EXPLOITING SYMMETRIES TO CONSTRUCT EFFICIENT MCMC ALGORITHMS WITH AN APPLICATION TO SLAM

by

ROSHAN SHARIFF

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

UNIVERSITY OF ALBERTA

© Roshan Shariff, 2015

ABSTRACT

Sampling from a given probability distribution is a key problem in many different disciplines. Markov chain Monte Carlo (MCMC) algorithms approach this problem by constructing a random walk governed by a specially constructed transition probability distribution. As the random walk progresses, the distribution of its states converges to the required target distribution. The Metropolis-Hastings (MH) algorithm is a generally applicable MCMC method which, given a proposal distribution, modifies it by adding an accept/reject step: it proposes a new state based on the proposal distribution and the existing state of the random walk, then either accepts or rejects it with a certain probability; if it is rejected, the old state is retained. The MH algorithm is most effective when the proposal distribution closely matches the target distribution: otherwise most proposals will be rejected and convergence to the target distribution will be slow. The proposal distribution should therefore be designed to take advantage of any known structure in the target distribution.

A particular kind of structure that arises in some probabilistic inference problems is that of symmetry: the problem (or its sub-problems) remains unchanged under certain transformations. A simple kind of symmetry is the choice of a coordinate system in a geometric problem; translating and rotating the origin of a plane does not affect the relative positions of any points on it. The field of group theory has a rich and fertile history of being used to characterize such symmetries; in particular, topological group theory has been applied to the study of both continuous and discrete symmetries. Symmetries are described by a group that acts on the state space of a problem, transforming it in such a way that the problem remains

unchanged. We consider problems in which the target distribution has factors, each of which has a symmetry group; each factor's value does not change when the space is transformed by an element of its corresponding symmetry group.

This thesis proposes a variation of the MH algorithm where each step first chooses a random transformation of the state space and then applies it to the current state; these transformations are elements of suitable symmetry groups. The main result of this thesis extends the acceptance probability formula of the textbook MH algorithm to this case under mild conditions, adding much-needed flexibility to the MH algorithm. The new algorithm is also demonstrated in the Simultaneous Localization and Mapping (SLAM) problem in robotics, in which a robot traverses an unknown environment, and its trajectory and a map of the environment must be recovered from sensor observations and known control signals. Here the group moves are chosen to exploit the SLAM problem's natural geometric symmetries, obtaining the first fully rigorous justification of a previous MCMC-based SLAM method. New experimental results comparing this method to existing state-of-the-art specialized methods on a standard range-only SLAM benchmark problem validate the strength of the approach.

PREFACE

This thesis is an original work by Roshan Shariff. Parts of it will appear in a forthcoming paper to be published in the *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR: W&CP*, San Diego, CA, USA, May 9–12 (Shariff, György, and Szepesvári, 2015, in press).

ACKNOWLEDGMENTS

I sincerely thank my adviser Csaba Szepesvári for his encouragement and patient support, which made this work possible. His guidance kept me on track and his enthusiasm kept me moving forward. I'm also grateful to András György for his suggestions and improvements, which have significantly benefited the readability of this thesis. Finally, I would like to acknowledge the assistance of the Alberta Innovates – Technology Futures Graduate Student Scholarship.

CONTENTS

Introduction 1

1 MOTIVATION AND PROBLEM STATEMENT 3

2 GROUPS AND HOMOGENEOUS SPACES 6
- 2.1 Continuous Group Actions 7
- 2.2 Orbits and Transitive Actions 8
- 2.3 Isomorphisms of Groups and G-spaces 9
- 2.4 Stabilizer Subgroups 11
- 2.5 Cosets and Quotient Spaces 12
- 2.6 Homogeneous Spaces 14
- 2.7 Some Topological Consequences 16
- 2.8 Invariant and Relatively Invariant Measures 17

3 METROPOLIS-HASTINGS WITH GROUP TRANSFORMATIONS 22
- 3.1 Metropolis-Hastings Based on Group Moves 23
- 3.2 Mixtures of Group Moves 25
- 3.3 Correctness 25
- 3.4 Proofs 27

4 EXPLOITING SYMMETRIES 35

5 THE SLAM PROBLEM 39
- 5.1 Symmetries of SLAM 40
- 5.2 The MCMC-SLAM Algorithm 41

6 EXPERIMENTS 43
- 6.1 Robot Models and Algorithm Details 45

7 CONCLUSIONS AND FUTURE WORK 47

LIST OF TABLES

Table 1 Comparison of Trajectory RMS Errors. 44

LIST OF FIGURES

Figure 1 A probability density on $\mathbb{R}^2$ with two factors. 3
Figure 2 An MCMC sampled trajectory and map. 44

INTRODUCTION

Symmetry is what we see at a glance;
based on the fact that there is no
reason for any difference ...

— Blaise Pascal, *Pensées*

Probabilistic reasoning plays a major role in many disciplines. Many state-of-the-art artificial intelligence (AI) approaches to major challenges depend on the manipulation of probability distributions (Korb and Nicholson, 2003; Russel and Norvig, 2009; Poole and Mackworth, 2010). In particular, probabilistic graphical models are widely used in computer vision (Prince, 2012), robotics (Thrun et al., 2005; Ferreira and Dias, 2014), speech and natural language processing (Manning and Schuetze, 1999), machine learning (Bishop, 2006; Murphy, 2012) and agent research (Xiang, 2002). A key step of working with probabilistic graphical models is inference: the computation of a posterior distribution given the model and some data. As the posterior can rarely be expressed in a closed form amenable to direct evaluation by a computer, one often must resort to approximate inference methods (Pearl, 1988; Darwiche, 2009; Koller and Friedman, 2009); Markov chain Monte Carl (MCMC) methods are often ideal for the task (Tierney, 1994). Many inference problems have a certain type of *symmetry* structure which manifests in the posterior distribution. In this thesis, we focus on a particular MCMC method, the Metropolis-Hastings (MH) algorithm, describing a version of it that is able to benefit from known (but possibly approximate) symmetries of the posterior.

The MH algorithm takes a target distribution and a user-chosen "proposal" kernel and transforms the latter into a new Markov transition kernel; the resulting Markov chain will have a limiting distribution equal to the target under mild conditions on the proposal (Metropolis et al., 1953; Hastings, 1970). While the MH algorithm gives substantial flexibility in choosing the proposal kernel, the calculations needed to implement the MH algorithm are simple only when it has certain special forms, such as the textbook case when the target and proposal measures have densities with respect to a common "reference" measure[1] (Tierney, 1994). In this thesis we describe two new classes of proposal kernels, based on *group transformations of the state space* and give the corresponding MH algorithms in closed form. The algorithms require basically the same amount of computation as the textbook MH algorithm, while we will argue that they significantly expand the scope of the MH algorithm. In particular, we will show that they allow us to exploit known symmetries in factors of the target distribution, which does indeed speed up computation; we will argue that this also improves the rate of convergence. We illustrate the results by specializing the algorithm to the simultaneous localization and mapping (SLAM) problem in robotics (Thrun et al., 2005) and argue that the algorithm essentially recovers the MCMC-SLAM method of Torma et al. (2010), providing much needed insight into the behavior of this method as well as the first fully rigorous proof of its correctness.[2] In fact, it was this method that served as the inspiration for the present thesis. In a new set of experiments, we demonstrate that this algorithm is competitive with state-of-the-art methods of robotics.

The thesis is organized as follows: In Chapter 1 we use an example to motivate our approach. Chapter 2 (which may be passed over on a first reading) sets up the mathematical background. The main contributions of this thesis lie in Chapter 3, wherein our approach is described and proved correct. Chapter 4 expands upon the example to illustrate how our approach can exploit symmetries. Chapter 5 is devoted to describing the SLAM problem, its symmetries, and how the general construction of Chapter 3 can be instantiated in this setting to recover the MCMC-SLAM algorithm. We close the thesis by providing experimental results on range-only SLAM (Chapter 6) followed by our conclusions (Chapter 7).

1 This restriction disallows even Gibbs sampling, since the target distribution typically has a density with respect to the Lebesgue measure on $\mathbb{R}^n$, which however is zero on the one-dimensional subspaces on which proposals are made (see Section 3.2). The target distribution must therefore be conditioned on the space of proposals, which is straightforward for Lebesgue measures and linear subspaces but requires the machinery of measure theory to be correct in general (Chang and Pollard, 1997).

2 Theorem 2 of Torma et al. (2010) is not correct when, in the notation of Section 3.1, $\Delta_r^G \neq 1$ or $\chi \neq 1$. However, this does not affect the special case of SLAM.

1

MOTIVATION AND PROBLEM STATEMENT

Suppose we want to draw samples from the simple two-dimensional probability distribution P of Fig. 1. Its density $p(x,y)$ has two factors: $p_1(x,y)$ and $p_2(x,y)$, which need not be probability densities themselves (e.g., p_2 is not integrable):

$$p_1(x,y) = c\, p_1(x,y)\, p_2(x,y)\,.$$

The MCMC approach is to construct a transition probability distribution that induces a random walk over $\mathbb{R}^2$, the distribution of which converges to P in the steady state. The Metropolis-Hastings (MH) algorithm allows us to specify a *proposal* distribution, and under mild conditions, constructs a suitable MCMC transition kernel by proposing a new state but rejecting it with some probability. With some *no-reject* proposal kernels the rejection probability is zero, which means the proposal kernel is itself suitable as a transition kernel. The MCMC algorithm will be efficient if the proposal kernel does not often propose low-probability states (which would increase the rejection rate) and quickly explores the high-probability states (speeding up convergence to the steady state).

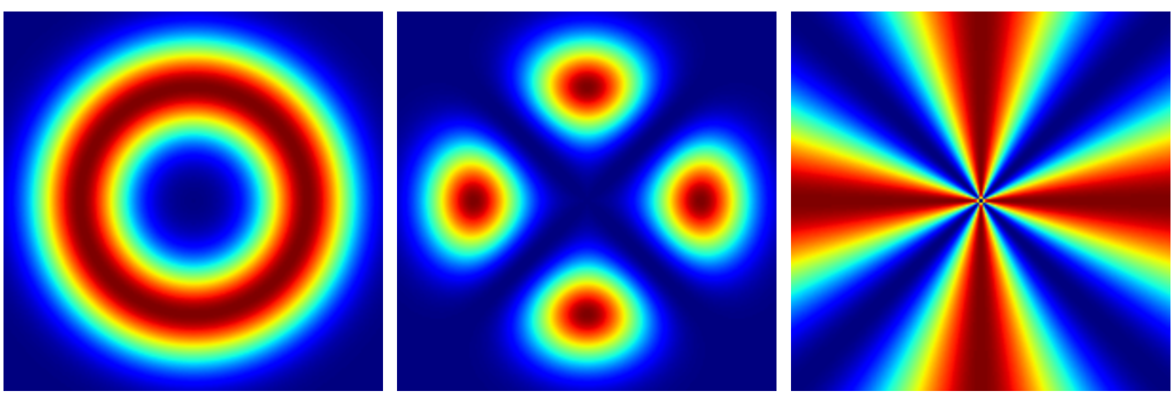

Figure 1: A probability density p on $\mathbb{R}^2$ (center) with factors p_1 (left) and p_2 (right).

Often, a proposal kernel updates the state by modifying one variable at a time (the canonical example is Gibbs sampling; some multivariate "slice sampling" kernels also do this). However, it is immediately apparent that such an update would be problematic for our example: it would be impossible to move between the $\pm X$ and $\pm Y$ modes of the distribution without transiting through a low-probability region. Another common approach is to change all the variables by a small delta, perhaps drawn from a multivariate normal distribution. However, the variance of this proposal kernel must be carefully tuned for each variable: too small and it will be confined to one mode in a multi-modal distribution like ours; too large and it will often propose points in the low-probability regions. In general, this idea does not work well with multi-modal distributions.

