbf{G} \ , \ B \in \mathcal{B}(\mathbf{X}) \ . \tag{2.19}$$

Both λ_ℓ^x and λ_r^x are unique up to normalization. Moreover, λ_r^x does not depend on the reference point x_0.

Proof. λ_ℓ^x is a Radon measure because λ_ℓ is a Radon measure and the group action is proper; likewise for λ_r^x. It's easy to check that λ_ℓ^x is invariant under the action of $\mathbf{G}$ on $\mathbf{X}$: for any positive

measurable $f\colon \mathbf{X} \to \mathbb{R}_+$ and $h \in \mathbf{G}$,

$$\int_{\mathbf{X}} f(hx)\lambda_\ell^x(dx) = \int_{\mathbf{G}} f(hgx_0)\lambda_\ell^x(dx) = \int_{\mathbf{G}} f(gx_0)\lambda_\ell^x(dx) = \int_{\mathbf{X}} f(x)\lambda_\ell^x(dx) \ ,$$

by the left-invariance of λ_ℓ. λ_ℓ^x is unique up to normalization; the proof is fairly involved [see Kal02, Thm. 2.29] and relies on properness of the group action to ensure that $\mathbf{G}_{x_0}$ is compact.

Similarly, we can check that λ_r^x obeys (2.19). For any positive measurable $f\colon \mathbf{X} \to \mathbb{R}_+$ and $h \in \mathbf{G}$,

$$\int_{\mathbf{X}} f(hx)\lambda_r^x(dx) = \int_{\mathbf{G}} f(hgx_0)\lambda_r(dg) = \int_{\mathbf{G}} f(gx_0)\Delta(h)\lambda_r(dg) = \int_{\mathbf{X}} \Delta(h)f(x)\lambda_r^x(dx) \ .$$

where the second equality is from (2.11). This is also unique up to normalization, by a slight adaptation of the proof of uniqueness up to normalization of λ_ℓ^x.

Now suppose we choose $x_0' = hx_0$ as our reference point. Let $\beta'(g) = gx_0'$ and $\lambda_r^{x\prime} = \beta'_\# \lambda_r$. Then by the right-invariance of λ_r,

$$\int_{\mathbf{X}} f(x)\lambda_r^{x\prime}(dx) = \int_{\mathbf{G}} f(gx_0')\lambda_r(dg) = \int_{\mathbf{G}} f(ghx_0)\lambda_r(dg) = \int_{\mathbf{G}} f(gx_0)\lambda_r(dg) = \int_{\mathbf{X}} f(x)\lambda_r^x(dx) \ ,$$

so λ_r^x is independent of x_0. Note, however, that the same argument cannot be made for λ_ℓ^x; we would get a factor of $\Delta(h)^{-1}$ if we changed the reference point. □

Note that an equivalent way of stating (2.19) is

$$g_\#^{-1}\lambda_r^x = \Delta(g)^{-1}\lambda_r^x \ , \quad g \in \mathbf{G} \ , \tag{2.20}$$

which is handy for integration.

2.3 Some odds and ends

We'll need a few more ideas that fit naturally here. A function $f\colon \mathbf{X} \to \mathbf{Y}$ is $\mathbf{G}$-invariant if $f(gx) = f(x)$ for each $g \in \mathbf{G}$ and $x \in \mathbf{X}$. It is a **maximal invariant** if $f(x) = f(x')$ implies that $x = gx'$ for some $g \in \mathbf{G}$. In short, a maximal invariant uniquely identifies the orbits of $\mathbf{X}$. Because of this, every $\mathbf{G}$-invariant function can be written as a function of a maximal invariant.

Proposition 2.3. *Suppose $m\colon \mathbf{X} \to \mathbf{Y}$ is a maximal invariant of $\mathbf{G}$ acting on $\mathbf{X}$. Then a function $h\colon X \to \mathbf{Z}$ is $\mathbf{G}$-invariant if and only if there is a function $t\colon \mathbf{Y} \to \mathbf{Z}$ such that $h(x) = t(m(x))$.*

If you took STAT 547C with me, this might look sort of familiar. There is a deeper measure theoretic result that implies the previous one.

A set $B \in \mathcal{B}(\mathbf{X})$ is $\mathbf{G}$-invariant if $g^{-1}B = B$ for all $g \in \mathbf{G}$. The collection of all $\mathbf{G}$-invariant sets is a σ-algebra, which we'll denote by $\mathcal{I}(\mathbf{X}) \subset \mathcal{B}(\mathbf{X})$.

Exercise 10 (Collection of $\mathbf{G}$-invariant sets is a σ-algebra —— **)**:**

Show that $\mathcal{I}(\mathbf{X})$ is a sub-σ-algebra of $\mathcal{B}(\mathbf{X})$.

We'll use $\mathbb{R}_+$ in the next theorem for convenience; we could use any standard Borel space.

Theorem 2.4. *A Borel-measurable function $f\colon \mathbf{X} \to \mathbb{R}_+$ is $\mathbf{G}$-invariant if and only if it is measurable with respect to $\mathcal{I}(\mathbf{X})$. Let $m\colon \mathbf{X} \to \mathbb{R}_+$ be a Borel-measurable maximal invariant of $\mathbf{G}$ acting on $\mathbf{X}$. Then the σ-algebra generated by m, $\sigma(m) := \{m^{-1}(B)\colon B \in \mathcal{B}(\mathbb{R}_+)\}$, is the same as $\mathcal{I}(\mathbf{X})$.*

An implication of the first part is that $\mathcal{I}(\mathbf{X})$ is generated by the collection of all $\mathbf{G}$-invariant functions.

Exercise 11 (A maximal invariant generates the invariant σ-algebra —— ****)**:**

Prove Theorem 2.4.

Optional easier version: Prove the second part of Theorem 2.4 only for the case when there are only countably many orbits generated by $\mathbf{G}$ acting on $\mathbf{X}$.

Corollary 2.5. *Let $m\colon \mathbf{X} \to \mathbb{R}_+$ be a Borel-measurable maximal invariant of $\mathbf{G}$ acting on $\mathbf{X}$. A Borel-measurable function $f\colon \mathbf{X} \to \mathbb{R}_+$ is $\mathbf{G}$-invariant if and only if there is some measurable function $t\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $f(x) = t(m(x))$.*

This follows from Theorem 2.4 and the Doob–Dynkin factorization [e.g., Kal02, Lemma 1.13].

3 Equivariant statistical analysis

Supplemental reading: Schervish [Sch95, Ch. 6.2], Eaton [Eat89, Ch. 6]

In this section, we'll generalize the basic location family setup from Section 1 to group-invariant estimation and decision problems.

3.1 Group-invariant decision problems

Generalizing our setup from earlier, assume we have:

- a sample space $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$;
- a parameterized family of Markov probability kernels $\{K_{X|\Theta}(\theta, \bullet) \colon \theta \in \mathbf{\Theta}\}$ from the parameter space parameter space $(\mathbf{\Theta}, \mathcal{B}(\mathbf{\Theta}))$ into $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$;
- a decision or action space, $(\mathbf{A}, \mathcal{B}(\mathbf{A}))$; and
- a loss function $L \colon \mathbf{\Theta} \times \mathbf{A} \to [0, \infty)$.

A (randomized) **decision rule** (or action rule) ρ is a Markov kernel from $\mathbf{X}$ into $\mathbf{A}$. Recall that a Markov kernel is a measurable function of its "input space" ($\mathbf{X}$ here) and a measure on its "output space" ($\mathbf{A}$ here). For this problem, we require ρ to be a *probability* measure on $\mathbf{A}$, which makes it appropriate as a model for a randomized decision/action in response to observing $X = x$. A decision rule is *nonrandomized* if it puts probability 1 at one point of $\mathbf{A}$, say $\zeta(x)$. This corresponds to $\rho(x, B) = \delta_{\zeta(x)}(B)$ for $B \in \mathcal{B}(\mathbf{A})$, i.e., ρ is a Dirac measure, and induces a measurable function $x \mapsto \zeta(x)$ (and vice versa).

The conditional risk of a randomized decision rule at θ is

$$R(\theta, \rho) = \int_{\mathbf{X}} \int_{\mathbf{A}} L(\theta, a) \rho(x, da) K_{X|\Theta}(\theta, dx) \, . \tag{3.1}$$

This is just a slight generalization of what we saw in Section 1 with nonrandomized decision rules, which we recover when $\rho(x, \bullet) = \delta_{\zeta(x)}(\bullet)$ as

$$R(\theta, \delta_a) = \int_{\mathbf{X}} \int_{\mathbf{A}} L(\theta, a) \delta_{\zeta(x)}(da) K_{X|\Theta}(\theta, dx) = \int_{\mathbf{X}} L(\theta, \zeta(x)) K_{X|\Theta}(\theta, dx) \, .$$

To set up the symmetry structure, different authors proceed in different ways. The most common is to start with a topological group $\mathbf{G}$ that acts on $\mathbf{X}$, and use it to induce homomorphisms of $\mathbf{G}$ that act on $\mathbf{\Theta}$ and on $\mathbf{A}$. For example, Schervish [Sch95, Ch. 6] takes this approach. We could proceed that way, but we'll take a shortcut that makes an extra assumption that the group homomorphisms turn out to be isomorphisms. This is in general a big assumption but it makes the presentation of the main ideas much cleaner. If we want to avoid some technical difficulties (which are important in general but not for getting the main ideas), we can either make this assumption now and carry around some extra notation, or make some other similar assumptions later on. So we'll make the assumption now and run with it.

So: let $\mathbf{G}$ be a lcscH group that acts measurably on each of $\mathbf{X}, \mathbf{\Theta}, \mathbf{A}$. The action may be different on each of the three spaces.

Definition 3.1. The decision problem with sample space, parameter space, etc., as listed above, is a **G-invariant decision problem** if:

(i) The family of Markov kernels $\{K_{X|\Theta}(\theta, \bullet)\colon \theta \in \mathbf{\Theta}\}$ is invariant, that is if for each $g \in \mathbf{G}$ and $\theta \in \mathbf{\Theta}$, we have that $g\theta \in \mathbf{\Theta}$ and

$$K_{X|\Theta}(g\theta, \bullet) = g_{\#}K_{X|\Theta}(\theta, \bullet)\ . \tag{3.2}$$

(ii) The loss function L is invariant, that is

$$L(g\theta, ga) = L(\theta, a)\ , \quad \text{each } g \in \mathbf{G}\ ,\ \theta \in \mathbf{\Theta}\ ,\ a \in \mathbf{A}\ .$$

Note that part (i) of the definition is slightly different from what is typical. The typical definition is that the family of probability measures $\{P_\theta\colon \theta \in \mathbf{\Theta}\}$ is invariant, i.e., $g\theta \in \mathbf{\Theta}$ for each $g \in \mathbf{G}$, $\theta \in \mathbf{\Theta}$, and that

$$P_{g\theta}(B) = P_\theta(g^{-1}B)\ , \quad g \in \mathbf{G}\ ,\ \theta \in \mathbf{\Theta}\ ,\ B \in \mathcal{B}(\mathbf{X})\ .$$

If, for each $B \in \mathcal{B}(\mathbf{X})$, we require the mapping $\theta \mapsto P_\theta(B)$ to be measurable, then it's easy to see that the two definitions are equivalent.