One might argue that we have overstated the difficulty of the problem. One sees at a glance that p_1 is radially symmetric and p_2 is scale invariant: we can make sampling much easier simply by re-parametrizing the state space using polar coordinates (r, θ) instead of Cartesian coordinates (x, y). Updating one variable at a time is then very effective: one can draw an independent sample from P just by sampling r according to p_1 and θ according to p_2. Indeed, our difficulties were simply because the Cartesian representation of the state space is mismatched with the independence and symmetry structure of the problem, whereas in the polar representation the r and θ variables are independent with distributions derived from p_1 and p_2, respectively.

In general, however, it is not always possible to come up with a parametrization that reflects so cleanly the symmetries of the factors. Instead, since the symmetries are more readily apparent than a suitable parametrization, we can sidestep the problem of re-parametrizing the state space and instead work directly with the known symmetries. To do this, we will use the mathematical tools of topological group theory, which have been extremely fertile in the study of continuous symmetries. As an ancillary advantage, the family of algorithms we describe will be independent of the representation of the state space, by construction. This avoids the problems noted above with algorithms that depend crucially on a favorable choice of parametrization.

The idea of using groups has been intensely studied in statistics (Eaton, 1989; Wijsman, 1990; Diaconis, 1988) and groups have also found their way to machine learning (Smola and Kondor, 2003; Kondor, 2008). Model symmetries are also exploited in the body of work on "lifted" probabilistic inference; these symmetries can be encoded by groups (Niepert, 2012a) and the problem has been approached with MCMC techniques (Niepert, 2012b). The focus of that work is on performing inference on a reduced space that collapses equivalence classes. Thus "lifting" is not applicable when the symmetries are approximate or when different symmetries apply to each sub-problem: then the representation cannot be reduced and states must be explored that are symmetric for one sub-problem but not for another. To the best of our knowledge the closest work to ours is that of Liu and colleagues (Liu and Wu, 1999; Liu and Sabatti, 2000; Liu, 2004), where the primary concern is generalizing Gibbs sampling so that it can work with group transformations (the main problem being the derivation of the right "conditional" distribution over the set of transformations considered). However, as in general in Gibbs sampling, it is left to the user to implement sampling from the derived distribution. In the present thesis, however, we start from the MH algorithm, giving the user the freedom to choose an easy to sample distribution over the transformations.

2

GROUPS AND HOMOGENEOUS SPACES

To reason about the symmetries of an object, we must first consider the set of transformations that leave it unchanged. We will find it natural to organize these transformations into a mathematical group:

DEFINITION 1: A *group* G is a set of elements along with an associative binary operation on them (which is not necessarily commutative). The group operation is conventionally written as multiplication or by juxtaposition: for any $a, b \in G$ we write $a \cdot b$ or $ab \in G$. Moreover, a group must have a *unit* element $e \in G$ such that $eg = ge = g$ for all $g \in G$. Each element $g \in G$ must also have a corresponding *inverse* element $g^{-1} \in G$ such that $g^{-1}g = gg^{-1} = e$.

Every group has a unique unit and every element has a unique inverse, so the notation is not ambiguous. For our purposes, the groups will comprise transformations of the state space of the Markov chain, with the unit being the identity transformation and the group operation being the composition of transformations. We then say that a group *acts* on the state space:

DEFINITION 2: The *action* of a group G on a set W is a function $T : G \times W \to W$ that has the following properties: $T(e, w) = w$ and $T(gh, w) = T(g, T(h, w))$ where $w \in W$, $g, h \in G$, and e is the unit of G.[1]

As a notational convenience, we will often write $gw := T(g, w)$ in any context where the group action is specified and it is clear that $g \in G$ and $w \in W$. We can even write ghw where $g, h \in G$ and $w \in W$, knowing that $(gh)w = g(hw)$ by the definition of an action.

1 These are left group actions; we can analogously define right group actions having the property $T(gh, w) = T(h, T(g, w))$ instead. We will only consider left group actions unless otherwise stated.

We will assume that the state space W is a *topological space*[2] and that the group action respects the topology: for all $g \in G$, the transformations of the state space $T_g := T(g, \cdot)$ are continuous[3] functions. By the properties of group actions, T_e is the identity function and $T_g \circ T_{g^{-1}} = T_{gg^{-1}} = T_e$ for all $g \in G$, so each T_g has the continuous inverse $T_{g^{-1}}$: it is a *homeomorphism*[4] from W to itself (a self-homeomorphism). The set of self-homeomorphisms of W forms the *homeomorphism group* $\mathrm{Homeo}(W)$ under composition. The property of being a group action is then equivalent to T being a *group homomorphism* from G to $\mathrm{Homeo}(W)$:

DEFINITION 3: A *homomorphism*[5] from group G to group H is a function $f : G \to H$ that preserves units and respects the group operation: $f(e_G) = e_H$ and $f(ab) = f(a)f(b)$ for any $a, b \in G$, where e_G and e_H are the units of G and H, respectively. It follows that f preserves inverses: $f(g^{-1}) = f(g)^{-1}$.

The group action T can be considered a way to use the elements of G to index those self-homeomorphisms of W that we are interested in, with the group structure of G reflecting the composition of homeomorphisms.

To accommodate continuous symmetries (like rotation and scaling in $\mathbb{R}^2$), we will assume that G is also equipped with a topology that is respected by its group structure:

DEFINITION 4: A *topological group* is a group that is also a topological space, with the group operation $(\cdot) : G \times G \to G$ and inversion $(g \mapsto g^{-1}) : G \to G$ being continuous functions.[6]

The group action must simultaneously respect the topologies of G and W:

2 A topological space is a set of points equipped with a *topology:* a class of subsets, called *open sets*, which includes the empty set and the space itself, and is closed under unions and finite intersections. The complements of open sets are called *closed sets*. A metric space like $\mathbb{R}^n$ is usually equipped with the *metric topology*, in which the open sets are the unions of open balls.

3 A function $f : A \to B$ between topological spaces is *continuous* if $f^{-1}(E) \subset A$ is open whenever $E \subset B$ is open, or equivalently if $f^{-1}(E)$ is closed whenever E is closed.

4 A homeomorphism between topological spaces A and B is a continuous function $f : A \to B$ that has a continuous inverse $f^{-1} : B \to A$.

5 Not to be confused with *homeomorphism*.

6 The Cartesian product $A \times B$ of two topological spaces is usually equipped with the *product topology*, with the open sets being unions of products of the open sets of A and B.

DEFINITION 5: A *continuous group action* of the topological group G on the topological space W is a group action $T : G \times W \to W$ that is also a continuous function. Then we say that W is a *G-space under T*.

Note that since $T(g, w)$ is continuous in both g and w simultaneously, T_g must be continuous for all $g \in G$; the converse is not necessarily true: for T to be continuous it is not sufficient for T_g to be continuous for all g. Intuitively, "small" changes to g must produce "similar" continuous transformations T_g.

2.2 ORBITS AND TRANSITIVE ACTIONS

For any point $w \in W$, the image of G under the function $U_w := T(\,\cdot\,, w)$ is the region of the state space that is accessible from w under the action of elements of G:

DEFINITION 6: The *orbit* of any point $w \in W$ under the action of a group G is the set $Gw := \{gw \in W \mid g \in G\}$.

Claim 1. The state space is partitioned by the orbits of the points in it, so that any two points are in the same partition when some group element transforms one to the other.

Proof. Every point $w \in W$ belongs to its own orbit Gw, since $ew = w$. Suppose two orbits overlap and $u \in Gv \cap Gw$ for some $u, v, w \in W$. In other words, $u = gv = hw$ for some $g, h \in G$, and so $Gu = Ggv = Gv$; at the same time, $Gu = Ghw = Gw$, so $Gv = Gw$. We have therefore shown that the orbits cover W and are either disjoint or equal, making them a partition of W. □

If the structure of G is rich enough, then any point of the state space can be transformed to any other by the action of some group element. Then all the points in the state space belong to the same orbit:

DEFINITION 7: The action T of group G on space W is *transitive* if for every $v, w \in W$ there is some $g \in G$ such that $T(g, v) = w$.

Claim 2. T is transitive if and only if the function $U_w := T(\,\cdot\,, w)$ is surjective on W for some $w \in W$ (and hence for all $w' \in W$).

Proof. If U_w is surjective, then for any $w', v \in W$ there are some $g, h \in G$ such that $gw = w'$ and $hw = v$. Then $hg^{-1}w' = hg^{-1}gw = hw = v$, so $U_{w'}$ is surjective for all $w' \in W$. It follows from the definition that T is transitive, as does the converse result. □

Even if G does not act transitively on W, it does on each of the orbits Gw: they satisfy the requirements of the following claim (since $GGw = Gw$) and the transitivity holds by definition.

Claim 3. T is a group action of G by restriction on any $V \subset W$ that is *stable* under G (which means that it satisfies $T(G, V) = V$). Moreover, if T is continuous and V has the subspace topology[7] then V is a G-space under T.

Proof. The group action axioms are satisfied trivially. To show that V is a G-space under $T|_{G\times V}$, suppose $U \subset V$ is open. Since V has the subspace topology, $U = V \cap U'$ for some open $U' \subset W$. It follows that $T|^{-1}_{G\times V}(U) = (G \times V) \cap T^{-1}(U')$ is open in the subspace topology of $G \times V$ because, since T is continuous, $T^{-1}(U')$ is open as a subset of $G \times W$. Thus $T|_{G\times V}$ is continuous as a function with co-domain V. □

2.3 Isomorphisms of groups and G-spaces

Recall that a *group homomorphism* is a function between groups that preserves their structure; if it is invertible, then it sets up a one-to-one correspondence between their elements: both groups must then have essentially the same structure.

Definition 8: An *isomorphism* between groups G and H is a homomorphism $f : G \to H$ that has an inverse $f^{-1} : H \to G$, which is then also a homomorphism. If the groups are topological, then an isomorphism between them is additionally required to be a homeomorphism: f and f^{-1} must be continuous.

An *automorphism* is an isomorphism between a group G and itself; these form the group $\mathrm{Aut}(G)$ under composition.

7 A subset V of a topological space W has the *relative* or *subspace topology* if $U \subset V$ is open just when $U = U' \cap V$ for some open $U' \subset W$; this will be our presumed topology for all subsets of topological spaces.

We will find use for automorphisms of G of the form $g \mapsto g^h := h^{-1}gh$, where $h \in H$. These are indeed automorphisms: they are continuous and satisfy $e^h = e$ and $(ab)^h = a^h b^h$, and have continuous inverses $g \mapsto g^{h^{-1}} := hgh^{-1}$.

DEFINITION 9: An automorphism of G the form $g \mapsto g^h := h^{-1}gh$ with $h \in G$ is called an *inner automorphism* or *conjugation by h*.

We now turn from groups to G-spaces, whose structure comes from the action of G upon them. Thus, to identify two G-spaces with each other, there must be continuous maps between them that preserve this structure:

DEFINITION 10: Let V and W be G-spaces under the actions S and T respectively. An *isomorphism of G-spaces* from V to W is a homeomorphism $f : V \to W$ that satisfies $f(gv) = gf(v)$ (i.e. $f(S(g, v)) = T(g, f(v))$) for any $g \in G$ and $v \in V$. It follows that $f^{-1}(gw) = gf^{-1}(w)$ for any $w \in W$.

We can verify that G itself is a transitive G-space under group multiplication on the left (i.e. under the canonical action $(g, h) \mapsto gh$). A natural question is whether every other transitive G-space W can be identified with G as such, by constructing an isomorphism between them.

For instance, we can fix some $w \in W$ as the "origin" and consider the map $U_w := T(\,\cdot\,, w)$ from G to W, which associates each transformation $g \in G$ with the image $gw \in W$ of the origin under it. It inherits continuity from T and it is surjective because we assumed that T is transitive (for every $v \in W$ there is some $g \in G$ such that $U_w(g) = gw = v$).

In general, however, U_w may not be injective: there may be several elements of G that map w to any given v. If we assume for a moment that U_w is indeed injective, then it is a continuous bijection from G to W. If we also assume that U_w is an open[8] map, then U_w^{-1} is continuous and so U_w is a homeomorphism from G to W. Additionally, we can verify that it respects the G-space structure of G and W by being an isomorphism of G-spaces.

Thus any transitive G-space W can indeed be identified with G itself as a G-space in the special case when U_w is injective and open. In the following sections, we will attempt to construct a similar isomorphism when U_w is not necessarily injective.

---

8 A function $f : A \to B$ is *open* if $f(E) \subset B$ is open whenever $E \subset A$ is open. *Closed* functions are defined analogously.

We observed in the previous section that in general, for any given $v, w \in W$, there may be several elements $g \in G$ satisfying $gw = v$: the map $U_w := T(\cdot, w)$ may not be injective from G to W. As we did there, we will fix some $w \in W$ as the "origin" and characterize the non-injectivity of U_w by first studying the set of group elements that map w to itself:

DEFINITION 11: Given a group G that acts on a set W, the *stabilizer* of any $w \in W$ is defined as $G_w := \{g \in G \mid gw = w\}$.

G_w contains the unit of G (since $ew = w$) and is closed under inversion (if $gw = w$ then $g^{-1}w = w$) and multiplication (if $gw = hw = w$ then $ghw = w$). In other words, any stabilizer G_w is a subgroup of G.

DEFINITION 12: A *subgroup* H of a group G is a subset that is itself a group under the same operation as G. If G is a topological group and H has the subspace topology then it is a *topological subgroup*.

Now select any $g \in G$ that transforms w to $v := gw$. Then any element of the set $gG_w := \{gh \mid h \in G_w\}$ also transforms w to v, since $ghw = gw = v$ for any $h \in G_w$. The converse is also true:

Claim 4. Any element $a \in G$ that transforms w to v can be written as gh for some $h \in G_w$ and therefore belongs to the set gG_w.

Proof. We see that $gw = aw$ if and only if $g^{-1}aw = w$. We can take $h := g^{-1}a \in G_w$ so that $gh = gg^{-1}a = a$. □

To summarize, the set $U_w^{-1}(v)$ (consisting of elements of G that transform w to v) is exactly gG_w for any $g \in U_w^{-1}(v)$. Thus $U_w(a) = U_w(b)$ if and only if $b \in aG_w$, which is to say, just when $a^{-1}b \in G_w$. In other words, the stabilizer subgroup G_w measures the non-injectivity of U_w: it is injective if and only if G_w is the trivial subgroup containing only the unit element.

Remark 1. Varying $w \in W$ results in different G_w. However, points in the same orbit have stabilizer subgroups with essentially the same structure (i.e., they are isomorphic groups; see Section 2.3). Indeed, the operation of *conjugation* (Definition 9 on page 10) transforms one to the other: if $h \in G_{gw}$ for some $g \in G$, then

$g^{-1}hgw = g^{-1}gw = w$, so $h^g := g^{-1}hg \in G_w$; the map $h \mapsto h^g := g^{-1}hg$ is an isomorphism of topological groups from G_{gw} to G_w. All topological and group properties of G_{gw} can thereby be transferred to G_w using conjugation by g.

2.5 COSETS AND QUOTIENT SPACES

In the previous section, we made use of sets of the form gG_w, where G_w is the stabilizer subgroup of w in G. We will now explore some properties of these sets for any arbitrary subgroup H of G.

DEFINITION 13: Given a group G, a *left coset* of any subgroup H is a set $gH := \{gh \mid h \in H\}$, where the element $g \in G$ is called a *coset representative*. A *right coset* is a set Hg.

The set of all left cosets of H is called its *left quotient space* and written G/H. Its *right quotient space* is defined analogously and written $H\backslash G$.

The left cosets of any subgroup H form a partition of the group G. Each element $g \in G$ belongs to the coset gH, because $e \in H$. Given two cosets aH and bH, they are either disjoint or identical[9]; in the latter case, it is necessary and sufficient that $a^{-1}b \in H$. Corresponding properties hold for right cosets, but we will only use left cosets and left quotient spaces unless otherwise stated.

It follows that the condition $a^{-1}b \in H$ is an equivalence relation on the elements of G, written $a \equiv b \pmod{H}$, whereby two elements are equivalent if they belong to the same coset. The quotient space G/H of cosets can be seen as a set of equivalence classes under this relation and equipped with the *quotient topology*, which is obtained from the topology of G by "collapsing" equivalent points.

DEFINITION 14: The *canonical quotient map* of any quotient space G/H is the surjective map $q : G \to G/H$ defined by $g \mapsto gH$, which assigns every element of G to its containing coset.

The *quotient topology* is the "finest" topology on G/H that makes q continuous: a subset $V \subset G/H$ is considered open if and only if $q^{-1}(V)$ is open in the topology of G.

9 The left cosets gH are the orbits of the elements of G with H acting on them by *right* multiplication (under the map $(h, g) \mapsto gh$). These properties therefore follow from our earlier discussion of orbits in Section 2.2.

Apart from being continuous, a quotient map has the following *universal property:* a function $f : G/H \to X$ from the quotient space to any topological space is continuous if and only if $f \circ q : G \to X$ is continuous. Corollary 2.5.2 strengthens this property using the topological group structure of G.

2.5.1 PROPOSITION: *Let G be a topological group and H a topological subgroup. Then the canonical quotient map $q : G \to G/H$ defined by $g \mapsto gH$ is an open map under the quotient topology on G/H.*

Proof. Given an open subset $E \subset G$, we must prove that $q(E)$ is open. By the definition of the quotient topology, a subset of G/H is open if and only if its pre-image under q is open in G. Therefore, if we show that $q^{-1}(q(E))$ is open in G, then it will follow that $q(E)$ is open in G/H.

q maps each element $g \in G$ to its coset $gH \in G/H$, so the pre-image of any coset is $q^{-1}(gH) = gH$ (where $gH := \{gh \in G \mid h \in H\}$ is considered a subset of G as well as an element of G/H). The image of E under q is $q(E) = \{gH \in G/H \mid g \in E\}$ and so $q^{-1}(q(E)) = \bigcup_{g \in E} gH = EH$. It only remains to show that EH is open.

EH can also be decomposed into a union as $EH = \bigcup_{h \in H} Eh$. The set Eg is open for any $g \in G$ because it is the pre-image of the open set E under the continuous map $h \mapsto hg^{-1}$. Thus EH is open as we wanted, being a union of open sets. □

2.5.2 COROLLARY: *A function $f : G/H \to X$ from a quotient space to any topological space is open (respectively, continuous) if and only if $f \circ q : G \to X$ is open (respectively, continuous).*

Proof. The claim about continuity is the previously stated universal property of the quotient map for any space with a quotient topology. We need only prove the claim about openness, which relies on G being a topological group.

If f is open, then $f \circ q$ is open because we showed that q is also open. Conversely, suppose $f \circ q$ is open and V is an open subset of G/H. Since q is surjective, it has a right inverse and $q(q^{-1}(V)) = V$. Since q is continuous, $q^{-1}(V)$ is an open subset of G. Thus $f(V) = (f \circ q)(q^{-1}(V))$ is an open subset of X when $f \circ q$ is an open map. □

2.5.3 COROLLARY: *The quotient space G/H is a transitive G-space under the canonical action $T(g, aH) = gaH$.*

Proof. G itself is a transitive G-space under the group multiplication map $(g, a) \mapsto ga$, and $ga \equiv gb \pmod{H}$ if and only if $a \equiv b \pmod{H}$. The G-space structure of G can therefore be "lifted" to the quotient space G/H. We will use the fact that $G \times G/H$ is homeomorphic to the quotient of the product group[10] $G \times G$ by the subgroup $\{e\} \times H$, so that

$$\begin{aligned} r : G \times G &\to G \times G/H \cong (G \times G)/(\{e\} \times H) \\ (g, a) &\mapsto (g, aH) \qquad \equiv \{g\} \times aH \end{aligned}$$

is a quotient map.

Take $S(g, a) := ga$, so that the map $(g, a) \mapsto gaH$ from $G \times G$ to G/H can be written either as $q \circ S$ or as $T \circ r$. It follows that T is well-defined[11] (since r is a quotient map and $q \circ S$ is well-defined) and that it is continuous (from the universal quotient property of r because $q \circ S$ is continuous).

T is also a group action: $T(ab, gH) = abgH = T(a, T(b, gH))$ and $T(e, gH) = egH = gH$. Since S is transitive (for every $a, b \in G$ there is $ba^{-1} \in G$ such that $(ba^{-1})a = b$) it follows that T is transitive (since $(ba^{-1})aH = bH$ for any cosets $aH, bH \in G/H$). □

2.6 HOMOGENEOUS SPACES

In Section 2.3 we showed that when W is a transitive G-space under the action T, the map $U_w := T(\,\cdot\,, w)$ is an isomorphism of G-spaces from G to W if it is injective and open. In Section 2.4 we characterized the non-injectivity of U_w by the stabiliser subgroup $G_w := \{g \in G \mid gw = w\}$. In Section 2.5 we "collapsed" cosets of subgroups to form quotient spaces, which were themselves transitive G-spaces under a canonical action (Corollary 2.5.3). We will now continue to characterize

10 The product of two (topological) groups $G \times H$ is itself a (topological) group with the operation being element-wise: $(g, h) \cdot (g', h') = (g \cdot g', h \cdot h')$ with $(g, h)^{-1} = (g^{-1}, h^{-1})$.

11 A function on cosets $f(\ldots, gH, \ldots)$ is said to be *well-defined* if its value depends only on the coset gH and not on the particular choice of representative g.

the structure of G-spaces by finding conditions under which they are isomorphic to appropriate quotient spaces:

DEFINITION 15: A G-space is *homogeneous* if it is isomorphic to a quotient space G/H with the canonical action $(g, aH) \mapsto gaH$, where H is a topological subgroup of G.

Our isomorphisms will be the maps $\phi_w : G/G_w \to W$ defined by $gG_w \mapsto gw$.

2.6.1 LEMMA: *If W is a transitive G-space, then ϕ_w is a well-defined continuous bijection from G/G_w to W. Additionally, it is a homeomorphism if and only if U_w is an open map.*

Proof. We showed that $aw = bw$ if and only if $a^{-1}b \in G_w$, or equivalently, $U_w(a) = U_w(b)$ if and only if $a \equiv b \pmod{G_w}$. Since $\phi_w(gG_w) := U_w(g)$, it follows that ϕ_w is well-defined (since U_w is constant on each coset of G_w) and injective (since U_w is distinct on different cosets of G_w).

By the transitivity property, U_w is surjective. Since $U_w = \phi_w \circ q$, its surjectivity is inherited by ϕ_w. From Corollary 2.5.2 we see that the continuity of U_w is also inherited by ϕ_w and, additionally, that ϕ_w is open (which is necessary and sufficient for it to be a homeomorphism) if and only if U_w is open. □

It is important to note that the bijection ϕ_w depends on the choice of the origin $w \in W$. The requirement that U_w be open for ϕ_w to be a homeomorphism does not limit the choice of w:

Claim 5. If U_w is an open map for some $w \in W$ then it is also open for any other $w' \in W$.

Proof. Since W is a transitive G-space, there is some $g \in G$ such that $gw = w'$. Suppose $V \subset G$ is open and $U_{w'}(V) = Vw' = Vgw$. Now Vg is open because it is the inverse image of the open set V under the continuous map $h \mapsto hg^{-1}$. Thus $U_w(Vg) = Vgw$ is an open set if U_w is open. □

This claim, combined with Claim 2 on page 8, means that if W is a G-space under T and if $T(\,\cdot\,, w)$ is surjective and open for some $w \in W$, then it is surjective and open for any other $w' \in W$. This is, in fact, all that is required for a G-space to be homogeneous:

2.6.2 PROPOSITION: *A G-space W is homogeneous if and only if it is transitive and U_w is an open map. Then ϕ_w is an isomorphism of G-spaces.*

Proof. To show the necessity of the condition, let W be a homogeneous G-space and identify it with the quotient space G/H under the canonical action for some subgroup $H \subset G$. Then W is transitive by Corollary 2.5.3 and the map $T(\,\cdot\,, eH)$ (given by $g \mapsto gH$) is open by Proposition 2.5.1; it follows from Claim 5 on page 15 that $T(\,\cdot\,, aH)$ is open for all $aH \in G/H$.