A decision rule ρ is a **G-equivariant decision rule** if

$$\rho(gx, \bullet) = g_{\#}\rho(x, \bullet)\ , \quad g \in \mathbf{G},\ x \in \mathbf{X}\ . \tag{3.3}$$

Observe that both (3.2) and (3.3) are special cases of the following, which defines a **G**-equivariant Markov kernel $Q_{Z|Y}$ from $\mathbf{Y}$ to $\mathbf{Z}$:

$$Q_{Z|Y}(gy, dz) = g_{\#}Q_{Z|Y}(y, dz)\ , \quad g \in \mathbf{G}\ ,\ y \in \mathbf{Y}\ . \tag{3.4}$$

Exercise 12 (Equivariant Dirac kernel —— **)**:**

Show that a nonrandomized decision rule $\rho(x, \bullet) = \delta_{\zeta(x)}(\bullet)$ satisfies (3.3) if and only if ζ is a **G-equivariant function**,

$$\zeta(gx) = g\zeta(x)\ , \quad g \in \mathbf{G}\ ,\ x \in \mathbf{X}\ .$$

Conditional risk in invariant decision problems

We'll first generalize the constant-risk theorem from Section 1.

Theorem 3.2. *Suppose we have an invariant decision problem as above. For any decision rule ρ,*

$$R(\theta, \rho) = \int_{\mathbf{X}}\int_{\mathbf{A}} L(g\theta, a)\ g_{\#}\rho(g^{-1}x, da)\ K_{X|\Theta}(g\theta, dx) \quad = R(g\theta, g_{\#}\rho(g^{-1}(\bullet)))\ , \quad \theta \in \mathbf{\Theta}\ ,\ g \in \mathbf{G}\ .$$

In particular, if ρ is **G***-equivariant then*

$$R(\theta, \rho) = R(g\theta, \rho)\ , \quad \theta \in \mathbf{\Theta}\ ,\ g \in \mathbf{G}\ .$$

If $\mathbf{G}$ acts transitively on $\mathbf{\Theta}$ then the risk of any equivariant decision rule is constant; otherwise, it is constant on each orbit of $\mathbf{\Theta}$.

Proof. This is just following the definitions above:

$$\begin{aligned}
R(\theta, \rho) &= \int_{\mathbf{X}} \int_{\mathbf{A}} L(\theta, a)\ \rho(x, da)\ K_{X|\Theta}(\theta, dx) \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} L(g\theta, ga)\ \rho(x, da)\ K_{X|\Theta}(\theta, dx) && \text{(invariant loss)} \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} L(g\theta, a)\ g_{\#}\rho(x, da)\ K_{X|\Theta}(\theta, dx) && \text{(pushforward of } a \text{ by } g) \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} L(g\theta, a)\ g_{\#}\rho(x, da)\ g_{\#}^{-1} K_{X|\Theta}(g\theta, dx) && \text{(equivariance of } K_{X|\Theta}) \\
&= \int_{\mathbf{X}} \int_{\mathbf{A}} L(g\theta, a)\ g_{\#}\rho(g^{-1}x, da)\ K_{X|\Theta}(g\theta, dx) && \text{(pushforward of } x \text{ by } g^{-1}).
\end{aligned}$$

Then if ρ is $\mathbf{G}$-equivariant then $g_{\#}\rho(g^{-1}x, da) = \rho(x, da)$, which yields $R(\theta, \rho) = R(g\theta, \rho)$. □

We'll build on this below.

3.2 Regularity/simplifying assumptions and some consequences thereof

For ease of exposition, notation, and a reduction of technical details, we'll assume that $\mathbf{G}$ acts *exactly transitively* on $\mathbf{\Theta}$. Recall that transitivity means that $\mathbf{\Theta}$ has one orbit: given any θ_1, θ_2, there is some $g \in \mathbf{G}$ such that $g\theta_1 = \theta_2$. Exact transitivity means that g is unique; thus we have that the stabilizer subgroup $\mathbf{G}_\theta = \{\mathrm{e}\}$ for each $\theta \in \mathbf{\Theta}$. We'll discuss the ramifications of this assumption after developing the results.

We will also assume that $\mathbf{G}$ acts on *each orbit* of $\mathbf{X}$ exactly transitively. This can usually be relaxed easily if the points of $\mathbf{X}$ for which $\mathbf{G}_x \neq \{\mathrm{e}\}$ have measure zero; then we can just remove them and consider our sample space to be $\mathbf{X} \setminus \{x\colon \mathbf{G}_x \neq \{\mathrm{e}\}\}$. Stronger violations of the exact transitivity assumption require more sophisticated methods.

All group actions are assumed to be **proper**.

Finally, we will make something like a *measurable factorization* assumption about $\mathbf{G}$ acting on $\mathbf{X}$:

- There exists a measurable **orbit selector** $\alpha\colon \mathbf{X} \to \mathbf{X}$ that maps any point x to a representative point of its orbit, so that $\alpha(x) = \alpha(x')$ if and only if $gx = x'$ for some $g \in \mathbf{G}$ (i.e., $x \sim x'$). We'll denote the **orbit representative** of x as $x_\alpha := \alpha(x)$.
- There is a measurable function $\tau\colon \mathbf{X} \to \mathbf{G}$ that inverts the orbit selector: $\tau(x)\alpha(x) = \tau(x)x_\alpha = x$, each $x \in \mathbf{X}$. We write τ_x as shorthand for the element of $\mathbf{G}$, and τ_x^{-1} for its inverse in $\mathbf{G}$. For lack of a better name, we'll call τ the **orbit actor**. (Suggestions for better names are welcome!)

The orbit selector α is an example of a maximal invariant.

Exercise 13 (Orbit selector as maximal invariant —— *)**:**

Show that the orbit selector as defined above is a maximal invariant.

Exercise 14 (Orbit actor is equivariant when group action is exact —— ***)**:**

Let x_α be the representative of an orbit in $\mathbf{X}$. Show that in general, even if the group action is not exact on the orbit, $(g\tau_x)^{-1}\tau_{gx}$ belongs to $\mathbf{G}_{x_\alpha}$, for each $g \in \mathbf{G}$ and $x \sim x_\alpha$. Conclude that if the group action is exact on the orbit then τ is a $\mathbf{G}$-equivariant function:

$$\tau_{gx} = g\tau_x \ , \quad g \in \mathbf{G} \ , \ x \sim x_\alpha \ . \tag{3.5}$$

Show that in that case, τ^{-1} (which maps x to the inverse group element of τ) is also $\mathbf{G}$-equivariant, in the sense that $\tau_{gx}^{-1} = \tau_x^{-1}g^{-1}$.

Exercise 15 (Equivariant function $\mathbf{X} \to \mathbf{G}$ is all you need —— ***)**:**

Show that if $\tilde{\tau}\colon \mathbf{X} \to \mathbf{G}$ is *some* $\mathbf{G}$-equivariant function then:

i) The function $m\colon \mathbf{X} \to \mathbf{X}$ defined by $m(x) := \tilde{\tau}_x^{-1}x$ is a maximal invariant.

ii) A function $h\colon \mathbf{X} \to \mathbf{X}$ is $\mathbf{G}$-equivariant if and only if there is some $\mathbf{G}$-invariant function $v\colon \mathbf{X} \to \mathbf{X}$ such that $h(x) = \tilde{\tau}_x v(x)$, $x \in \mathbf{X}$.

With these functions in hand, each point of $\mathbf{X}$ can be written as $t(x) := (\tau_x, x_\alpha)$. Under the exact transitivity assumption, this is unique; t is bimeasurable and one-to-one. So, we'll write $\mathbf{X}_\alpha$ to denote an orbit, using α to index orbits.

We've assumed that $\mathbf{G}$ is lcscH; denote the left Haar measure by λ_ℓ and corresponding right Haar measure λ_r. Fixing an orbit $\mathbf{X}_\alpha$, we can induce an invariant measure on the orbit, as follows. Let $\beta_\alpha\colon \mathbf{G} \to \mathbf{X}_\alpha$ be the function defined by

$$\beta_\alpha(g) = gx_\alpha \ . \tag{3.6}$$

Since the orbit selector $\alpha(\,\bullet\,)$ is assumed to be measurable, and the group action is measurable, then so is β_α (it's a particular section of the group action).

Recall from Theorem 2.2 that, viewing $\mathbf{X}_\alpha$ as a set on which $\mathbf{G}$ acts transitively and properly, the measure on $\mathbf{X}_\alpha$ defined by $\lambda_r^\alpha := \beta_{\alpha\#}\lambda_r$ is right-invariant, in the sense that $g_\#\lambda_r^\alpha = \Delta(g)\lambda_r^\alpha$. It also does not depend on the choice of orbit representative x_α.

Now, we can use the orbit selector to disintegrate $K_{X|\Theta}(\theta, dx)$ into $(\alpha_\# K_{\alpha|\Theta})(\theta, dx_\alpha)K_{X|\alpha,\Theta}(\theta, x_\alpha, dx)$. However, we know that since α is a maximal invariant and there is only one orbit in $\mathbf{\Theta}$, $\alpha(X) \perp\!\!\!\perp \Theta$, so that $(\alpha_\# K_{\alpha|\Theta})(\theta, dx_\alpha) =: P_\alpha(dx_\alpha)$ for all θ. That is, for integrable f,

$$\int_{\mathbf{X}} K_{X|\Theta}(\theta, dx)f(x, \alpha(x)) = \int_{\alpha(\mathbf{X})} P_\alpha(dx_\alpha) \int_{\mathbf{X}_\alpha} K_{X|\alpha,\Theta}(\theta, x_\alpha, dx)f(x, x_\alpha) \ . \tag{3.7}$$

In words, this says that an expectation with respect to $K_{X|\Theta}(\theta, dx)$ (left-hand side) can be performed orbit-by-orbit (right-hand side).

Exercise 16 (Orbit selector is ancillary —— ***)**:**

Show that $\alpha_{\#}K_{X|\Theta}$ does not depend on θ. (Show this directly.)

Exercise 17 (Conditional kernel is equivariant —— ****)**:**

Show that $K_{X|\alpha,\Theta}$ inherits the $\mathbf{G}$-equivariance of $K_{X|\Theta}$. That is, for each $g \in \mathbf{G}$, $\theta \in \mathbf{\Theta}$, and P_α-almost all $x_\alpha \in \alpha(\mathbf{X})$ (the exceptional null set not depending on g),

$$K_{X|\alpha,\Theta}(g\theta, x_\alpha, \bullet) = g_{\#}K_{X|\alpha,\Theta}(\theta, x_\alpha, \bullet)\ . \tag{3.8}$$

Finally, assume that $K_{X|\alpha,\Theta}(\theta, x_\alpha, dx)$ is absolutely continuous with respect to λ_r^α, for each $\theta \in \mathbf{\Theta}, x_\alpha \in \alpha(\mathbf{X})$, with density $k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta)$.