Conversely, if W is a G-space and U_w is open, then ϕ_w is a well-defined continuous bijection by Lemma 2.6.1; it is also an isomorphism of G-spaces: $h\phi_w(gG_w) = \phi_w(hgG_w) = hgw$. □

2.7 SOME TOPOLOGICAL CONSEQUENCES

The topology of a topological group, since it is interwoven with its group structure, gains several strong topological properties. We will describe some of the more useful ones in this section. Recall some preliminary definitions from topology: A *neighborhood* of a point in a topological space is any set with an open subset containing the point. A subset of a topological space is *compact* if every collection of open sets whose union contains that set (an *open cover*) has a finite sub-collection that is also an open cover. If the whole topological space is compact, it is called a *compact space*. On the other hand, a space may be only *locally compact*, if every point has a compact neighborhood. A topological space may satisfy several *separation axioms:* It is a T_0 (*Kolmogorov*) space if, given any two distinct points, there is a neighborhood of one that doesn't include the other (i.e., all points are *topologically distinguishable*). A space is T_1 (*Fréchet*) if any set containing only a single point is closed. A space is T_2 (*Hausdorff*) if any two distinct points have disjoint neighborhoods.

2.7.1 LEMMA: *Let G be a topological group and U an open neighborhood of the unit element e. Then U contains an open neighborhood V that includes e and satisfies $V^{-1}V \subset U$, where $V^{-1}V := \{v^{-1}v' \mid v, v' \in V\}$.*

Proof. Let E be the pre-image of U under the continuous map $(g,h) \mapsto g^{-1}h : G \times G \to G$, so that $(e,e) \in E$ maps to $e \in U$; since U is open, E must also be open.

The product topology on $G \times G$ is generated by basis sets of the form $V_0 \times V_1$, where V_0 and V_1 are open subsets of G. Thus there must be some open V_0 and V_1 such that $(e, e) \in V_0 \times V_1 \subset E$. We can now take $V := V_0 \cap V_1$; it is open, contains e, and has the property that $V \times V \subset V_0 \times V_1 \subset E$, so $V^{-1}V \subset U$. $\square$

2.7.2 PROPOSITION: *Let G be a topological group and H a subgroup. Then the quotient space G/H is Hausdorff if and only if H is closed.*

Proof. If the quotient space is Hausdorff, it is also T_1: regarding H as a point in G/H, the set $\{H\} \subset G/H$ is a singleton and thus closed, and its pre-image under the quotient map is $H \subset G$, which must therefore also be closed. Conversely, suppose H is closed and aH and bH are distinct, so that $e \notin aHb^{-1}$ (because $a^{-1}b \notin H$) and aHb^{-1} is closed (being the pre-image of H under $g \mapsto a^{-1}gb$). Applying Lemma 2.7.1 gives us an open neighborhood V containing e such that $V^{-1}V$ is disjoint from aHb^{-1}. Then VaH and VbH are disjoint open neighborhoods in G/H of aH and bH, respectively: G/H is thus Hausdorff. $\square$

2.7.3 PROPOSITION: *For topological groups, the T_0, T_1, and T_2 conditions are equivalent.*

Proof. Suppose G is a T_0 topological group and $g \in G$. If $e \neq g$, there must be an open set U_g containing g but not e. Indeed, otherwise, by the T_0 property there is an open set V_g containing e but not g. Then take $U_g = gV_g^{-1} := \{gv^{-1} \mid v \in V_g\}$: it is open, since it is the pre-image of V_g under the continuous map $h \mapsto h^{-1}g$, and it doesn't contain e since that would imply $g \in V_g$; then U_g is the required neighborhood of g. Thus every point in $G \setminus \{e\}$ has an open neighborhood, making $\{e\}$ closed; indeed, any singleton $\{g\}$ is closed, being the pre-image of $\{e\}$ under $h \mapsto g^{-1}h$. In other words, G is T_1.

If G is T_1, then the trivial subgroup $\{e\}$ is closed: $G \cong G/\{e\}$ must therefore be Hausdorff by Proposition 2.7.2. Finally, any Hausdorff space is T_0 by definition. We have therefore shown the required equivalence. $\square$

2.8 INVARIANT AND RELATIVELY INVARIANT MEASURES

The Lebesgue measure on $\mathbb{R}^n$ has the defining property that it is *invariant* under translation: if $E \subset \mathbb{R}^n$, $c \in \mathbb{R}^n$, and λ is the Lebesgue measure, then $\lambda(E) =$

$\lambda(E+c)$, where $E+c$ denotes $\{e+c \mid e \in E\}$. The Lebesgue measure is, in fact, a special case of invariant measures on more general topological groups:

DEFINITION 16: A measure μ on a group G is called *left invariant* or a *left Haar measure* if it is invariant under the operation of left multiplication:

$$\mu(gH) := \mu(\{gh \mid h \in H\}) = \mu(H)$$

for any $g \in G$ and measurable $H \subset G$. *Right invariant* or *right Haar measures* are defined analogously.

Just as the Lebesgue measure is a natural choice as a reference measure for probability distributions on $\mathbb{R}^n$, we will use distributions on topological groups that have densities with respect to Haar measures. This section sketches the major results we will need for the rest of this thesis; a more complete development may be found in Nachbin (1965); Wijsman (1990); Bourbaki (2004).

2.8.1 PROPOSITION (Haar, 1933; Weil, 1940; Cartan, 1940): *On every locally compact Hausdorff*[12] *topological group G (equipped with the Borel σ-algebra*[13]*) there exists a left invariant measure $\mu \neq 0$ satisfying:*

(i) $\mu(K) < \infty$ *for any compact* $K \subset G$,

and for any measurable $H \subset G$

(ii) outer regularity: $\mu(H) = \inf\{\mu(U) \mid \textit{open } U \supset H\}$,

(iii) inner regularity: $\mu(H) = \sup\{\mu(K) \mid \textit{compact } K \subset H\}$.

Such a measure is unique except for a strictly positive factor of proportionality; that is, if $\nu \neq 0$ is another left invariant measure on G satisfying these conditions, there exists a real number $c > 0$ such that $\nu = c\mu$.

An analogous proposition holds for right Haar measures; if μ is a left Haar measure, then $\nu(H) := \mu(H^{-1})$ is a right Haar measure, where $H^{-1} := \{h^{-1} \mid h \in H\}$: indeed, $\nu(Hg) = \mu(g^{-1}H^{-1}) = \mu(H^{-1}) = \nu(H)$ for any $g \in G$ and

12 A topological group need not be Hausdorff, but many authors define it to be: this gives the Haar measure its extra properties beyond left invariance; it is a mild condition (Proposition 2.7.3).

13 The Borel σ-algebra on a topological space contains exactly those sets that can be formed from open (equivalently, closed) sets by complements and countable unions and intersections.

measurable $H \subset G$. While we will not explicitly need right Haar measures, we will find use for the Radon-Nikodym derivative $d\mu/d\nu$. Before we proceed, note that $E \mapsto \mu(Eg)$ is another left Haar measure (since $hE \mapsto \mu(hEg) = \mu(Eg)$); it is therefore proportional to μ:

DEFINITION 17: The *right-hand modulus* $\Delta_r^G : G \to \mathbb{R}_+$ of a topological group G is defined by

$$\mu(Hg) = \Delta_r^G(g)\, \mu(H)$$

for any left Haar measure μ and measurable $H \subset G$. The *left-hand modulus* Δ_l^G is defined analogously.

One can show that the left- and right-hand moduli are continuous group homomorphisms from G to $\mathbb{R}_+^\times$, the group of positive real numbers under multiplication: $\Delta_r^G(gh) = \Delta_r^G(g) \cdot \Delta_r^G(h)$ and $\Delta_r^G(g^{-1}) = 1/\Delta_r^G(g)$ for any $g, h \in G$; the same holds for Δ_l^G. In fact, Δ_r^G is the Radon-Nikodym derivative $d\mu/d\nu$ (see Nachbin, 1965, Chapter 2, Propositions 7 and 8) and thus $(d\nu/d\mu)(g) = 1/\Delta_r^G(g) = \Delta_r^G(g^{-1})$:

$$\int f\, d\nu := \int f(g^{-1})\, \mu(dg) = \int f(g)\Delta_r^G(g^{-1})\, \mu(dg)\,. \tag{2.8.1}$$

It is now clear that $\Delta_r^G \cdot \Delta_l^G = 1$ identically and that Δ_r^G and Δ_l^G do not depend on the choice of the left Haar measure μ; indeed, choosing any $c\mu$ instead would produce the corresponding right Haar measure $c\nu$ and the Radon-Nikodym derivative would remain unchanged. The left- and right-hand moduli are therefore unique and characteristic of a topological group. In many cases $\Delta_r^G = \Delta_l^G = 1$ identically: if G is discrete, or commutative (abelian), or its topology is compact, for example; then G is said to be *unimodular* and the left and right Haar measures coincide.

Proposition 2.6.2 showed us that the G-space structure of a group could be carried over to any homogeneous space acted upon by the group; the next step is to carry over the Haar measure on G to give an "invariant" measure on W. We will, however, be satisfied with a weaker notion of invariance:

DEFINITION 18: If a group G acts on a set W, then a measure λ on W is said to be *χ-relatively invariant* for $\chi : G \to \mathbb{R}_+^\times$ if

$$\lambda(gV) = \chi(g)\, \lambda(V)$$

for any measurable $V \subset W$ and $g \in G$; then χ is called the *modulus* of λ and is a continuous group homomorphism.[14]

Note that an invariant measure is a special case of a relatively invariant measure with modulus $\chi = 1$ identically. One may view the left Haar measure on G as being invariant under the action of left multiplication and Δ_r^G-relatively invariant under the action of right multiplication (which is, of course, a *right* group action). The following result gives conditions for the existence of relatively invariant measures:

2.8.2 PROPOSITION (Weil, 1940): *Let G be a locally compact Hausdorff topological group, H a subgroup, and $\chi : G \to \mathbb{R}_+^\times$ a continuous group homomorphism. In order that there exist a χ-relatively invariant measure λ on the homogeneous space G/H under the action of G to the left, it is necessary and sufficient that $\Delta_r^H(h) = \chi(h)\,\Delta_r^G(h)$ for every $h \in H$. Then λ is unique up to a strictly positive factor of proportionality.*

The relatively invariant measure on G/H is constructed as a *quotient measure* of the Haar measures on G and H; while we do not present the details here, we do use the quotient measure construction in the proof of Lemma 3.4.4.

We will henceforth assume that the state space of our Markov chain possesses a measure that is relatively invariant under the action of one or more groups, possibly with different moduli. We will not, however, explicitly assume that it is a homogeneous space; the preceding proposition shows when such relatively invariant measures may be constructed on homogeneous spaces. In our proofs, the chief manifestation of the χ-relative invariance of a measure λ on W under the action of G will be in simplifying the integral of a function $f : W \to \mathbb{R}$ as follows:

$$\int f(gw)\,\lambda(dw) = \chi(g^{-1}) \int f(w)\,\lambda(dw), \qquad g \in G. \tag{2.8.2}$$

One can verify this result when $f(w)$ is the characteristic function $\mathbb{1}\{w \in V\}$ of some measurable $V \subset W$; then the characteristic function of $g^{-1}V$ is $\mathbb{1}\{w \in g^{-1}V\} = \mathbb{1}\{gw \in V\} = f(gw)$.

Remark 2. We close this section with an observation about continuous group homomorphisms $\phi : G \to \mathbb{R}_+^\times$ such as the moduli of (relatively) invariant measures.

14 That the modulus is a homomorphism follows from a simple computation; that it is continuous is the subject of Proposition 7.2.2 of Wijsman (1990).

In particular, if H is a compact subgroup of G, then $\phi(H)$ must be a compact subgroup of $\mathbb{R}_+^\times$. The only compact subgroup of $\mathbb{R}_+^\times$, however, is the trivial subgroup $\{1\}$: indeed, if $1 \neq x \in \phi(H)$, then without loss of generality $x > 1$ (otherwise take $1/x$) and $x^n \in \phi(H)$ for all $n \in \mathbb{N}$, so $\phi(H)$ is unbounded; the compact sets in $\mathbb{R}_+$, however, are exactly those that are closed and bounded. In other words, $\phi|_H = 1$.

Since the left- and right-hand moduli of Haar measures are just such continuous group homomorphisms, this observation justifies our earlier statement that all compact groups are unimodular. We also see that any homogeneous space G/H with compact H must necessarily have a relatively invariant measure: $\Delta_r^H = \Delta_r^G = \chi = 1$ identically on H, so the requirement of Proposition 2.8.2 is trivially satisfied.

3

METROPOLIS-HASTINGS WITH GROUP TRANSFORMATIONS

A Markov Chain Monte Carlo (MCMC) algorithm to sample from a probability distribution P over a state space W is specified by a *transition kernel* $Q(dw' \mid w)$, which gives rise to a Markov chain $U_0, U_1, U_2, \ldots$ where U_0 is sampled according to some initial distribution P_0 and each U_i after that is sampled according to $Q(\,\cdot \mid U_{i-1})$. Under appropriate conditions on Q, the random variables U_n converge in distribution to P as $n \to \infty$; P is then called a steady state distribution of the Markov chain. A convenient condition to force P to be a steady state distribution of Q is *detailed balance:*

$$P(du)\, Q(dv \mid u) = P(dv)\, Q(du \mid v); \tag{3.0.1}$$

the Markov chain is then said to be *reversible*. Indeed, the meaning of (3.0.1) is that if (U, V) is sampled from the joint in (3.0.1) then we cannot tell whether (U, V) was generated by first choosing U from P and then following Q to generate V, or whether it was generated by first choosing V from P and then following Q to generate U. Under additional conditions on Q, such as Q being ϕ-irreducible and aperiodic, P is the unique steady state distribution of Q and the Markov chain (U_i) sampled from Q will indeed converge in distribution to P regardless of P_0 (see, e.g., Roberts and Rosenthal, 2004, Theorem 4).

The MH algorithm is one way to construct reversible transition kernels: given a *proposal kernel* $Q'(dw' \mid w)$, the MH kernel first samples U'_{n+1} according to $Q'(\,\cdot \mid U_n)$ and then *accepts* U'_{n+1} as U_{n+1} with probability $\alpha(U_n, U'_{n+1})$; otherwise U_{n+1} is taken to be U_n. With an appropriate choice of the *acceptance probability* function $\alpha : W \times W \to [0, 1]$, the MH transition kernel satisfies detailed balance (Tierney, 1998). However, we will call any transition kernel obtained via the above procedure an MH transition kernel regardless of whether it satisfies detailed balance or whether its stationary distribution matches the target distribution.

Building upon the mathematical setting outlined in Chapter 2, we will take W to be a topological space and G a topological group acting continuously upon it (i.e.,

W is a G-space). The MH proposal kernel Q' will sample an element according to a distribution over G and propose the state that results from its action on the current state of the chain. The group structure of G will ensure that the composition of transformations of the state space is expressible as the action of an element of G, as is the inverse of any transformation: this latter property will allow us to reason about the reversibility of the Markov chain. The topologies of W and G will reflect the continuity of the state space and its transformations; the discrete topology may be used when suitable.

Working in this general setting will allow our algorithm and its correctness results to rely only on the operational notion of transforming the state space in certain ways, and the resulting algorithms will remain unchanged under different parametrizations of the state space. The state representation can be chosen freely, guided only by practical implementation concerns. However, as a guide to intuition, the reader can imagine the state space W to be a subset of the Euclidean space $\mathbb{R}^n$ using an arbitrary choice of parametrization. The group G can be taken to be the invertible continuous maps, or even just the invertible affine transformations. One must only keep in mind that an algorithm constructed under these restrictions must be explicitly proven to be invariant under re-parametrization; it is not automatically invariant by construction as in the general setting we adopt.

3.1 METROPOLIS-HASTINGS BASED ON GROUP MOVES

The proposal kernel can be defined in terms of a conditional distribution $Q_G(dg \,|\, w)$ over the group G; it samples $g \sim Q_G(\,\cdot\,|\, w)$ and proposes the new state gw. We will assume that the proposal distribution Q_G has the density q with respect to the Haar measure μ on G and that the target distribution has a density p with respect to a measure λ on W which is χ-relatively invariant under the action of G (see Section 2.8):

$$P(dw) = p(w)\,\lambda(dw), \quad Q_G(dg \,|\, w) = q(g \,|\, w)\,\mu(dg)\,.$$

We will also assume that the initial state of the Markov chain lies within the support of P. Our MH transition kernel Q based on Q_G is defined (for w in the support of P) by the following procedure; that it is "correct" (in the sense that it is in detailed balance with P) will be the subject of Theorem 3.3.3.

PROCEDURE 1: Given the current state $w \in W$, sample the new state w' as follows:

1. Sample $g \sim Q_G(\,\cdot\,|\,w)$.
2. Calculate $\alpha := \dfrac{\chi(g)\,p(gw)\,q'(g^{-1}\,|\,gw)}{\Delta_r^G(g)\,p(w)\,q'(g\,|\,w)}$.
3. Accept $w' = gw$ with probability $\min\{1, \alpha\}$.

In the procedure we use the function q' (derived from q) to account for the possibility that many different moves $g \in G$ may result in the same w'. In particular, q' is defined as follows: Recall from Section 2.4 that for $w \in W$, the *stabilizer subgroup* $G_w := \{g \in G \,|\, gw = w\}$ measures the injectivity of the map $g \mapsto gw$: for any $g \in G$, the set of all g' that also satisfy $g'w = gw$ is exactly gG_w. Under mild conditions on the action of G on W, G_w will be seen to be compact, implying that there exists a unique Haar measure β_w on G_w with $\beta_w(G_w) = 1$. Then

$$q'(g\,|\,w) = \int_{G_w} q(gh\,|\,w)\,\beta_w(dh)\,.$$

Remark 3. It follows from this definition that $q'(\,\cdot\,|\,w)$ is constant on each gG_w. Moreover, if q itself has this property then $q' = q$.

We note, in closing, that Procedure 1 encompasses the standard MH algorithm defined for Euclidean spaces. Indeed, if the state space W and group G are both $\mathbb{R}^n$ with $gw = g + w$, without loss of generality one can rewrite the proposal in terms of the move $g = w' - w$. Then the Lebesgue measure m serves as both λ and μ. Since m is invariant, $\chi = 1$. Furthermore, since vector addition is commutative, $\Delta_r^G = 1$. Finally, for any $x, y \in \mathbb{R}^n$ there is a unique $g = y - x$ such that $x + g = y$, so $q' = q$ and

$$\alpha = \frac{p(w')\,q(w - w'|w')}{p(w)\,q(w' - w|w)}.$$

The proposal kernel described above is often too restrictive, in that Q_G may not have a density with respect to the Haar measure μ on G. For example, the Gibbs sampler on $\mathbb{R}^n$ updates the state space by modifying one coordinate at a time. Its proposal distribution is therefore concentrated on the coordinate axes (which have zero Lebesgue measure on $\mathbb{R}^n$) and so does not have a density with respect to that measure.

One way to increase flexibility is to allow several different groups $G_1, G_2, \ldots, G_n$ to act on the state space W, each associated with a kernel $Q_i(dg_i \,|\, w)$ ($i = 1, \ldots, n$). Each Q_i will be assumed to have a density q_i w.r.t. the Haar measure μ_i on G_i. We will choose λ to be a measure on W that is simultaneously relatively invariant under all the groups: χ_i-relatively invariant under each G_i, respectively. The proposal kernel Q' will be a mixture of the Q_i with coefficients $a(i \,|\, w) > 0$, $i = 1, \ldots, n$, with $\sum_{i=1}^{n} a(i \,|\, w) = 1$ for all $w \in W$. The MH transition kernel based on Q' is defined by the following procedure; that it is in detailed balance with P will be the subject of Theorem 3.3.5.

PROCEDURE 2: Given the current state $w \in W$, sample the new state w' as follows:

1. Sample $i \sim a(\,\cdot \,|\, w)$ and $g \sim Q_i(\,\cdot \,|\, w)$.
2. Calculate $\alpha := \dfrac{\chi_i(g)\, a(i \,|\, gw)\, p(gw)\, q_i'(g^{-1} \,|\, gw)}{\Delta_r^{G_i}(g)\, a(i \,|\, w)\, p(w)\, q_i'(g \,|\, w)}$.
3. Accept $w' = gw$ with probability $\min\{1, \alpha\}$.

3.3 CORRECTNESS

We will assume that the proposals are chosen in such a way that ϕ-irreducibility holds: in particular, this is easy to verify in the case of SLAM below. To prove that the MCMC transition kernels described in Procedures 1 and 2 satisfy detailed balance, we will require some technical conditions on the space W and the groups G or G_i.

3.3.1 ASSUMPTION: *The state space W and the groups G and G_i (for Procedures 1 and 2, respectively) are locally compact and Hausdorff.*[1]

The local compactness condition on the groups G and G_i guarantees the existence of the Haar measures on them. The Hausdorff property implies that every compact set in a space is also closed, and thus singleton sets are also closed.

The second assumption is designed to exclude certain pathological examples of group actions:

3.3.2 ASSUMPTION: *The actions of the groups G and G_i (for Procedures 1 and 2, respectively) on the state space W are* proper: *the map $\theta : W \times G \to W \times W$ defined by $(w, g) \mapsto (w, gw)$ preserves compactness of pre-images, so $\theta^{-1}(K)$ is compact in $W \times G$ for every compact $K \subset W \times W$.*[2]

A group G acting properly on the space W has several desirable properties. Most importantly for our immediate purposes, the stabilizer subgroups G_w of G at $w \in W$ are compact and thus also locally compact. Thus there is a finite Haar measure β_w on each G_w which, without loss of generality, is normalized: $\beta_w(G_w) = 1$.

As noted earlier, for any $g \in G$, the set of all g' that also satisfy $g'w = gw$ is exactly gG_w. Thus, if the action of G on W is proper, we are assured that the structure of G is not too rich in relation to the space it acts upon: gG_w is compact and thus not too "large". With this, we can state our first main result:

3.3.3 THEOREM: *If the state space W and group G satisfy Assumptions 3.3.1 and 3.3.2, then the Markov transition kernel defined by Procedure 1 satisfies detailed balance* (3.0.1).

To show the correctness of Procedure 2, we will need to assume that the image of w under any two G_i, G_j overlap only negligibly. To do this, we will assume that all the G_i are, in fact, subgroups of some overarching group K, so that we can define intersections of the G_i:

1 See Section 2.7; in particular Proposition 2.7.3 shows that the Hausdorff condition is very mild for topological groups.

2 Alternatively, $f : X \to Y$ is said to be proper if it is *universally closed:* $f \otimes \mathrm{id}_Z : X \times Z \to Y \times Z$ is closed for every topological space Z, on which id_Z is the identity function. Our definition coincides with this one when Assumption 3.3.1 holds.

3.3.4 ASSUMPTION: *For every* $1 \leq i, j \leq n$ *with* $i \neq j$*, the condition*

$$p(w) \int \mathbb{1}\{g \in (G_i \cap G_j) G_{k,w}\}\, q'(g \mid w)\, \mu_k(dg) = 0, \quad w \in W$$

is satisfied with either $k = i$ *or* $k = j$*, where* $G_{k,w}$ *is the stabilizer subgroup of* G_k *at* $w \in W$*.*

3.3.5 THEOREM: *If the state space* W *and each* G_i *(*$1 \leq i \leq n$*) satisfy Assumptions 3.3.1, 3.3.2 and 3.3.4, then the Markov transition kernel defined by Procedure 2 satisfies detailed balance* (3.0.1)*.*

3.4 PROOFS

We will begin by restating some results of Tierney (1998) for our own use.