Lemma 3.3. *Assume the assumptions above. For every $g \in \mathbf{G}$ and $\theta \in \mathbf{\Theta}$, and P_α-almost all $x_\alpha \in \alpha(\mathbf{X})$,*

$$k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) = k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g)^{-1}\ , \quad \lambda_r^\alpha\text{-a.e.} \tag{3.9}$$

Proof. Let $f\colon \mathbf{X}_\alpha \to \mathbb{R}_+$ be any positive measurable function. Then

$$\begin{aligned}
\int_{\mathbf{X}_\alpha} f(x)k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta)\lambda_r^\alpha(dx) &= \int_{\mathbf{X}_\alpha} f(x)K_{X|\alpha,\Theta}(\theta, x_\alpha, dx) \\
&= \int_{\mathbf{X}_\alpha} f(x)g_{\#}^{-1}K_{X|\alpha,\Theta}(g\theta, x_\alpha, dx) \\
&= \int_{\mathbf{X}_\alpha} f(g^{-1}x)k_{X|\alpha,\Theta}(x \mid x_\alpha, g\theta)\lambda_r^\alpha(dx) \\
&= \int_{\mathbf{X}_\alpha} f(x)k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)(g_{\#}^{-1}\lambda_r^\alpha)(dx) \\
&= \int_{\mathbf{X}_\alpha} f(x)k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g^{-1})\lambda_r^\alpha(dx)\ .
\end{aligned}$$

By the a.e.-uniqueness of densities (i.e., Radon–Nikodym derivatives), (3.9) follows. □

For technical reasons (they will be apparent to the astute observer below), we'll need to assume that there is a version of the density $k_{X|\alpha,\Theta}$ that satisfies (3.9) for *all* $x \in \mathbf{X}$ (rather than λ_r^α-a.e.). It is possible to show that the subset of $\mathbf{X} \times \mathbf{\Theta}$ on which (3.9) is $\mathbf{G}$-invariant, i.e., $gB^* = B^*$ for all $g \in \mathbf{G}$ and $B^* := \{(x,\theta)\colon k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) = k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g)^{-1}\}$.

Proposition 3.4. *The following sets are $\mathbf{G}$-invariant:*

- $B^* := \{(x,\theta)\colon k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) = k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g)^{-1}$ *for all* $g \in \mathbf{G}\}$*; and*
- $B^{**} := \{x\colon k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) = k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g)^{-1}$ *for all* $g \in \mathbf{G},\ \theta \in \mathbf{\Theta}\}$.

The upshot is that if (3.9) holds for some $x \in \mathbf{X}$ then it holds for all $x' \in \mathbf{G}x$, i.e., over the entire orbit of x. So if we assume that (3.9) holds for each $x_\alpha \in \alpha(\mathbf{X})$ then it holds for each $x \in \mathbf{X}$.

3.3 Statistical properties of G-equivariant decision rules

Set an arbitrary $\theta_0 \in \mathbf{\Theta}$, and define the counterpart of β on $\mathbf{\Theta}$ as

$$\gamma(g) = g\theta_0 \ , \quad g \in \mathbf{G} \ . \tag{3.10}$$

We'll use this to define a prior on $\mathbf{\Theta}$, $\pi := \gamma_\# \lambda_r$. Again by Theorem 2.2, this is right-invariant and does not depend on the choice of θ_0. We will also assume the existence of the $\mathbf{\Theta}$-counterparts of α and τ. Since $\mathbf{\Theta}$ is assumed to have exactly one orbit, we only need the counterpart of τ, denoted $\eta\colon \mathbf{\Theta} \to \mathbf{G}$, so that $\theta = \eta_\theta \theta_0 = (\eta_\theta, \theta_0)$.

For simplicity, we'll study nonrandomized decision rules; the extension to randomized decision rules is straightforward.

Theorem 3.5. *Assume the assumptions above. Let ρ be a nonrandomized $\mathbf{G}$-equivariant decision rule, and set the prior on Θ to be $\pi := \gamma_\# \lambda_r$. Then the risk at θ, conditional on $\alpha(X) = x_\alpha$, is constant in θ, and is equal to the posterior risk of ρ when any $X = x \sim x_\alpha$ is observed, which is constant on the orbit $\mathbf{G}x_\alpha$.*

Proof. By the equivariance of $K_{X|\alpha,\Theta}$ and of ρ, it is easy to show that the conditional risk is constant in θ (as in Theorem 3.2). So we can take θ_0, and

$$\begin{aligned}
&R(\theta_0,\rho \mid \alpha(X) = x_\alpha) \\
&\quad = \int_{\mathbf{X}_\alpha} L(\theta_0, \rho(x)) K_{X|\alpha,\Theta}(\theta_0, x_\alpha, dx) && \text{(definition)} \\
&\quad = \int_{\mathbf{X}_\alpha} L(\theta_0, \rho(x)) k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta_0) \lambda_r^\alpha(dx) && \text{(density)} \\
&\quad = \int_{\mathbf{G}} L(\theta_0, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, \theta_0) \lambda_r(dg) && \text{(definition of } \lambda_r^\alpha) \\
&\quad = \int_{\mathbf{G}} L(g^{-1}\theta_0, \rho(x_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, \theta_0) \lambda_r(dg) && \text{(equivariance of } \rho\text{; invariance of } L) \\
&\quad = \int_{\mathbf{G}} L(g^{-1}\theta_0, \rho(x_\alpha)) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, g^{-1}\theta_0) \Delta(g) \lambda_r(dg) && \text{(Lemma 3.3)} \\
&\quad = \int_{\mathbf{G}} L(g\theta_0, \rho(x_\alpha)) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, g\theta_0) \lambda_r(dg) && (\lambda_\ell(dg^{-1}) = \lambda_r(dg),\ \Delta(g^{-1})\lambda_\ell(dg) = \lambda_r(dg)) \\
&\quad = \int_{\mathbf{\Theta}} L(\theta, \rho(x_\alpha)) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta) \pi(d\theta) \ .
\end{aligned}$$

Note that the last line contains what we hope is the posterior distribution,

$$K_{\Theta|X,\alpha}(x, x_\alpha, d\theta) = k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) \pi(d\theta) \ . \tag{3.11}$$

We'll need to check that it is. We'll do that in the exercise below.

Then the last line above represents the posterior risk of ρ when we observe $X = x_\alpha$. We still need to prove the claim that the posterior risk is constant on each orbit $\mathbf{X}_\alpha$. Picking up from the last

line above, we have that

$$\begin{aligned}
&\int_{\mathbf{\Theta}} L(\theta, \rho(x_\alpha)) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\pi(d\theta) \\
&\quad = \int_{\mathbf{\Theta}} L(g\theta, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\pi(d\theta) && \text{(equivariance of } \rho\text{, invariance of } L) \\
&\quad = \int_{\mathbf{\Theta}} L(g\theta, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, g\theta)\Delta(g)^{-1}\pi(d\theta) && \text{(Lemma 3.3)} \\
&\quad = \int_{\mathbf{\Theta}} L(\theta, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, \theta)\Delta(g)^{-1}(g_{\#}\pi)(d\theta) \\
&\quad = \int_{\mathbf{\Theta}} L(\theta, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, \theta)\Delta(g)^{-1}\Delta(g)\pi(d\theta) && \text{Theorem 2.2 applied to } \pi \\
&\quad = \int_{\mathbf{\Theta}} L(\theta, \rho(gx_\alpha)) k_{X|\alpha,\Theta}(gx_\alpha \mid x_\alpha, \theta)\pi(d\theta) && .
\end{aligned}$$

Here, g is arbitrary so the equality holds for all $x \sim x_\alpha$. □

Exercise 18 ("Posterior" indeed is a posterior distribution —— ***)**:**

Show that $K_{\Theta|X,\alpha}(x, x_\alpha, d\theta)$ as defined above in (3.11) is the posterior of Θ given $X = x$.

Note that conditioning on $(X, \alpha(X))$ is the same as conditioning only on X, and since $\alpha(X) \perp\!\!\!\perp \Theta$, the marginal distribution of $\alpha(X)$ doesn't appear on the right-hand side of (3.11).

The equalities above can be summarized by

$$\begin{aligned}
E_\theta[L(\theta, \rho(X)) \mid \alpha(X) = x_\alpha] &= \int_{\mathbf{X}_\alpha} L(\theta, \rho(x)) K_{X|\alpha,\Theta}(\theta, x_\alpha, dx) \qquad (3.12) \\
&= \int_{\mathbf{\Theta}} L(\theta, \rho(x_\alpha)) K_{\Theta|X,\alpha}(x, x_\alpha, d\theta) \\
&= E[L(\Theta, \rho(X)) \mid X = x_\alpha] = r(\rho; x_\alpha) \\
&= r(\rho; gx_\alpha)\ , \quad g \in \mathbf{G}\ .
\end{aligned}$$

Pivotal quantity

Suppose we set the decision/action space to be $\mathbf{A} = \mathbf{G}$. For $g \in \mathbf{G}$, $B \in \mathcal{B}(\mathbf{G})$, set $L(\theta, g) = \mathbf{1}_B(\eta_\theta^{-1} g)$ then we get the following corollary.

Corollary 3.6. *Define $Q(\theta, x) := \eta_\theta^{-1}\tau_x$. Then the conditional distribution of $Q(\theta, X)$ given $\alpha(X) = a$ is the same as the posterior distribution (with prior π as in Theorem 3.5) of $Q(\Theta, x)$.*

Exercise 19 (Pivotal quantity —— ****)**:**

Prove Corollary 3.6.

This is called a **pivotal quantity** because we can switch between what we treat as random without having to do separate probability calculations (i.e., change the distribution we're working with).

Fisher called the common distribution the "fiducial" distribution, and group-invariant decision theory setup is one of the only known cases in which Fisher's desired fiducial statistics actually works. This line of thought was developed by Fraser [Fra61] and Hora and Buehler [HB66], and others.

MRE decision rules

Theorem 3.7. *Assume that the formal Bayes rule with respect to the prior $\pi = \gamma_{\#}\lambda_r$ exists. Let $\rho_\alpha(x_\alpha)$ be the decision rule that minimizes the posterior risk if $X = x_\alpha$ is observed. That is,*

$$\rho_\alpha(x_\alpha) = \operatorname*{arg\,min}_{a\in\mathbf{A}} \int_{\mathbf{\Theta}} L(\theta, a) K_{\Theta|X}(x_\alpha, d\theta) = \operatorname*{arg\,min}_{a\in\mathbf{A}} \int_{\mathbf{\Theta}} L(\theta, a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\pi(d\theta) \,,$$

for each $x_\alpha \in \alpha(\mathbf{X})$. Then the equivariant decision rule $\rho^*(x) := \tau_x \rho_\alpha(\alpha(x))$ is the formal Bayes rule, minimum risk equivariant (MRE) and MRE conditional on $\alpha(X) = x_\alpha$.

Proof. First, observe that ρ^* is $\mathbf{G}$-equivariant: $\rho^*(gx) = \tau_{gx}\rho_\alpha(\alpha(gx)) = g\tau_x\rho_\alpha(\alpha(x)) = g\rho^*(x)$. (This is just a special case of Exercise 15.)

To show that it is the formal Bayes rule, the posterior risk at $a \in \mathbf{A}$ when x is observed is

$$\begin{aligned}
\int_{\mathbf{\Theta}} L(\theta, a) k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta)\pi(d\theta) &= \int_{\mathbf{\Theta}} L(\tau_x^{-1}\theta, \tau_x^{-1}a) k_{X|\alpha,\Theta}(\tau_x x_\alpha \mid x_\alpha, \theta)\pi(d\theta) && \text{(invariance of } L\text{)} \\
&= \int_{\mathbf{\Theta}} L(\tau_x^{-1}\theta, \tau_x^{-1}a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \tau_x^{-1}\theta)\Delta(\tau_x)\pi(d\theta) && \text{(Lemma 3.3)} \\
&= \int_{\mathbf{\Theta}} L(\theta, \tau_x^{-1}a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\Delta(\tau_x)(\tau_{x\#}^{-1}\pi)(d\theta) \\
&= \int_{\mathbf{\Theta}} L(\theta, \tau_x^{-1}a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\pi(d\theta) \,.
\end{aligned}$$

This is just the posterior risk of the decision rule $\tau_x^{-1}a$ when x_α is observed, which by assumption is minimized for $\tau_x^{-1}a = \rho_\alpha(x_\alpha)$. Rearranging yields $\rho^*(x) = \tau_x\rho_\alpha(x_\alpha)$. This is true for each $x \in \mathbf{X}$, yielding the decision rule ρ^*.