3.4.1 PROPOSITION (Tierney, 1998, Proposition 1): *Let* $\mu(dx, dy)$ *be a sigma-finite measure on the product space* $(E \times E, \mathcal{E} \otimes \mathcal{E})$ *and let* $\mu^T(dx, dy) = \mu(dy, dx)$*. Then there exists a symmetric set* $R \in \mathcal{E} \otimes \mathcal{E}$ *such that* μ *and* μ^T *are mutually absolutely continuous on* R *and mutually singular on the complement of* R*,* R^C*. The set* R *is unique up to sets that are null for both* μ *and* μ^T*. Let* μ_R *and* μ_R^T *be the restrictions of* μ *and* μ^T *to* R*. Then there exists a version of the density*

$$r(x, y) = \frac{\mu_R(dx, dy)}{\mu_R^T(dx, dy)}$$

such that $0 < r(x, y) < \infty$ *and* $r(x, y) = 1/r(y, x)$ *for all* $x, y \in E$*.*

3.4.2 PROPOSITION (Tierney, 1998, Theorem 2): *A Metropolis-Hastings transition kernel satisfies the detailed balance condition* (3.0.1) *if and only if the following two conditions hold.*

(i) The function α *is* μ*-almost everywhere zero on* R^C*.*

(ii) The function α *satisfies* $\alpha(x, y) r(x, y) = \alpha(y, x)$ μ*-almost everywhere on* R*.*

The Metropolis-Hastings acceptance probability

$$\alpha(x,y) = \begin{cases} \min\{1, r(y,x)\}, & \text{if } (x,y) \in R, \\ 0, & \text{if } (x,y) \notin R. \end{cases}$$

satisfies these conditions by construction.

The image of $W \times G$ under the map $\theta : (w,g) \mapsto (w,gw)$ (see Assumption 3.3.2) is the set $E := \{(w,gw) \mid w \in W, g \in G\}$, which is closed in $W \times W$ because θ is a proper (hence closed) map and $W \times G$ is closed. If we restrict the co-domain of θ to E, it becomes a surjective, continuous, and closed map: it is a quotient map. In other words, any set $U \subset E$ is open in the subspace topology inherited by E from $W \times W$ if and only if $\theta^{-1}(U)$ is open in $W \times G$. Furthermore, θ has the following universal property: if Z is any topological space and $f : W \times G \to Z$ is a continuous function satisfying $f(w,g) = f(w',g')$ whenever $\theta(w,g) = \theta(w',g')$, then there is a unique continuous function $\bar{f} : E \to Z$ such that $f = \bar{f} \circ \theta$. We see that $\theta(w,g) = \theta(w',g')$ if and only if $w = w'$ and $g' \in gG_w$ (i.e., $gw = g'w$). The equivalence classes under θ are therefore sets of the form $\{w\} \times gG_w$. Our strategy will be to use θ to work in the space $W \times G$ as a proxy for $E \subset W \times W$, taking care to account for the non-injectivity of θ.

DEFINITION 19: Given a measurable function $f : A \to B$ and a measure ρ on A, the *push-forward measure* $f(\rho)$ is a measure on B defined by $f(\rho)(E) = \rho(f^{-1}(E))$ for any measurable $E \subset B$. Equivalently, for integrals, $\int g \, df(\rho) = \int g \circ f \, d\rho$ for any integrable $g : B \to \mathbb{R}$.

Note. From now on, the notation $f(\rho)$ always means the push-forward of ρ under f, as long as f is an $A \to B$ map and ρ is a measure on A. In particular, the parentheses in a setting like this will never be used for grouping. To help in parsing the formulas, we will also occasionally write $g \cdot \rho$ to denote the measure whose density w.r.t. ρ is g, where ρ is a measure on A and $g : A \to [0, \infty)$ is ρ-integrable.

Given a topological subgroup $H \subset G$, let β be a left Haar measure on it and $\pi : G \to G/H$ the canonical quotient map. Consider the following construction with $f : G \to \mathbb{R}$: let $f'(g) := \int_H f(gh)\,\beta(dh)$ so that for any $g' \in gH$

$$f'(g') = \int_H f(gg^{-1}g'h)\,\beta(dh) = \int_H f'(gh)\,\beta(dh) = f'(g),$$

since β is invariant under left-translation by $g^{-1}g' \in H$. Thus f' is constant on each coset gH and there is a unique $\widetilde{f} : G/H \to \mathbb{R}$ such that $f' = \widetilde{f} \circ \pi$ (defined by $gH \mapsto f'(g)$).

DEFINITION 20: ν is a *quotient measure* μ/β on G/H if $\mu(f) = \nu(\widetilde{f})$ for $f : G \to \mathbb{R}$.

The existence and uniqueness of quotient measures in general is the subject of Theorem 7.3.3 of Wijsman (1990). We need only consider the case of compact H:

3.4.3 PROPOSITION (Wijsman, 1990, Proposition 7.3.5): *Let the compact group H act continuously on the right of a l.c. space X and let β be normalized Haar measure on H, i.e., $\beta(H) = 1$. If the measure μ on X is invariant, then the quotient measure μ/β coincides with the induced measure $\pi(\mu)$, where π is the orbit projection $X \to X/H$.*

We will apply this proposition taking X to be G and H to be the stabilizer subgroups G_w ($w \in W$), which indeed act upon G to the right by $(h, g) \mapsto gh$ ($g \in G$, $h \in G_w$). Moreover, G_w is compact as a consequence of Assumption 3.3.2, so $\Delta_r^G|_{G_w} = 1$ (by Remark 2) and μ is thus invariant under right-translation by G_w. For the same reason β_w (which we defined to be the normalized left Haar measure on G_w) is also a right Haar measure. Finally, the orbit projection is $\pi_w : G \to G/G_w$, the canonical quotient map $g \mapsto gG_w$.

Now consider a measure Γ on $W \times G$ having density γ with respect to $\lambda \otimes \mu$. We will calculate the densities of various measures induced from Γ.

3.4.4 LEMMA: *The density of the push-forward measure $\theta(\Gamma)$ with respect to $\theta(\lambda \otimes \mu)$ is*

$$\widetilde{\gamma}(\theta(w, g)) := \int_{G_w} \gamma(w, gh)\,\beta_w(dh).$$

Proof. For any $f : W \times W \to \mathbb{R}$,

$$\begin{aligned}\int f\, d\theta(\Gamma) &= \int f \circ \theta \, d\Gamma \\ &= \int_W \lambda(dw) \int_G \mu(dg)\, \gamma(w,g)\, f(\theta(w,g))\end{aligned}$$

Let $\nu_w := \mu/\beta_w$ be the quotient measure, so that by definition:

$$= \int_W \lambda(dw) \int_{G/G_w} \nu_w(dg) \int_{G_w} \beta_w(dh)\, \gamma(w,gh)\, f(\theta(w,gh))$$

We can take $f \circ \theta$ out of the innermost integral because $\theta(w,gh) = \theta(w,g)$ for any $h \in G_w$:

$$= \int_W \lambda(dw) \int_{G/G_w} \nu_w(dg)\, f(\theta(w,g)) \int_{G_w} \beta_w(dh)\, \gamma(w,gh)$$

We define $\gamma'(w,g) := \int_{G_w} \beta_w(dh)\, \gamma(w,gh)$, so that $\gamma'(w,\cdot)$ is constant on each coset gG_w and there is some $\widetilde{\gamma} : E \to \mathbb{R}$ such that $\gamma' = \widetilde{\gamma} \circ \theta$:

$$= \int_W \lambda(dw) \int_{G/G_w} \nu_w(dg)\, f(\theta(w,g))\, \widetilde{\gamma}(\theta(w,g))$$

The inner integrand is well-defined because it depends on g only through its coset $\pi_w(g) = gG_w$. By Proposition 3.4.3, $\nu_w = \pi_w(\mu)$, so we can replace ν_w by μ:

$$\begin{aligned}&= \int f(\theta(w,g))\, \widetilde{\gamma}(\theta(w,g))\, \lambda(dw)\, \mu(dg) \\ &= \int f\widetilde{\gamma}\, d\theta(\lambda \otimes \mu)\,. \qquad \square\end{aligned}$$

Reversing a Markov chain corresponds to the operation of *transposition* on $W \times W$, defined by the map $T : (w,w') \mapsto (w',w)$ and occasionally written as $(w,w')^T$. We note that T is continuous and is its own inverse. Furthermore, T maps the set E to itself: for any $(w,gw) \in E$ we have $T(w,gw) = (gw,w) = (gw, g^{-1}gw) \in E$. We can define an analogous operation on $W \times G$ under the correspondence with E provided by θ: the continuous map $t : (w,g) \mapsto (gw, g^{-1})$; it is also its own inverse: $t(t(w,g)) = t(gw,g^{-1}) = (g^{-1}gw, g) = (w,g)$. The following commutative diagram illustrates the relationship between t, T, and θ:

$$\begin{array}{ccccc} W \times G & \xrightarrow{\theta} & E & \hookrightarrow & W \times W \\ \updownarrow t & & \updownarrow T|_E & & \updownarrow T \\ W \times G & \xrightarrow{\theta} & E & \hookrightarrow & W \times W \end{array}$$

Indeed, if $\theta(w, g) = \theta(w, g')$ (i.e., $gw = g'w$) then $t(w, g) = (gw, g^{-1})$ and $t(w, g') = (g'w, g'^{-1})$, where $g'^{-1}g'w = w = g^{-1}gw$ and thus $\theta(t(w, g')) = \theta(t(w, g))$. Conversely, if $\theta(t(w, g)) = \theta(t(w', g'))$ then by the previous result $\theta(t(t(w, g))) = \theta(t(t(w', g')))$, and since t is its own inverse, we have shown that $\theta(t(w, g)) = \theta(t(w', g')) \iff \theta(w, g) = \theta(w', g')$. In other words, $\theta \circ t : W \times G \to E$ is constant on the equivalence classes of θ, so there is some continuous $\tau : E \to E$ such that $\theta \circ t = \tau \circ \theta$; we can verify that τ is simply T restricted to E.

3.4.5 LEMMA: *The density of the push-forward measure $t(\Gamma)$ with respect to $\lambda \otimes \mu$ is*

$$\gamma_t(w, g) \coloneqq \varphi(g)\, \gamma(t(w, g)), \qquad \text{where } \varphi(g) = \chi(g) / \Delta_r^G(g).$$

Proof. For any $f : W \times G \to \mathbb{R}$,

$$\int f\, dt(\Gamma) = \int f \circ t\, d\Gamma$$
$$= \int f(gw, g^{-1})\, \gamma(w, g)\, \lambda(dw)\, \mu(dg)$$

changing g^{-1} to g using (2.8.1)

$$= \int f(g^{-1}w, g)\, \Delta_r^G(g^{-1})\, \gamma(w, g^{-1})\, \lambda(dw)\, \mu(dg)$$

changing $g^{-1}w$ to w using (2.8.2)

$$= \int f(w, g)\, \chi(g)\, \Delta_r^G(g^{-1})\, \gamma(gw, g^{-1})\, \lambda(dw)\, \mu(dg)$$

using the definition of t

$$= \int f(w, g)\, \chi(g)\, \Delta_r^G(g^{-1})\, \gamma(t(w, g))\, \lambda(dw)\, \mu(dg). \qquad \square$$

3.4.6 THEOREM: *Let W, G, λ, μ, $(\beta_w)_{w \in W}$ be as stated in this chapter. Then, for any measure Γ on $W \times G$ having density γ w.r.t. $\lambda \otimes \mu$,*

$$\frac{d\theta(\Gamma)}{dT(\theta(\Gamma))}(w, gw) = \frac{\Delta_r^G(g)\, \widetilde{\gamma}(w, gw)}{\chi(g)\, \widetilde{\gamma}(gw, w)}, \qquad w \in W, g \in G,$$

where $\theta(w,g) = (w,gw)$ *and* $T(w,w') = (w',w)$ *for any* $w,w' \in W$, $g \in G$ *and*

$$\widetilde{\gamma}(w,gw) = \int_{G_w} \gamma(w,gh)\,\beta_w(dh), \qquad w \in W, g \in G.$$