Now, Theorem 3.5 showed that the posterior risk of an equivariant decision rule is constant within any orbit $\mathbf{X}_\alpha$, so if an equivariant decision rule minimizes the posterior risk at x_α then it does so at all $x \sim x_\alpha$. The same theorem shows that the posterior risk is equal to the conditional risk given $\alpha(X) = x_\alpha$, so the conditional risk is minimized (among $\mathbf{G}$-equivariant decision rules) by ρ^*, which makes it the MRE conditional on $\alpha(X) = x_\alpha$.

Since ρ^* minimizes the risk uniformly in x_α, we see that

$$R(\theta_0, \rho) = \int_{\alpha(\mathbf{X})} P_\alpha(dx_\alpha) R(\theta_0, \rho \mid x_\alpha)$$

is minimized at ρ^*. So it is the MRE decision rule. □

3.4 An examples

The following example is taken from [Eat89].

Example: Underparameterized location-scale normal [Eat89, Ex. 6.1]

Consider i.i.d. $X_1, \ldots, X_n$ with distribution $\mathcal{N}(\theta, \theta^2)$, $\theta > 0$. The random vector $X = (X_1, \ldots, X_n)$ in $\mathbb{R}^n$ has distribution $\mathcal{N}(\theta\mathbf{1}, \theta^2\mathbf{I})$. With $\mathbf{\Theta} = \mathbf{A} = (0, \infty)$, the loss function

$$L(\theta, a) = \frac{(\theta - a)^2}{\theta} ,$$

is invariant under the multiplicative group $\mathbf{G} = (\mathbb{R}_+, \times)$. With $\mathbf{G}$ acting on $\mathbb{R}^n$ by coordinate-wise multiplication, the decision problem is invariant.

Exercise 20 (Verifying invariance of decision problem —— *)**:**

Check that the decision problem as specified is invariant.

The action of $\mathbf{G}$ on $\mathbb{R}^n$ is not proper (why not?) but it is proper on $\mathbb{R}^n \setminus \{0\mathbf{1}\}$. Here, $\mathbf{\Theta} = \mathbf{G}$, so specifying the prior π is the same as finding a right Haar measure λ_r on $\mathbf{G}$.

Exercise 21 (Right Haar measure on the multiplicative group —— *)**:**

Show that if dg is the Lebesgue measure on $\mathbb{R}_+$ then $\frac{dg}{g}$ is a right Haar measure on $\mathbf{G} = (\mathbb{R}_+, \times)$.

Thus, we take our improper prior $\pi(d\theta) = \frac{d\theta}{\theta}$, where $d\theta$ is Lebesgue measure. If $k_{X|\Theta}$ is the density of $K_{X|\Theta}$ with respect to Lebesgue measure then the posterior density (again w.r.t. Lebesgue measure) is

$$k_{\Theta|X}(\theta \mid x)d\theta = \frac{\theta^{-1}k_{X|\Theta}(x \mid \theta)}{\int_0^\infty \theta^{-1}k_{X|\Theta}(x \mid \theta)d\theta}d\theta .$$

Finding the MRE estimator then boils down to solving for each $x \in \mathbf{X}$

$$\arg\min_{a \in \mathbb{R}_+} \int_{\mathbb{R}_+} L(\theta, a)k_{\Theta|X}(\theta \mid x)d\theta .$$

For the loss function above, this is solved by

$$\rho^*(x) = \frac{E[\Theta^{-1} \mid X = x]}{E[\Theta^{-2} \mid X = x]} ,$$

which can be estimated numerically.

Eaton [Eat89] includes some other examples in Ch. 6 to demonstrate the method. They're worth reading through—not necessarily for the specific problems, but for the techniques used to set up and solve the problem.

3.5 Some thoughts on the big picture

We'll end our study of the classical literature on equivariant estimation here. I'll end with few pointers to related literature and some thoughts on the applicability of the theory.

Equivariant prediction

The problem of equivariant prediction, say predicting a random element $Y \in \mathbf{Y}$ from an observation X, is structurally very similar to equivariant estimation. In the case of prediction, we would augment the decision problem presented at the beginning of this section to include $\mathbf{G}$-equivariant conditional distributions of $Y|X$, i.e.,

$$K_{Y|X}(gx, dy) = g_{\#}K_{Y|X}(x, dy) \,, \quad g \in \mathbf{G},\ x \in \mathbf{X} \,, \tag{3.13}$$

assuming, of course, that $\mathbf{G}$ acts on $\mathbf{Y}$. Our loss function would also need to be modified as

$$L(g\theta, gy, ga) = L(\theta, y, a) \,, \quad g \in \mathbf{G},\ y \in \mathbf{Y},\ a \in \mathbf{A} \,.$$

In "pure prediction" problems, our loss would not depend on θ, and just compare y with our prediction $\hat{y}$, i.e., $L(y, \hat{y})$. Eaton and Sudderth have studied these types of problems, establishing analogous results to the ones for equivariant estimation in a series of papers, e.g., [ES99; ES01; ES04].

Applicability/limitations

The theory in this section is presented in the literature as being generally applicable. Although we assumed that we had a group $\mathbf{G}$ acting on each of $\mathbf{X}, \mathbf{\Theta}, \mathbf{A}$, that was mostly for convenience of presentation. We really only need a group acting on $\mathbf{X}$, and under certain conditions we can *induce* actions on $\mathbf{\Theta}$ and $\mathbf{A}$. The following is Theorem 2.6 of [Eat89].

Theorem 3.8. *Consider a group $\mathbf{G}$ acting on $\mathbf{Y}$ and suppose that F is a function defined on $\mathbf{Y} \times \mathbf{Z}$, for some sets $\mathbf{Y}, \mathbf{Z}$. Assume that for each $g \in \mathbf{G}$ and each $z \in \mathbf{Z}$, there exists a unique $z' \in \mathbf{Z}$ such that*

$$F(gy, z') = F(y, z) \,, \quad y \in \mathbf{Y} \,.$$

Then $\mathbf{G}$ acts on $\mathbf{Z}$ through the defined group action $gz := z'$. With this group action,

$$F(gy, gz) = F(x, z) \,, \quad g \in \mathbf{G},\ y \in \mathbf{Y},\ z \in \mathbf{Z} \,,$$

so F is $\mathbf{G}$-invariant.

Starting with a group $\mathbf{G}$ acting on the sample space $\mathbf{X}$, we can apply this to our model family by considering the Markov kernel $K_{X|\Theta}(\theta, B)$, $B \in \mathcal{B}(\mathbf{X})$, as a function $\mathbf{\Theta} \times \mathcal{B}(\mathbf{X}) \to [0, 1]$. Then the action of $\mathbf{G}$ on $\mathbf{\Theta}$ is defined by $g\theta := \theta'$, where θ' is the unique element of $\mathbf{\Theta}$ (if it exists) such that

$$K_{X|\Theta}(\theta', B) = K_{X|\Theta}(\theta, g^{-1}B) \,, \quad g \in \mathbf{G},\ B \in \mathcal{B}(\mathbf{X}) \,.$$

See [Eat89, Example 2.20] for more details. If we have a $\mathbf{G}$ action on $\mathbf{\Theta}$ (possibly induced) then we can also apply this to our loss function $L \colon \mathbf{X} \times \mathbf{A} \to \mathbb{R}_+$ to induce an action on $\mathbf{A}$, and we know that our loss function is invariant by construction of the action on $\mathbf{A}$. Schervish [Sch95] also addresses both of these cases separately in Lemmas 6.28 and 6.31.

In short, if we have a group acting on the sample space, we can just induce an action on $\mathbf{\Theta}, \mathbf{A}$ and be in the equivariant setting. We could also just construct our model so that we're in the equivariant setting; this approach was advocated for by Fraser [Fra61; Fra67] under the name

"structural inference." As a somewhat modern example, we could just use $\mathbf{G}$ as our parameter space, specify some simple "base distribution" like $\mathcal{N}(0, \mathbf{I})$ as $K_{X|G}(\mathrm{e}, dx)$, and construct the model family as $K_{X|G}(g, dx) := g_{\#}K_{X|G}(e, dx)$. This looks like a particular kind of normalizing flow.

So this seems pretty general, right? Why isn't it considered to be on equal footing with frequentist and Bayesian approaches to inference? I haven't come across anything in the classical literature that addresses this, other than saying that the approach causes philosophical discomfort. There are also good unresolved questions about how to choose the group $\mathbf{G}$, because it must be fixed from the start. But that hasn't stopped, for example, frequentists from adopting Bayesian methods/criteria (and vice versa) when the performance is better.

Here are some of my thoughts on the matter. The concerns about how to choose $\mathbf{G}$ are important, though we can reason that as a rule of thumb, we'd like $\mathbf{G}$ to be large enough so that its action on $\mathbf{\Theta}$ is transitive (otherwise the best we can do is "slice" up the parameter space, which is a simplification but still presents difficulties; see [Wes59]), but small enough that its action on each of $\mathbf{X}, \mathbf{A}, \mathbf{\Theta}$ is proper (or satisfies some other regularity condition).

Besides the problem of choosing $\mathbf{G}$, there's a very difficult problem smuggled into the generality statements: determining an induced action is not trivial, and in many cases may defeat the point of using a group to simplify the problem. Consider operationalizing Theorem 3.8 in order to induce an action on $\mathbf{A}$ through L. Each time a new (g, a) pair is encountered, one would be required to solve some sufficiently large system of equations $L(g\theta, a') = L(\theta, a)$ to determine which $a' = ga$, and it would have to be consistent with all previously computed actions. We might pre-compute this, in which case we might think it of as a "lookup table action".[6] In some nice cases this might work out, but it seems prohibitive to carry out computationally. Inducing an action on $\mathbf{\Theta}$ is equally, if not more, difficult; using a family of densities $k_{X|\Theta,\alpha} \ll \lambda_r^\alpha$ requires the same approach via Theorem 3.8 and defining $g\theta = \theta'$ for θ' satisfying

$$k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta') = k_{X|\alpha,\Theta}(gx \mid x_\alpha, g\theta)\Delta(g)^{-1} , \quad g \in \mathbf{G},\ x \in \mathbf{X} .$$

To pull on this thread a little more, consider the following approaches to putting the theory into practice, in order to find an MRE estimator.

Approach 1: Using τ_x

Assume we know how $\mathbf{G}$ acts on each of $\mathbf{X}, \mathbf{\Theta}, \mathbf{A}$, and we have a $\mathbf{G}$-equivariant function $\tau\colon \mathbf{X} \to \mathbf{G}$. Then we can find ρ_α for each $x_\alpha \in \alpha(X)$ and set $\rho^*(x) = \tau_x \rho_\alpha(\alpha(x))$, as in Theorem 3.7. Operationally, this probably involves approximating the posterior in order to estimate the posterior risk,

$$\int_{\mathbf{\Theta}} L(\theta, a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, \theta)\pi(d\theta) = \int_{\mathbf{G}} L(g\theta_0, a) k_{X|\alpha,\Theta}(x_\alpha \mid x_\alpha, g\theta_0)\lambda_r(dg) .$$

This could be performed using, e.g., MCMC or variational inference, or, if $\mathbf{G}$ is compact, estimating the integral on the right-hand side by sampling group elements and applying them to θ_0 (which, recall, is arbitrary).

[6] Thanks to Geoffrey Woollard for the name.

This approach provides substantial benefits in terms of simplifying the problem, but requires a problem setup that is "just right" in terms of specifying the required group actions, and seems less flexible/general than the following two approaches.

Approach 2: Posterior risk minimization

As discussed by Eaton [Eat89, pp. 85–87], we can minimize the posterior risk with respect to $a \in \mathbf{A}$ for each $x \in \mathbf{X}$, i.e.,

$$\rho^*(x) := \operatorname*{arg\,min}_{a \in \mathbf{A}} \int_{\mathbf{\Theta}} L(\theta, a) k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) \pi(d\theta) \; . \tag{3.14}$$

This approach does not require knowing how $\mathbf{G}$ acts on $\mathbf{A}$; it is induced through the minimization problem. In some special cases we may be able to solve the minimization problem analytically, though in general that will not be true. Without further structure, we actually have to carry out this *for every* $x \in \mathbf{X}$ that we're interested in. We could, of course, do this lazily, i.e., only carry it out when we observe an x for which we need an action, but this seems to greatly diminish the biggest utility of using groups: compressing the problem complexity through the algebraic rules of the group.

If we do know the action of $\mathbf{G}$ on $\mathbf{A}$ then this approach is equivalent to Approach 1 based on τ_x.

Approach 3: Conditional risk minimization

Somewhat less common in the literature as a general approach is finding ρ^* by minimizing the conditional risk over the class of $\mathbf{G}$-equivariant functions, $\mathcal{E}_{\mathbf{G}}(\mathbf{X}, \mathbf{A})$,

$$\begin{aligned} \rho^* &= \operatorname*{arg\,min}_{\rho \in \mathcal{E}_{\mathbf{G}}(\mathbf{X},\mathbf{A})} \int_{\mathbf{X}_\alpha} L(\theta, \rho(x)) k_{X|\alpha,\Theta}(x \mid x_\alpha, \theta) \lambda_r^\alpha(dx) \\ &= \operatorname*{arg\,min}_{\rho \in \mathcal{E}_{\mathbf{G}}(\mathbf{X},\mathbf{A})} \int_{\mathbf{G}} L(\theta, \rho(g\alpha(x))) k_{X|\alpha,\Theta}(g\alpha(x) \mid x_\alpha, \theta) \lambda_r(dg) \; . \end{aligned}$$

This is the approach of most examples I've come across. The main sticking point for implementing it in practice is that in order to determine $\mathcal{E}_{\mathbf{G}}(\mathbf{X}, \mathbf{A})$, we need to know the action (or a set of candidate actions) of $\mathbf{G}$ on $\mathbf{A}$. Classically, even when we do know that, if $\mathcal{E}_{\mathbf{G}}(\mathbf{X}, \mathbf{A})$ is "too big" then optimizing over it analytically or computationally was not realistic, and was prohibitive for many classical examples. However, with the advent of fairly general $\mathbf{G}$-equivariant convolutional neural networks (CNNs), it seems less daunting a challenge. I have not come across this in the recent literature. (Research project?)

The application to prediction problems in which the conditional distribution of $Y|X$ is assumed to be $\mathbf{G}$-invariant, i.e., $K_{Y|X}(gx, dy) = K_{Y|X}(x, dy)$, $g \in \mathbf{G}$, rather than equivariant, does seem to be fairly widespread, albeit under a different name. In that case, if we have a $\mathbf{G}$-invariant predictor

$\rho\colon \mathbf{X} \to \mathbf{Y}$, then

$$\begin{aligned} R(\theta, \rho) &:= \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(x)) K_{Y|X}(x, dy) K_{X|\Theta}(\theta, dx) \\ &= \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(g^{-1}x)) K_{Y|X}(g^{-1}x, dy) K_{X|\Theta}(g\theta, dx) \\ &= \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(x)) K_{Y|X}(x, dy) K_{X|\Theta}(g\theta, dx) \\ &= R(g\theta, \rho) \ . \end{aligned}$$

In typical machine learning practice, we would not specify a parameterized set of Markov kernels $K_{X|\Theta}$, but instead assume some distribution P on $\mathbf{X}$ and then use $(g_{\#}P)_{g\in\mathbf{G}}$ as our model family, along with some $\mathbf{G}$-invariant Markov kernel $K_{Y|X}$, to get a family of joint distributions, $\{g_{\#}P \otimes K_{Y|X} \colon g \in \mathbf{G}\}$. In this case, we can take $\mathbf{G}$ as the parameter space.

If we don't start with a $\mathbf{G}$-invariant predictor but would like to find one from a more general (including non-invariant ones) class of predictors, then **data augmentation** minimizes the augmented risk,

$$\begin{aligned} \int_{\mathbf{G}} R(g, \rho) \lambda_r(dg) &= \int_{\mathbf{G}} \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(x)) K_{Y|X}(x, dy) (g_{\#}P)(dx) \lambda_r(dg) \\ &= \int_{\mathbf{G}} \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(gx)) K_{Y|X}(gx, dy) P(dx) \lambda_r(dg) \\ &= \int_{\mathbf{G}} \int_{\mathbf{X}\times\mathbf{Y}} L(y, \rho(gx)) K_{Y|X}(x, dy) P(dx) \lambda_r(dg) \\ &= \int_{\mathbf{X}\times\mathbf{Y}} \int_{\mathbf{G}} L(y, \rho(gx)) \lambda_r(dg) K_{Y|X}(x, dy) P(dx) \ , \end{aligned}$$

where we used Fubini's theorem to interchange the integrals.

We can use the empirical distribution of an observed i.i.d. dataset $(x_i, y_i)_{i=1}^n$ to approximate the risk with the empirical risk,

$$\hat{R}(\rho) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} L(y_i, \rho(g_j x_i)) \ ,$$

where g_j are chosen to approximate the integral over $\mathbf{G}$. In practice, $\mathbf{G}$ is assumed to be compact, or well approximated by a compact group, so that we can sample i.i.d. group elements $G_j \sim \lambda$.

Note that there's no guarantee that the resulting ρ will be $\mathbf{G}$-invariant, but this procedure "encourages" it to be. A related technique, **feature averaging** averages ρ over $\mathbf{G}$, which does produce an invariant function. Both seem to work well in practice, though our theoretical understanding of these techniques is still nascent. For statistical theory related to data augmentation, see [CDL20; HOA22]. The latter, in particular, shows that data augmentation is not always beneficial, something noticed empirically by Lyle et al. [Lyl+20], which also has some (loose) results comparing feature averaging to data augmentation.

4 Some benefits of incorporating symmetry

Supplemental reading: Jacod and Protter [JP04, Ch. 22-23]; Elesedy [Ele21]

In this section, we'll try to develop a basic understanding of why incorporating symmetry into our models can be a good idea—if the symmetry is appropriate for the problem. This won't be comprehensive, but it should give us a decent idea of the general picture. The short explanation is that if a symmetry is appropriate for a problem, then accounting for it (incorporating it into the model) adds beneficial constraints and/or reduces the learning/estimation/inference problem.

One way to show this is to start with something (e.g., a prediction function) that is not symmetric (invariant or equivariant, as appropriate), turn it into a symmetric one, and show that the symmetric one is better. The most convenient method for obtaining a symmetric thing from a not symmetric thing is by taking an appropriate average over the group.

We'll focus on compact groups here. In some cases, the ideas can be extended to locally compact groups but the math gets hairy; in other cases, it's not clear if they can be extended. To understand the results, we need a bit of background on Hilbert spaces.

4.1 Background: Hilbert spaces and projections

We'll cover the basics here; Ch. 22 of Jacod and Protter [JP04] has a nice coverage of Hilbert spaces.

From linear algebra, we're familiar with the following properties of $\mathbb{R}^d$:

- The **inner product** between two vectors, $x, y \in \mathbb{R}^d$, is $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$.
- $\mathbb{R}^d$ is a **linear space**: for $a, b \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$, $ax + by \in \mathbb{R}^d$. Moreover, the inner product is **bilinear**:

$$\langle ax + by, z \rangle = \langle ax, z \rangle + \langle by, z \rangle \ .$$

- $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$, which defines a norm,

$$||x|| := \sqrt{\langle x, x \rangle} \ ,$$

  is just the Euclidean norm. By the Cauchy–Schwarz inequality ($|\langle x, y \rangle| \leq ||x|| ||y||$), we see that this satisfies the triangle inequality,

$$||x + y||^2 = ||x||^2 + 2\langle x, y \rangle + ||y||^2 \leq ||x||^2 + 2||x|| \ ||y|| + ||y||^2 = (||x|| + ||y||)^2 \ ,$$

  so it is a true norm.

- It is **complete**: if a sequence (x_n) is Cauchy under $|| \bullet ||$ then it has a limit in $\mathbb{R}^d$. (A sequence (x_n) is Cauchy if $||x_n - x_m|| \to 0$ as both $m, n \to \infty$.) This basically says that $\mathbb{R}^d$ doesn't have any holes; nothing goes missing when we take limits.

These are the defining properties of a Hilbert space, of which $\mathbb{R}^d$ is a special case.

Definition 4.1. A **Hilbert space** is a complete normed linear space with an inner product $\langle x, x \rangle_{\mathcal{H}}^{1/2} = ||x||_{\mathcal{H}}$, for each $x \in \mathcal{H}$.

Example 4.1 L^2 **is a Hilbert space.** We will be interested in a particular kind of Hilbert space. Let $(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ be a standard Borel measure space, and $L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ the space of square-integrable functions. That is,

$$L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu) := \left\{ f \colon \mathbf{X} \to \mathbb{R} \ : \ \int_{\mathbf{X}} |f(x)|^2 \mu(dx) < \infty \right\} .$$

Since we won't be changing $\mathbf{X}$, we'll just denote this by $L^2(\mu)$. Here, we identify all functions that are equal μ-a.e. as the same element of $L^2(\mu)$. We define the inner product by

$$\langle f, h \rangle_{L^2(\mu)} := \int_{\mathbf{X}} f(x) h(x) \mu(dx) . \tag{4.1}$$

It's straightforward to check that $L^2(\mu)$ is a normed linear space with an inner product defined by (4.1) that defines the norm as $||f||_{L^2(\mu)} = \sqrt{\langle f, f \rangle_{L^2(\mu)}}$. Completeness is harder to show; see Thm. 22.2 in [JP04].

The rest of this subsection will discuss things in terms of a general Hilbert space $\mathcal{H}$, and we'll eventually apply them to $L^2(\mu)$.

Two elements $x, y \in \mathcal{H}$ are **orthogonal** if $\langle x, y \rangle_{\mathcal{H}} = 0$. As with $\mathbb{R}^d$, we can study projections onto subspaces of a Hilbert space. A **subspace** of $\mathcal{H}$ is a subset $\mathcal{L} \subset \mathcal{H}$ that is linear ($x, y \in \mathcal{L}$ implies that $ax + by \in \mathcal{L}$) and closed (if a sequence $(x_n) \in \mathcal{L}$ converges to $x \in \mathcal{H}$ then $x \in \mathcal{L}$). $\mathcal{L}^\perp$ is the set of all elements of $\mathcal{H}$ that are orthogonal to each $y \in \mathcal{L}$.

For a point $x \in \mathcal{H}$ and a subspace $\mathcal{L} \subset \mathcal{H}$, the distance from x to $\mathcal{L}$ is defined as

$$d(x, \mathcal{L}) := \inf_{y \in \mathcal{L}} ||x - y|| .$$

This is always well defined, but it's not obvious that there is a point $y^* \in \mathcal{L}$ such that $d(x, \mathcal{L}) = ||x - y^*||$. In fact, such a point does exist and it is unique. (See Thm. 22.6 of [JP04].)

Definition 4.2. The (orthogonal) **projection** of an element $x \in \mathcal{H}$ onto a subspace $\mathcal{L} \subset \mathcal{H}$ is the unique $y \in \mathcal{L}$ which is closest to x. We denote the **projection operator** $\Pi_{\mathcal{L}}$, expressed as

$$\Pi_{\mathcal{L}} x = \arg\min_{y \in \mathcal{L}} ||x - y|| . \tag{4.2}$$

Note that in general, a "projection" may not be orthogonal (it just satisfies $\Pi^2 = \Pi$). We'll only be concerned with orthogonal projection here, so I'll drop "orthogonal" but keep in mind that it's implicit. Note that orthogonal projection is unique, but there are possibly many non-unique non-orthogonal projections.

The projection operator satisfies some very useful properties. This is a collection of results that can be found in [JP04, Ch. 22].

Theorem 4.3. *The projection operator* Π *of* $\mathcal{H}$ *onto* $\mathcal{L}$ *satisfies the following properties:*

i) Π *is* ***idempotent****:* $\Pi^2 = \Pi$.

ii) $\Pi x = x$ *for* $x \in \mathcal{L}$*, and* $\Pi x = 0$ *for* $x \in \mathcal{L}^\perp$.

iii) For every $x \in \mathcal{H}$, $x - \Pi x$ is orthogonal to every element of $\mathcal{L}$.

iv) Every $x \in \mathcal{H}$ can be written as a unique decomposition $x = \Pi x + (x - \Pi x)$, as the sum of one element of $\mathcal{L}$ and one element of $\mathcal{L}^\perp$. $(x - \Pi x)$ is the projection of x onto $\mathcal{L}^\perp$ and $(\mathcal{L}^\perp)^\perp = \mathcal{L}$.

v) $\langle \Pi x, y \rangle = \langle x, \Pi y \rangle$, and Π is a linear operator: $\Pi(ax + by) = a\Pi x + b\Pi y$.

vi) Πx is the unique element of $\mathcal{L}$ such that $\langle \Pi x, y \rangle = \langle x, y \rangle$ for all $y \in \mathcal{L}$.

vii) If T is an operator that maps $\mathcal{H}$ onto $\mathcal{L}$ such that $x - Tx$ is orthogonal to all of $\mathcal{L}$ for each $x \in \mathcal{H}$, then $T = \Pi$.

Example 4.2. We saw above that $L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ is a Hilbert space. Now suppose we have a sub-σ-algebra $\mathcal{A} \subset \mathcal{B}(\mathbf{X})$. Then $L^2(\mathbf{X}, \mathcal{A}, \mu)$ is also a Hilbert space; it is a (closed Hilbert) subspace of $L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ consisting of all $\mathcal{A}$-measurable and square-integrable functions. When μ is a probability measure (not necessarily $\mathbf{G}$-invariant), we can define conditional expectation this way. For random variables $Y \colon \mathbf{X} \to \mathbb{R}$, $Z \colon \mathbf{X} \to \mathbf{Z}$ defined on the probability space $(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$, $E_\mu[Y \mid \sigma(Z)]$ is the unique element $\hat{Y} \in L^2(\mathbf{X}, \sigma(Z), \mu)$ such that

$$E[\hat{Y}W] = E[YW] \,,$$

for all $W \in L^2(\mathbf{X}, \sigma(Z), \mu)$. Finally, recall that a random variable W is $\sigma(Z)$-measurable if and only if it can be written $W = f(Z)$, for some function f, so $L^2(\mathbf{X}, \sigma(Z), \mu)$ is just the collection of all square-integrable functions of Z. (Square-integrability here is in the sense that $f \circ Z \in L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$.) By Theorem 4.3 vi), we can see that $\hat{Y}$ is just the projection of Y onto $L^2(\mathbf{X}, \sigma(Z), \mu)$. This generalizes to all sub-σ-algebras, not just those generated by a random variable.

4.2 Symmetry through averaging over the group

Recall that a function $f \colon \mathbf{X} \to \mathbf{Y}$ is $\mathbf{G}$-invariant if $f(gx) = f(x)$ for all $g \in \mathbf{G}$ and $x \in \mathbf{X}$. The function is $\mathbf{G}$-equivariant if $f(gx) = gf(x)$ for all $g \in \mathbf{G}$ and $x \in \mathbf{X}$. Note that invariance is a special case of equivariance, in which the group action on $\mathbf{Y}$ is trivial.

Assume that $\mathbf{G}$ is compact. Recall that this implies that Haar measure is left- and right-invariant so that $\lambda_\ell = \lambda_r = \lambda$; that $\lambda(\mathbf{G}) < \infty$; and that $\mathbf{G}$ is unimodular: $\Delta(g) = 1$. So we just denote Haar measure as λ and assume without loss of generality that $\lambda(\mathbf{G}) = 1$.

Here's a clever way to obtain an equivariant function from any old function. We'll assume for simplicity that $\mathbf{Y} = \mathbb{R}^d$, but these ideas generalize to when $\mathbf{Y}$ is a Hilbert space. Let $R \colon \mathbf{G} \to U(\mathbf{Y})$ be a unitary representation of $\mathbf{G}$ on $\mathbf{Y}$. A **unitary representation** is a linear representation of $\mathbf{G}$ such that $R(g)$ is a unitary operator for every $g \in \mathbf{G}$. A **unitary operator** R is a surjective bounded operator on $\mathbf{Y}$ that preserves the inner product, $\langle Ry, Ry' \rangle_{\mathbf{Y}} = \langle y, y' \rangle_{\mathbf{Y}}$, for all $y, y' \in \mathbf{Y}$. For $\mathbf{Y} = \mathbb{R}^d$, these are just the $d \times d$ orthogonal matrices, $O(d)$, which preserve the Euclidean inner product. Note that a unitary representation is a linear action that preserves the inner product; from this it follows that $\langle R(g)y, y' \rangle_{\mathbf{Y}} = \langle y, R(g^{-1})y' \rangle_{\mathbf{Y}}$ for all $g \in \mathbf{G}$, $y, y' \in \mathbf{Y}$. When $\mathbf{G}$ is compact, any linear representation R can be made unitary with respect to a new inner product defined by

$$\langle y, y' \rangle_{R\ \mathbf{Y}} := \int_{\mathbf{G}} \langle R(g)y, R(g)y' \rangle_{\mathbf{Y}} \lambda(dg) \,.$$

It's easy to check that R is unitary with respect to $\langle \bullet, \bullet \rangle_{R\ \mathbf{Y}}$. (This is known as the "unitarian trick," widely attributed to the mathematician and theoretical physicist Hermann Weyl.)

We can also extend the Hilbert space construction from the previous subsection so that $L^2(\mu)$ is a Hilbert space of functions $f\colon \mathbf{X} \to \mathbf{Y}$ with inner product,

$$\langle f, f' \rangle_{L^2(\mu)} := \int_{\mathbf{X}} \langle f(x), f'(x) \rangle_{\mathbf{Y}}\ \mu(dx)\ . \tag{4.3}$$

A function $f\colon \mathbf{X} \to \mathbf{Y}$ is in $L^2(\mu)$ if $||f||_{L^2(\mu)} < \infty$. For notational convenience, I'll omit the subscript $\mathbf{Y}$ from the inner product, and I'll denote $\langle \bullet, \bullet \rangle_{L^2(\mu)}$ by $\langle \bullet, \bullet \rangle_\mu$. I will also assume for simplicity that $\mathbf{Y} = \mathbb{R}^d$ with Euclidean inner product, so R is just an orthogonal representation of $\mathbf{G}$.

Let μ be a $\mathbf{G}$-invariant σ-finite measure on $\mathbf{X}$, and define $\mathcal{Q}$ as an operator on $L^2(\mu)$ by

$$(\mathcal{Q}f)(x) := \int_{\mathbf{G}} R(g^{-1}) f(gx) \lambda(dg)\ . \tag{4.4}$$

It's equivariant:

$$\begin{aligned}(\mathcal{Q}f)(hx) &= \int_{\mathbf{G}} R(g^{-1}) f(ghx) \lambda(dg) = \int_{\mathbf{G}} R(hg^{-1}) f(gx) \lambda(dg) \\ &= R(h) \int_{\mathbf{G}} R(g^{-1}) f(gx) \lambda(dg) \\ &= R(h)(\mathcal{Q}f)(x)\ .\end{aligned}$$

Note that this construction would work even for non-compact groups if we use right Haar measure. However, the following may fail for non-compact groups.

Theorem 4.4. *The operator $\mathcal{Q}$ is the (orthogonal) projection operator in the Hilbert space $L^2(\mu)$ onto the subspace of $\mathbf{G}$-equivariant functions (where equivariance is with respect to the representation R).*

In this form, the theorem is due to Elesedy and Zaidi [EZ21]. Our proof will be essentially the same, but with explicit reliance on the properties of Hilbert spaces as given in the previous subsection.

Lemma 4.5. *Let $f, f' \in L^2(\mu)$. The operator $\mathcal{Q}$ (corresponding to orthogonal representation R) has the following properties:*

i) $\mathcal{Q}$ is linear.

ii) $\langle \mathcal{Q}f, f' \rangle_\mu = \langle f, \mathcal{Q}f' \rangle_\mu$.

iii) $f \in L^2(\mu)$ is $\mathbf{G}$-equivariant if and only if $\mathcal{Q}f = f$. Therefore, $\mathcal{Q}^2 f = \mathcal{Q}f$.

iv) $||\mathcal{Q}f||_\mu \leq ||f||_\mu$. Therefore, $||\mathcal{Q}||_\mu \leq 1$, and if $f \in L^2(\mu)$ then $\mathcal{Q}f \in L^2(\mu)$.

Proof. Linearity just follows from linearity of the integral. For ii),

$$\begin{aligned}
\langle \mathcal{Q}f, f' \rangle_\mu &= \int_{\mathbf{X}} \langle (\mathcal{Q}f)(x), f'(x) \rangle \mu(dx) \\
&= \int_{\mathbf{X}} \int_{\mathbf{G}} \langle R(g^{-1}) f(gx), f'(x) \rangle \lambda(dg) \mu(dx) \quad \text{(linearity of integral and inner product)} \\
&= \int_{\mathbf{X}} \int_{\mathbf{G}} \langle f(gx), R(g) f'(x) \rangle \lambda(dg) \mu(dx) \qquad R \text{ is unitary} \\
&= \int_{\mathbf{X}} \int_{\mathbf{G}} \langle f(x), R(g) f'(g^{-1}x) \rangle \lambda(dg) \mu(dx) \qquad \text{Fubini; } \mathbf{G}\text{-invariance of } \mu\text{; Fubini} \\
&= \int_{\mathbf{X}} \int_{\mathbf{G}} \langle f(x), R(g^{-1}) f'(gx) \rangle \lambda(dg) \mu(dx) \qquad \text{unimodularity of } \mathbf{G} \\
&= \int_{\mathbf{X}} \langle f(x), (\mathcal{Q}f')(x) \rangle \mu(dx) \\
&= \langle f, \mathcal{Q}f' \rangle_\mu \ .
\end{aligned}$$

Note that the same holds for locally compact $\mathbf{G}$ when $\mathcal{Q}$ is defined with right Haar measure λ_r and μ is "right-invariant" (as in Section 3), with $g_\# \mu = \Delta(g)\mu$. That proves ii).

For iii), suppose that $\mathcal{Q}f = f$. Then by the right-invariance of $\mathbf{G}$,

$$f(hx) = (\mathcal{Q}f)(hx) = \int_{\mathbf{G}} R(g^{-1}) f(ghx) \lambda(dg) = \int_{\mathbf{G}} R(hg^{-1}) f(gx) \lambda(dg) = R(h)(\mathcal{Q}f)(x) = R(h) f(x) \ ,$$

so f is $\mathbf{G}$-equivariant. (Note that, again, this would hold for a locally compact group and λ_r.) Conversely, suppose that f is $\mathbf{G}$-equivariant. Then

$$\mathcal{Q}f(x) = \int_{\mathbf{G}} R(g^{-1}) f(gx) \lambda(dg) = \int_{\mathbf{G}} R(g^{-1}) R(g) f(x) \lambda(dg) = f(x) \lambda(\mathbf{G}) = f(x) \ .$$

That proves iii). (Note that in the non-compact case, if f is equivariant then $\mathcal{Q}f = +\infty$, and we lose the ability to use $\mathcal{Q}$ as a projection.)

Now, ii), iii), and Cauchy–Schwarz imply that

$$\begin{aligned}
||\mathcal{Q}f||_\mu^2 &= \langle \mathcal{Q}f, \mathcal{Q}f \rangle_\mu = \langle f, \mathcal{Q}^2 f \rangle_\mu \\
&= \langle f, \mathcal{Q}f \rangle_\mu \\
&\leq ||f||_\mu ||\mathcal{Q}f||_\mu \ ,
\end{aligned}$$

and therefore $||\mathcal{Q}f||_\mu \leq ||f||_\mu$. The operator norm $||\mathcal{Q}||_\mu$ is defined as $\inf\{c \geq 0 \colon ||\mathcal{Q}f||_\mu \leq c||f||_\mu \text{ for all } f \in L^2(\mu)\}$. Since the previous inequality holds for arbitrary $f \in L^2(\mu)$, we can see that $||\mathcal{Q}||_\mu \leq 1$. □

Lemma 4.6. *Let $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$ be the set of all $\mathbf{G}$-equivariant functions (relative to R) in $L^2(\mu)$. $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$ is a closed linear subset of $L^2(\mu)$, i.e., it is a subspace, and $\mathcal{Q}$ is the projection of $L^2(\mu)$ onto it.*

Proof. Linearity is easy to check. If f, f' are both equivariant functions and $a, b \in \mathbb{R}$, then $af + bf'$ is also equivariant: for $h \in \mathbf{G}$, $x \in \mathbf{X}$,

$$af(hx) + bf'(hx) = aR(h)f(x) + bR(h)f'(x) = R(h)(af(x) + bf'(x)) \; .$$

The fact that it is closed is only slightly more involved. Let (f_n) be a sequence in $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$ that converges to $f \in L^2(\mu)$. We need to show that $f \in \mathcal{E}_R(\mathbf{X}, \mathbf{Y})$.

$$\begin{aligned} ||f_n - f||_\mu^2 &= ||(\mathcal{Q}f_n - \mathcal{Q}f) + (\mathcal{Q}f - f)||_\mu^2 \\ &= ||\mathcal{Q}f_n - \mathcal{Q}f||_\mu^2 + 2\langle \mathcal{Q}(f_n - f), \mathcal{Q}f - f\rangle_\mu + ||\mathcal{Q}f - f||_\mu^2 \; . \end{aligned}$$

As $n \to \infty$, the LHS goes to zero because $f_n \to f$. The first term on the RHS, $||\mathcal{Q}f_n - \mathcal{Q}f||_\mu^2 = ||\mathcal{Q}||_\mu^2 ||f_n - f||_\mu^2$, also goes to zero because $||\mathcal{Q}||_\mu \leq 1$. The second term is

$$\begin{aligned} \langle \mathcal{Q}(f_n - f), \mathcal{Q}f - f\rangle_\mu &= \langle (f_n - f), \mathcal{Q}^2 f - \mathcal{Q}f\rangle_\mu \\ &= \langle (f_n - f), \mathcal{Q}f - \mathcal{Q}f\rangle_\mu = 0 \; , \end{aligned}$$

for all n. So we have

$$0 = \lim_{n\to\infty} ||f_n - f||_\mu^2 = ||\mathcal{Q}f - f||_\mu^2 \; ,$$

so $||\mathcal{Q}f - f||_\mu^2 = 0$, which implies that $\mathcal{Q}f = f$ and therefore by Lemma 4.5 iii) that $f \in \mathcal{E}_R(\mathbf{X}, \mathbf{Y})$. So $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$ is a Hilbert subspace of $L^2(\mu)$. If we can show that

$$\langle f - \mathcal{Q}f, \bar{f}\rangle_\mu = 0 \; , \quad \text{all } f \in L^2(\mu), \; \bar{f} \in \mathcal{E}_R(\mathbf{X}, \mathbf{Y}) \; ,$$

then $\mathcal{Q}$ is the projection of $L^2(\mu)$ onto $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$, by Theorem 4.3 vii). To that end,

$$\begin{aligned} \langle f - \mathcal{Q}f, \bar{f}\rangle_\mu &= \langle f - \mathcal{Q}f, \mathcal{Q}\bar{f}\rangle_\mu \\ &= \langle \mathcal{Q}(f - \mathcal{Q}f), \bar{f}\rangle_\mu \\ &= \langle \mathcal{Q}f - \mathcal{Q}^2 f, \bar{f}\rangle_\mu \\ &= \langle \mathcal{Q}f - \mathcal{Q}f, \bar{f}\rangle_\mu = 0 \; . \end{aligned}$$

This is true for arbitrary $f \in L^2(\mu)$ and $\bar{f} \in \mathcal{E}_R(\mathbf{X}, \mathbf{Y})$, so $\mathcal{Q}$ is the projection. □

That proves Theorem 4.4.

G-invariant functions

Note that $\mathcal{Q}$ depends implicitly on the representation R of $\mathbf{G}$, so a version of Theorem 4.4 holds for any $\mathcal{Q}$ based on an orthogonal representation of $\mathbf{G}$. Inequivalent representations may lead to different subspaces $\mathcal{E}_R(\mathbf{X}, \mathbf{Y})$, and hence different projections $\mathcal{Q}_R$. In particular, if R is the trivial representation $R(g) = \mathbf{I}$, $\mathcal{Q}_\mathbf{I}$ is the projection onto the subspace of $\mathbf{G}$-*invariant* functions, $\mathcal{E}_\mathbf{I}(\mathbf{X}, \mathbf{Y})$, where

$$\mathcal{Q}_\mathbf{I} f = \int_\mathbf{G} f(gx)\lambda(dg) \; .$$

Corollary 4.7. $\mathcal{Q}_\mathbf{I}$ *is the projection of* $L^2(\mu)$ *onto* $\mathcal{E}_\mathbf{I}(\mathbf{X}, \mathbf{Y})$.

An example of the benefit of model symmetry

This example is taken from Elesedy and Zaidi [EZ21]. Suppose that $X \sim \mu$, where μ is a $\mathbf{G}$-invariant probability distribution, and $\mathbf{G}$ is compact. Let $Y = f^*(X) + \epsilon \in \mathbb{R}^d$, where ϵ is random independent (of X) noise in $\mathbb{R}^d$ with zero mean and finite variance. Finally, assume that f^* is $\mathbf{G}$ equivariant with respect to some orthogonal representation R, in the sense that $f^*(gx) = R(g)f^*(x)$. Define the **generalization gap** as

$$\delta(f, \mathcal{Q}f) := E[||Y - f(X)||_2^2] - E[||Y - \mathcal{Q}f(X)||_2^2] \,. \tag{4.5}$$

This is just the "extra error" we incur when using $f \in L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ instead of $\mathcal{Q}f$, as measured by mean-square error (MSE).

Proposition 4.8. *For the model described above, the generalization gap is*

$$\delta(f, \mathcal{Q}f) = ||f - \mathcal{Q}f||_\mu^2 \,.$$

Here, the extra error is just the norm of the part of f that is orthogonal to its equivariant part.

Proof. The result follows from the properties we established above: since the noise is mean zero, finite variance, and independent, we have, after cancelling shared terms,

$$\begin{aligned}
E[||Y - f(X)||_2^2] - E[||Y - \mathcal{Q}f(X)||_2^2] &= E_\mu[||f^*(X) - f(X)||_2^2] - E_\mu[||f^*(X) - \mathcal{Q}f(X)||_2^2] \\
&= E_\mu[||f^*(X)||^2] + E_\mu[||f(X)||^2] - 2E_\mu[\langle f^*(X), f(X)\rangle] \\
&\quad - E_\mu[||f^*(X)||^2] - E_\mu[||\mathcal{Q}f(X)||^2] + 2E_\mu[\langle f^*(X), \mathcal{Q}f(X)\rangle] \\
&= E_\mu[||f(X) - \mathcal{Q}f(X)||_2^2] - 2E_\mu[\langle f^*(X) - \mathcal{Q}f(X), f(X) - \mathcal{Q}f(X)\rangle] \,.
\end{aligned}$$

The second term in the last line is the inner product in $L^2(\mu)$ between an equivariant function, $f^*(X) - \mathcal{Q}f(X)$, and a function that is orthogonal to all equivariant functions, $f(X) - \mathcal{Q}f(X)$. So that inner product is equal to zero and the result follows. □

This is enough to show that there's a benefit in terms of MSE, and it is strictly greater than zero if f has a non-zero orthogonal component. Note that it also applies to situation in which f^* is $\mathbf{G}$-invariant. Quantifying the benefit for specific models still requires considerable further work; see [EZ21; Ele21] for examples.

I'll note that this approach is limited in two ways: first, we required $\mathbf{G}$ to be compact; second, we required the distribution of X to be $\mathbf{G}$-invariant. We'll see if we can address these with other methods.

4.3 Symmetry through conditioning

Detour: Conditional expectation

Example 4.2 is formalized as follows. I've modified the typical random variable notation to be in terms of measurable functions $f \colon \mathbf{X} \to \mathbb{R}$, but recall that a random variable is nothing but a measurable function, so that for a probability measure ν on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$,

$$E_\nu[f] = \int_{\mathbf{X}} f(x)\nu(dx) \,.$$

The following theorem is Theorem IV.1.11 in [Çin11], whose proof just relies on the Hilbert space/-subspace properties of $L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ and $L^2(\mathbf{X}, \mathcal{A}, \mu)$, with $\mathcal{A}$ a sub-σ-algebra of $\mathcal{B}(\mathbf{X})$.

Theorem 4.9. *Let ν be a probability measure on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, and $\mathcal{A}$ a sub-σ-algebra of $\mathcal{B}(\mathbf{X})$. For every $f \in L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)$, there exists a unique (up to a.s.-equivalence) $\bar{f} \in L^2(\mathbf{X}, \mathcal{A}, \nu)$ such that*

$$E_\nu[|f - \bar{f}|^2] = \inf_{f' \in L^2(\mathbf{X}, \mathcal{A}, \nu)} E_\nu[|f - f'|^2] \,. \tag{4.6}$$

Moreover, $f - \bar{f}$ is orthogonal to $L^2(\mathbf{X}, \mathcal{A}, \nu)$, in the sense that for every function $k \in L^2(\mathbf{X}, \mathcal{A}, \nu)$,

$$E_\nu[k(f - \bar{f})] = 0 \,. \tag{4.7}$$

Note that saying that $\bar{f} \in L^2(\mathbf{X}, \mathcal{A}, \nu)$ means that $\bar{f}$ is measurable with respect to $\mathcal{A}$ and $\mathcal{B}(\mathbb{R})$, as well as being square-integrable.

Theorem 4.9 makes the following definition unique up to a.s.-equivalence.

Definition 4.10. Let $\mathcal{A}$ be a sub-σ-algebra of $\mathcal{B}(\mathbf{X})$. For any $f \in L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)$, the **conditional expectation** of f given $\mathcal{A}$ is the unique element $E_\nu[f \mid \mathcal{A}] := \bar{f}$ of $L^2(\mathbf{X}, \mathcal{A}, \nu)$ such that

$$E_\nu[f \cdot f'] = E_\nu[\bar{f} \cdot f'] \,,$$

for all $f' \in L^2(\mathbf{X}, \mathcal{A}, \nu)$. (Here, $\cdot$ is just multiplication.)

Clearly, $\bar{f}$ as defined this way is the orthogonal projection of f onto $L^2(\mathbf{X}, \mathcal{A}, \nu)$.

This definition extends to functions in $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)_+$ (note the restriction to positive functions) word-for-word, and the two agree for functions in $L^2(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)$. One can then use the decomposition of a function into its positive and negative parts, $f = f_+ - f_-$ to extend to all of $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)$. Note that even in $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \nu)$, the orthogonality property (4.7) holds for all bounded $\mathcal{A}$-measurable functions k. See Chap. IV.1 in [Çin11] and Chap. 23 in [JP04] (particularly the latter for the Hilbert space perspective) for a more detailed presentation.

Back to invariance with benefits

We saw in Theorem 2.4 that the invariant σ-algebra, denoted $\mathcal{I}(\mathbf{X})$, is generated by the set of all invariant functions $\mathbf{X} \to \mathbb{R}$, i.e., $L^2(\mathbf{X}, \mathcal{I}(\mathbf{X}), \nu)$ is the set of all ν-square-integrable and $\mathbf{G}$-invariant functions. If μ is a $\mathbf{G}$-invariant probability measure, we see that if $f \colon \mathbf{X} \to \mathbb{R}$ is in $L^2(\mu)$, then

$$\mathcal{Q}_\mathbf{I} f = E_\mu[f \mid \mathcal{I}(\mathbf{X})] \,.$$

Note that establishing $\mathcal{Q}_\mathbf{I}$ as the orthogonal projection onto $L^2(\mathbf{X}, \mathcal{I}(\mathbf{X}), \mu)$ required μ to be $\mathbf{G}$-invariant and that $\mathbf{G}$ is compact. The conditional expectation $E_\nu[\bullet \mid \mathcal{I}(\mathbf{X})]$ is not bound by the same constraints.

This lets us extend the generalization gap analysis of the regression model above—at least in the case where f^* is $\mathbf{G}$-invariant—to situations where the distribution of X is ν, possibly not $\mathbf{G}$-invariant, and where $\mathbf{G}$ is non-compact (say lcscH). So: Let $Y = f^*(X) + \epsilon \in \mathbb{R}^d$, where ϵ is random independent (of X) noise in $\mathbb{R}^d$ with zero mean and finite variance, and $f^* \colon \mathbf{X} \to \mathbb{R}^d$ is

some square-integrable (w.r.t. ν) $\mathbf{G}$-invariant function. Let $\bar{f} := E_\nu[f \mid \mathcal{I}(\mathbf{X})]$ for f such that $E_\nu[||\bar{f}||^2] < \infty$.[7] The generalization gap here is

$$\begin{aligned}
\delta(f, \bar{f}) &= E[||Y - f(X)||^2] - E[||Y - \bar{f}(X)||^2] \\
&= E_\nu[||f^* - f||^2] - E_\nu[||f^* - \bar{f}||^2] \\
&= E_\nu[|||f^*|||^2 + E_\nu[||f||^2] - 2E_\nu[\langle f^*, f\rangle] \\
&\qquad - E_\nu[||f^*||^2] - E_\nu[||\bar{f}||^2] + 2E_\nu[\langle f^*, \bar{f}\rangle] \\
&= E_\nu[||f - \bar{f}||^2] - 2E_\nu[\langle f^* - \bar{f}, f - \bar{f}\rangle] \,.
\end{aligned}$$

Here, we make two observations. First, a vector-valued function $f\colon \mathbf{X} \to \mathbb{R}^d$ can be written as $f(x) = (f_1(x), \ldots, f_d(x))$, and is $\mathbf{G}$-invariant if and only if each of the component functions f_i are $\mathbf{G}$-invariant, and therefore each component of $f^* - \bar{f}$ is $\mathbf{G}$-invariant and thus $\mathcal{I}(\mathbf{X})$-measurable. Secondly, the orthogonality property in (4.7) holds for each component, i.e., $E_\nu[(f^* - \bar{f})_i\ (f - \bar{f})_i] = E_\nu[(f^* - \bar{f})_i\ (f_i - \bar{f}_i)] = 0$ for each $i = 1, \ldots, d$, because the orthogonal projection is a linear operator. So,

$$E_\nu[\langle f^* - \bar{f}, f - \bar{f}\rangle] = \sum_{i=1}^{d} E_\nu[(f^* - \bar{f})_i\ (f_i - \bar{f}_i)] = 0 \,,$$

and therefore

$$\delta(f, \bar{f}) = \delta(f, E_\nu[f \mid \mathcal{I}(\mathbf{X})]) = E_\nu[||f - \bar{f}||^2] \,.$$

This is a substantial generalization of the result in Proposition 4.8, which was based on $\mathcal{Q}_\mathbf{I}$. The flip-side is that rather than a constructive definition of the orthogonal projection (as we had with $\mathcal{Q}$), the definition of conditional expectation relies on the existence and uniqueness of the orthogonal projection in L^2 so that it is well defined, but passive in that all it instructs us to do is to solve the minimization problem in (4.6). One potential way around this is to find the Markov kernel corresponding to ν conditioned on $\mathcal{I}(\mathbf{X})$, say $K_{\mid X \mid \mathcal{I}(\mathbf{X})}(x, \bullet)$. It is straightforward to show that this conditional distribution is $\mathbf{G}$-invariant in both arguments, which would let us (at least in the compact $\mathbf{G}$ case) construct an averaging operator analogous to $\mathcal{Q}_\mathbf{I}$.

Finally, I don't know of an analogous approach (based on conditional expectation) to obtain an equivariant function.

[7] Note that the square-integrability requirement here is not required for the conditional expectation, but because we're using MSE.

References

[AB06] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, 2006.

[CDL20] S. Chen, E. Dobriban, and J. H. Lee. "A Group-Theoretic Framework for Data Augmentation". *Journal of Machine Learning Research* 21.245 (2020), pp. 1–71. URL: http://jmlr.org/papers/v21/20-163.html.

[Çin11] E. Çinlar. *Probability and Stochastics*. Springer New York, 2011.

[Eat89] M. L. Eaton. *Group invariance in applications in statistics*. Vol. 1. Regional Conference Series in Probability and Statistics. Haywood, CA and Alexandria, VA: Institute of Mathematical Statistics and American Statistical Association, 1989.

[ES99] M. L. Eaton and W. D. Sudderth. "Consistency and Strong Inconsistency of Group-Invariant Predictive Inferences". *Bernoulli* 5.5 (1999), pp. 833–854.

[ES01] M. L. Eaton and W. D. Sudderth. "Best invariant predictive distributions". *Contemporary Mathematics*. Ed. by M. A. G. Viana and D. S. P. Richards. Vol. 287. American Mathematical Society, 2001, pp. 49–62.

[ES04] M. L. Eaton and W. D. Sudderth. "Properties of Right Haar Predictive Inference". *Sankhya: The Indian Journal of Statistic* 66.3 (2004), pp. 487–512.

[Ele21] B. Elesedy. "Provably Strict Generalisation Benefit for Invariance in Kernel Methods". *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 17273–17283. URL: https://proceedings.neurips.cc/paper/2021/file/8fe04df45a22b63156ebabbb064fcd5e-Paper.pdf.

[EZ21] B. Elesedy and S. Zaidi. "Provably Strict Generalisation Benefit for Equivariant Models". *Proceedings of the 38th International Conference on Machine Learning*. 2021, pp. 2959–2969.

[Fol16] G. B. Folland. *A Course in Abstract Harmonic Analysis*. 2nd ed. Vol. 29. CRC press, 2016.

[Fra61] D. A. S. Fraser. "The Fiducial Method and Invariance". *Biometrika* 48.3/4 (1961), pp. 261–280.

[Fra67] D. A. S. Fraser. "Statistical Models and Invariance". *The Annals of Mathematical Statistics* 38.4 (Aug. 1967), pp. 1061–1067.

[HB66] R. B. Hora and R. J. Buehler. "Fiducial Theory and Invariant Estimation". *The Annals of Mathematical Statistics* 37.3 (1966), pp. 643–656.

[HOA22] K. H. Huang, P. Orbanz, and M. Austern. "Quantifying the Effects of Data Augmentation" (Feb. 2022). eprint: 2202.09134. URL: https://arxiv.org/pdf/2202.09134.pdf.

[JP04] J. Jacod and P. Protter. *Probability Essentials*. 2nd. Springer-Verlag Berlin Heidelberg, 2004.

[Kal02] O. Kallenberg. *Foundations of Modern Probability*. 2nd. Springer-Verlag New York, 2002.

[Kal17] O. Kallenberg. *Random Measures, Theory and Applications*. Springer International Publishing, 2017.

[LC98] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. third. New York, NY: Springer–Verlag, 1998.

[Lyl+20] C. Lyle et al. "On the Benefits of Invariance in Neural Networks" (2020). arXiv: 2005.00178 [cs.LG].

[Rob07] C. P. Robert. *The Bayesian Choice*. 2nd ed. Springer New York, 2007.

[Sch95] M. J. Schervish. *Theory of Statistics*. Springer-Verlag New York, 1995.

[Wes59] O. Wesler. "Invariance Theory and a Modified Minimax Principle". *The Annals of Mathematical Statistics* 30.1 (1959), pp. 1–20.

[Wij90] R. A. Wijsman. *Invariant measures on groups and their use in statistics*. Vol. 14. Lecture Notes–Monograph Series. Hayward, CA: Institute of Mathematical Statistics, 1990.

Index