Proof. We apply Lemma 3.4.4 to the density for $t(\Gamma)$ from Lemma 3.4.5 to get a density for $\theta(t(\Gamma))$ with respect to $\theta(\lambda \otimes \mu)$:

$$\theta(t(\Gamma)) = \theta(\gamma_t \cdot (\lambda \otimes \mu)) = \widetilde{\gamma}_t \cdot \theta(\lambda \otimes \mu)\,,$$

where

$$\begin{aligned}
\widetilde{\gamma}_t(\theta(w,g)) &:= \int_{G_w} \gamma_t(w,gh)\,\beta_w(dh) \\
&\overset{(a)}{=} \int_{G_w} \varphi(gh)\gamma(t(w,gh))\,\beta_w(dh) \\
&\overset{(b)}{=} \varphi(g)\int_{G_w} \gamma(ghw,h^{-1}g^{-1})\,\beta_w(dh) \\
&\overset{(c)}{=} \varphi(g)\int_{G_w} \gamma(gw,g^{-1}gh^{-1}g^{-1})\,\beta_w(dh) \\
&\overset{(d)}{=} \varphi(g)\int_{G_{gw}} \gamma(gw,g^{-1}h^{-1})\,\beta_{gw}(dh) \\
&\overset{(e)}{=} \varphi(g)\int_{G_{gw}} \gamma(gw,g^{-1}h)\,\beta_{gw}(dh) \\
&\overset{(f)}{=} \varphi(g)\widetilde{\gamma}(\theta(gw,g^{-1})) = \varphi(g)\widetilde{\gamma}(T(\theta(w,g)))\,.
\end{aligned}$$

Here, the various equalities hold for the following reasons: (a) Definition of γ_t; (b) Since φ is a continuous group homomorphism, $\phi(gh) = \phi(g)\phi(h)$, and since G_w is compact, $\varphi|_{G_w} = 1$ identically by Remark 2 on page 20; (c) $hw = w$ since $h \in G_w$; (d) By Remark 1 on page 11, $c_g(h) = h^g := g^{-1}hg$ is a $G_{gw} \to G_w$ group isomorphism: $c_g(\beta_{gw})$ is thus the unique normalized Haar measure on G_w, so it must be equal to β_w; (e) Since G_{gw} is compact, β_{gw} remains unchanged under the change of variables $h \mapsto h^{-1}$; (f) Definition of $\widetilde{\gamma}$.

Thus $\varphi(g)\,\widetilde{\gamma}(T(\theta(w,g)))$ is a density for $\theta(t(\Gamma))$ (and hence for $T(\theta(\Gamma))$, since $T \circ \theta = \theta \circ t$) with respect to $\theta(\lambda \otimes \mu)$. Since, by Lemma 3.4.4, the density for $\theta(\Gamma)$ with respect to the same measure is $\widetilde{\gamma}$, we see that the Radon-Nikodym derivative $d\theta(\Gamma)/dT(\theta(\Gamma))$ is $\widetilde{\gamma}(w,gw)/\varphi(g)\widetilde{\gamma}(gw,w)$ at $(w,gw) \in E$. □

Proof of Theorem 3.3.3. Procedure 1 describes an MH kernel based on the proposal $Q'(dw' \mid w)$ that, given a state w, samples $g \sim Q_G(\,\cdot \mid w)$ and proposes gw. In other words, $Q'(\,\cdot \mid w)$ is the push-forward of $Q_G(\,\cdot \mid w)$ under the map $g \mapsto gw$, making $P(dw)\, Q'(dw' \mid w)$ the push-forward of $P(dw)\, Q_G(dg \mid w)$ under the map $\theta(w, g) = (w, gw)$. We can now apply Theorem 3.4.6 by taking $\Gamma(dw, dg) := P(dw)\, Q_G(dg \mid w)$ with density $\gamma(w, g) = p(w)\, q(g \mid w)$, so that $P(dw)\, Q'(dw' \mid w) = \theta(\Gamma)$ and

$$\begin{aligned} r(w, gw) &:= \frac{d\theta(P(dw)\, Q_G(dg \mid w))}{dT(\theta(P(dw)\, Q_G(dg \mid w)))}(w, gw) \\ &= \frac{\Delta_r^G(g)\, \widetilde{\gamma}(w, gw)}{\chi(g)\, \widetilde{\gamma}(gw, w)} \qquad\qquad w \in W, g \in G \end{aligned}$$

where

$$\begin{aligned} \widetilde{\gamma}(w, gw) &= \int_{G_w} p(w)\, q(gh \mid w)\, \beta_w(dh) \\ &= p(w) \int_{G_w} q(gh \mid w)\, \beta_w(dh) \\ &= p(w)\, q'(g \mid w). \end{aligned}$$

Define

$$R := \big\{ (w, gw) \in E \;\big|\; p(w)\, q'(g \mid w) > 0 \text{ and } p(gw)\, q'(g^{-1} \mid gw) > 0 \big\}.$$

The image of θ is E, so both $\theta(\Gamma)$ and $T(\theta(\Gamma))$ are zero outside E. Thus they are mutually singular outside $R \subset E$ and mutually absolutely continuous on R. We can define $r(w, w') = 1$ outside R, and by inspection we can verify that $r(w', w) = 1/r(w, w')$. Thus we have satisfied all the conditions for Proposition 3.4.1 and by Proposition 3.4.2 the MH kernel with acceptance probability $\alpha(w, w') := \min\{1, r(w', w)\}$ on R satisfies detailed balance. Since we assume that the initial state is within the support of P, and the acceptance probability is always zero for proposals outside the support, α will never be evaluated outside the set R. □

Proof of Theorem 3.3.5. Procedure 2 describes an MH kernel based on a proposal Q' which is a mixture of the types of proposals seen in Procedure 1:

$$Q'(dw' \mid w) = \sum_{i=1}^{n} a(i \mid w) Q'_i(dw' \mid w)$$
$$P(dw)\, Q'(dw' \mid w) = \sum_{i=1}^{n} a(i \mid w)\, P(dw)\, Q'_i(dw' \mid w).$$

Now define $\Gamma_i(dw, dg) = a(i \mid w)\, P(dw)\, Q_i(dg \mid w)$. By a similar argument to the previous proof it follows that $P(dw)\, Q'(dw' \mid w) = \sum_{i=1}^{n} \theta(\Gamma_i)$. As before, we can define a function r_i that is the Radon-Nikodym derivative $d\theta(\Gamma_i)/dT(\theta(\Gamma_i))$ restricted to a set R_i where both those measures are mutually absolutely continuous, and mutually singular outside it. Since $\theta(\Gamma_i)$ is zero outside the set $E_i := \theta(W, G_i)$, we see that $R_i \subset E_i$. Now the problem arises that the E_i may not be disjoint; however, we will show that we can take the R_i to be disjoint without loss of generality.

For each $1 \leq i \leq n$, define V_i to contain all the $1 \leq j \leq n$ such that Assumption 3.3.4 is satisfied for i and j with $k = i$. Now for any $j \in V_i$ the pre-image of $E_i \cap E_j$ under θ is $\{(w, g) \mid w \in W, g \in G_{i,j} G_{i,w}\}$. Applying the assumption, this set has zero measure under Γ_i so $E_i \cap E_j$ has zero measure under $\theta(\Gamma_i)$. Then $\bigcup_{j \in V_i} E_i \cap E_j$ has zero measure under $\theta(\Gamma_i)$ and is symmetric, so it has zero measure under $T(\theta(\Gamma_i))$ as well. Thus, without loss of generality, we can take R_i to be a subset of $E_i \setminus \bigcup_{j \in V_i} E_j$ since it is only unique up to $\theta(\Gamma_i)$-null sets. By the assumption, for any $i \neq j$ either $i \in V_j$ or $j \in V_i$, so the R_i are disjoint. We have found a collection of disjoint sets R_i such that each $\theta(\Gamma_i)$ is mutually absolutely continuous on R_i and mutually singular outside R_i, with $d\theta(\Gamma_i)/d(T(\theta(\Gamma_i))) = r_i$ restricted to R_i. We can now define r so that it takes on the value r_i on E_i, with $R := \bigcup_{i=1}^{n} R_i$. This r is the Radon-Nikodym derivative for Proposition 3.4.1.

It only remains to note that by Assumption 3.3.4 for any w in the support of P and $w' = gw$ sampled according to $Q_i(\,\cdot \mid w)$, $(w, gw) \in R_i$ with probability 1. Thus if the algorithm samples from some Q_i then the probability that r is evaluated outside E_i is zero. $\square$

4 EXPLOITING SYMMETRIES

Judiciously choosing the groups G_i and proposal kernels Q_i allows the MH kernel with group transformations (Procedure 2) to take advantage of symmetries of the target distribution. Consider a distribution P with a density that can be factored as follows:

$$p(w) = \prod_{i=1}^{m} p_i(w), \qquad \text{where } p_i(hw) = p_i(w) \text{ for all } h \in H_i;$$

we say that each group H_i is a symmetry of the factor p_i, or that p_i is *invariant* under the action of H_i. For concreteness, we present a variation on the example of Chapter 1: p is a density with respect to the Lebesgue measure λ on $W = \mathbb{R}^2 \setminus \{(0,0)\}$ with $m = 3$ factors, p_1 and p_2 are as described earlier, and we add another factor p_3 with no useful symmetries; thus H_1 and H_2 are, respectively, the groups that rotate and scale $\mathbb{R}^2$ around its origin, and H_3 is the trivial group (containing only the identity transformation).

To apply Procedure 2 to this example, take $n = 2$, $G_1 = H_2$, $G_2 = H_1$, and $a(i \mid w) = 1/2$ for $i = 1,2$ and all w. In this example, for $i = 1,2$, $\Delta_r^{G_i} = 1$ identically (since both groups are commutative) and $q_i' = q_i$ (by Remark 3 on page 24, since the isotropy subgroups are trivial). The proposed state is $w' = gw$ for some $g \in G_i$, so we see immediately that $p_j(w') \neq p_j(w)$ is only possible for $j \in \{i,3\}$. Thus, in the $i = 1$ case, the p_2 factor cancels out of the acceptance probability:

$$\alpha|_{i=1} = \frac{\chi_1(g)\, p_1(gw)\, \cancel{p_2(gw)}\, p_3(gw)\, q_1(g^{-1} \mid gw)}{p_1(w)\, \cancel{p_2(w)}\, p_3(w)\, q_1(g \mid w)}. \tag{4.0.1}$$

Next we choose q_1, attempting to cancel the χ_1 and p_1 factors as well. Since G_1 acts by scaling $\mathbb{R}^2$, we can identify it with $\mathbb{R}_+^\times$: the group of positive real numbers under multiplication (i.e., composition of scaling factors). Then $g \in \mathbb{R}_+^\times$ acts on $\mathbb{R}^2$ by $(x,y) \mapsto (gx, gy)$, the corresponding effect on the Lebesgue measure (area) on the plane is described by $\chi_1(g) = g^2$, and $\mu_1(dg) = g^{-1}dg$ is a Haar measure on $\mathbb{R}_+^\times$. The obvious choice is to set $q_1(g \,|\, w) \propto \chi_1(g)\, p_1(gw)$ with a normalizing constant $c_1(w)$; then for any $w \in W$, since q_1 must be a probability kernel, we use the definitions of μ_1 and χ_1 to get

$$\int_0^\infty q_1(g \,|\, w)\, g^{-1} dg \;=\; c_1(w) \int_0^\infty p_1(gw)\, g\, dg \;=\; 1. \tag{4.0.2}$$

A simple calculation using (4.0.2) yields $c_1(gw) = g^2 c_1(w) = \chi_1(g)\, c_1(w)$, which we substitute into (4.0.1):

$$\alpha|_{i=1} = \frac{\cancel{\chi_1(g)}\cancel{p_1(gw)}\, p_3(gw)\, \cancel{\chi_1(g)}\cancel{c_1(w)}\cancel{\chi_1(g^{-1})}\cancel{p_1(w)}}{\cancel{p_1(w)}\, p_3(w)\, \cancel{c_1(w)}\cancel{\chi_1(g)}\cancel{p_1(gw)}}.$$

An analogous derivation can be carried out for the $i = 2$ case, identifying G_2 with $[0, 2\pi)$ as the set of rotation angles under the operation of addition (mod 2π). Then $\chi_2 = 1$ and μ_2 is just the Lebesgue measure on G_2; again we get $\alpha|_{i=2} = p_3(gw)/p_3(w)$. In fact, the same technique works in general for any target distribution P, even if $\Delta_r^{G_i} \neq 1$, as long as $\chi_i(g)\, p_i(gw)$ is μ_i-integrable:

4.0.1 **PROPOSITION:** *Suppose $q_i(g \,|\, w) := c_i(w)\, \chi_i(g)\, p_i(gw)$ ($g \in G_i$, $w \in W$) is a probability kernel density for some appropriately chosen normalizer c_i. Then $q_i' = q_i$ and*

$$\frac{\chi_i(g)\, p_i(gw)\, q_i(g^{-1} \,|\, gw)}{\Delta_r^{G_i}(g)\, p_i(w)\, q_i(g \,|\, w)} = 1.$$

Proof. If $q_i(h \,|\, gw) := c_i(gw)\, \chi_i(h)\, p_i(hgw)$ is a probability kernel density for any $g \in G_i$ and $w \in W$, then

$$\begin{aligned} 1 &= c_i(gw) \int_{G_i} \chi_i(h)\, p_i(hgw)\, \mu_i(dh) \\ &= c_i(gw) \int_{G_i} \chi_i(hgg^{-1})\, p_i(hgw)\, \mu_i(dh) \end{aligned}$$

μ_i, being a left Haar measure, is $\Delta_r^{G_i}$-relatively invariant under *right* multiplication: since $f(hg)$ is the *right-translation* of $f(h)$ by g^{-1} (see (2.8.2) and ensuing discussion), it satisfies $\int_{G_i} f(hg)\, \mu_i(dh) = \Delta_r^{G_i}(g^{-1}) \int_{G_i} f(h)\, \mu_i(dh)$; we therefore replace hg with h in the integrand

$$= c_i(gw)\, \Delta_r^{G_i}(g^{-1}) \int_{G_i} \chi_i(hg^{-1})\, p_i(hw)\, \mu_i(dh)$$

and since $\chi_i : G_i \to \mathbb{R}_+^\times$ is a group homomorphism, $\chi_i(hg^{-1}) = \chi_i(h)\, \chi_i(g^{-1})$

$$\begin{aligned}
&= c_i(gw)\, \Delta_r^{G_i}(g^{-1})\, \chi_i(g^{-1}) \int_{G_i} \chi_i(h)\, p_i(hw)\, \mu_i(dh) \\
&= \frac{c_i(gw)\, \chi_i(g^{-1})}{c_i(w)\, \Delta_r^{G_i}(g)}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{\chi_i(g)\, p_i(gw)\, q_i(g^{-1} \mid gw)}{\Delta_r^{G_i}(g)\, p_i(w)\, q_i(g \mid w)} &= \frac{\chi_i(g)\, p_i(gw)\, \cancel{c_i(gw)}\, \cancel{\chi_i(g^{-1})}\, p_i(g^{-1}gw)}{\cancel{\Delta_r^{G_i}(g)}\, p_i(w)\, \cancel{c_i(w)}\, \chi_i(g)\, p_i(gw)} \\
&= \frac{\cancel{\chi_i(g)}\, \cancel{p_i(gw)}\, \cancel{p_i(w)}}{\cancel{p_i(w)}\, \cancel{\chi_i(g)}\, \cancel{p_i(gw)}}.
\end{aligned}$$

To show that $q_i' = q_i$, suppose $w \in W$ and $g' \in gG_{i,w}$, so that $g' = gh$ for some $h \in G_{i,w}$ and $q_i(g' \mid w) = c_i(w)\, \chi_i(gh)\, p_i(ghw)$. Now $\chi_i(gh) = \chi_i(g)\, \chi_i(h)$ and $\chi_i(h) = 1$ for all h in the compact subgroup $G_{i,w}$ (see Remark 2 on page 20), so $\chi_i(g') = \chi_i(g)$. Similarly, $ghw = gw$ by the definition of the stabilizer subgroup, so $p_i(g'w) = p_i(gw)$. Thus $q_i(\,\cdot \mid w)$ is constant on the cosets of $G_{i,w}$, and the result follows by Remark 3 on page 24. □

We conclude that when Procedure 2 is applied to a target distribution having factors p_i invariant under H_i, the proposals in the mixture should be chosen so that (a) $G_i \subset H_j$ for as many $j \neq i$ as possible, eliminating the p_j terms from the acceptance probability, and (b) $q_i(g \mid w) \propto \chi_i(g)\, p_i(gw)$ to eliminate the χ_i, $\Delta_r^{G_i}$, and p_i terms; the constraint is that the G_i transformations sampled according to Q_i must collectively be rich enough to be able to explore the support of P. Indeed, ideally only the non-symmetric factors of p appear in the acceptance probability, as we saw in the example. If we had $p_3 = 1$ as in Chapter 1, we would recover the no-reject algorithm that produces independent samples every time it performs a rotation and a scaling. The simpler acceptance probability also means that only the non-symmetric factors contribute to the time required to compute it.

Note. The preceding discussion remains valid when the target distribution conforms only *approximately* to the structure described in this chapter. For example, the factorization may be approximate, so that $p(w) \approx \prod p_i(w)$; the symmetries H_i may also be approximate, so that $p_i(w) \approx p_i(hw)$. The MH algorithm gracefully handles such imperfectly symmetric target distributions: the calculation of the acceptance probability accounts for them. Although the computational advantages are then lost, we expect exploiting even approximate symmetries to improve convergence rates.

5

THE SLAM PROBLEM

The SLAM problem is concerned with a robot navigating an unknown environment under the effect of sensor and control noise. The goal is to determine the robot's trajectory as well as the map of the environment based on the robot's observations. The environment comprises N landmarks; the position of each is denoted by a variable Y_i ($i = 1, \ldots, N$) taking values in a space $\mathcal{Y}$. Let X_t ($t = 0, \ldots, T$) denote the *pose* (typically, position and orientation) of the robot at time step t and take values in space $\mathcal{X}$. At every time step the robot can observe the landmarks, and at time step t the observation of landmark i is denoted by Z_t^i taking values in $\mathcal{Z}$. For simplicity, we assume that all landmarks can always be observed and the robot can distinguish the landmarks. The goal of the SLAM problem is to estimate the trajectory $X = (X_0, \ldots, X_T)$ and the landmark positions $Y = (Y_1, \ldots, Y_N)$ based on the observations $Z = (Z_t^i)_{0 \leq t \leq T, 1 \leq i \leq N}$ (our notation consistently refers to time steps and landmarks with subscripted and superscripted indices, respectively).

We use the Bayesian formulation of SLAM, in which the robot's trajectory, environment, and observations are random variables and are assumed to evolve according to the following dynamical system: (a) X_0 and Y are independent with known densities; (b) at each time step $t = 0, 1, 2, \ldots$, each observation Z_t^i depends only on X_t and Y_i via the conditional density $p_{Z_t^i | X_t, Y_i}$, and (c) the pose of the robot X_t depends only on X_{t-1} and the previous observations $Z_{<t} := (Z_0, \ldots, Z_{t-1})$ via the conditional density $p_{X_t | X_{t-1}, Z_{<t}}$ (where $Z_t = (Z_t^1, \ldots, Z_t^N)$). That is, we make the following Markov assumptions: (a) Z_t^i is conditionally independent of $X_{<t}$ and Y_j ($j \neq i$) given X_t and Y_i, and (b) X_t is conditionally independent of $X_{<t-1}$ and Y given X_{t-1} and $Z_{<t}$. Also, we assume throughout that conditional densities exist relative to some dominating measure, usually an appropriate Lebesgue or Haar measure.

The SLAM posterior is the conditional density $p_{X,Y|Z}(\,\cdot\,|\,z)$ over trajectories and environments given observations $Z = z$. We first factor the joint density $p_{X,Y,Z}$ as $p_Y(y)\,p_{X,Z|Y}(x,z\,|\,y)$. Then, under the above Markov assumptions, we obtain

$$\begin{aligned} p_{X,Z|Y}(x,z\,|\,y) &= \prod_{t=0}^{T} p_{X_t|X_{t-1},Z_{<t}}(x_t\,|\,x_{t-1},z_{<t}) \cdot p_{Z_t|X_t,Y}(z_t\,|\,x_t,y) \\ p_{X,Y|Z}(x,y\,|\,z) &= \frac{p_Y(y)p_{X,Z|Y}(x,z\,|\,y)}{p_Z(z)}. \end{aligned} \tag{5.0.1}$$

We consider the SLAM problem in which the robot moves on a two-dimensional plane. Then its position and orientation are fully specified by the rigid (i.e., distance-preserving and non-reflecting) transformation of $\mathbb{R}^2$ from the robot's body-local coordinate system to the global coordinates. Any rigid transformation can be decomposed into a rotation around the origin followed by a translation; the set of such transformations under composition forms the *special Euclidean group* $\mathrm{SE}(2)$. The space of poses is therefore $\mathcal{X} := \mathrm{SE}(2)$. The landmarks are specified by their positions on the plane, so $\mathcal{Y} := \mathbb{R}^2$.

5.1 SYMMETRIES OF SLAM

We assume that, apart from the landmarks, the environment is essentially homogeneous (we will elaborate upon what this means), giving rise to certain symmetries in the factors of the SLAM posterior distribution. If a robot has pose $x \in \mathcal{X}$, in its body-local frame the coordinates of another pose $x' \in \mathcal{X}$ are $x^{-1}x'$ and those of a landmark $y \in \mathcal{Y}$ are $x^{-1}y$. One can verify that these local coordinates do not change if x, x', and y are all transformed by some $g \in \mathcal{G} := \mathrm{SE}(2)$ to gx, gx', and gy, respectively. The assumption that the environment is homogeneous means, firstly, that the motion of the robot is not affected by its location in a way undetectable to its sensors. In particular, for a given value of $Z_{<t}$, the motion model $p_{X_t|X_{t-1},Z_{<t}}$ depends only on the relative movement $X_{t-1}^{-1}X_t$ and not on the global coordinates. Secondly, since the sensors are fixed to the robot's body, the observation of a landmark depends only on its local coordinates in the robot's frame: $p_{Z_t^i|X_t,Y_i}$ depends only on Z_t^i and $X_t^{-1}Y_i$. Thirdly, the landmarks and the robot's initial pose are *a priori* equally likely to be anywhere in the environment: p_{Y_i} and p_{X_0} are invari-

ant under $\mathcal{G}$. The homogeneity of the environment thus implies that no reference frame is privileged, and that being transformed by $\mathcal{G}$ does not affect the likelihood of a SLAM solution. To resolve the resulting ambiguity, without loss of generality we work in the coordinate system whose origin is the robot's initial pose (i.e., X_0 is the identity transformation). Note that p_{Y_i} must then be a distribution that is invariant under rigid transformations of the plane: the Lebesgue measure; however, it is not a probability distribution. We may nevertheless use it for inference as a so-called *improper* prior as long as the resulting posterior is indeed a probability distribution.

Thus, for our purposes, the SLAM posterior is a distribution over the state space $W := \mathcal{X}^T \times \mathcal{Y}^N$ of all possible trajectories (that start at the origin) and environments. The group $K := \mathcal{G}^T \times \mathcal{G}^N$ acts on W, with the $g_t, g^i \in \mathcal{G}$ components acting on $w_t \in \mathcal{X}$ and $w^i \in \mathcal{Y}$, respectively (by our convention, the subscripts and superscripts refer to the pose and landmark components, respectively). Using the terminology of Chapter 4, the $p_{X_t|X_{t-1},Z_{<t}}$ factors are invariant under the subgroups $H_t := \{g \in K \mid g_{t-1} = g_t\}$ and the $p_{Z_t^i|X_t,Y_i}$ factors under $H_t^i := \{g \in K \mid g_t = g^i\}$.

5.2 THE MCMC-SLAM ALGORITHM

We now specify how Procedure 2 may be applied to the problem of sampling from the SLAM posterior. First, we select a function $b : \{1, \ldots, N\} \to \{1, \ldots, T\}$, which "anchors" each landmark to one of the time steps at which it was observed. The proposal is a mixture of $T + N$ kernels, indexed with subscripts or superscripts as before. The mixture component corresponding to time step t transforms W by an element of $G_t := \left(\bigcap_{s \neq t} H_s\right) \cap \left(\bigcap_i H^i_{b(i)}\right)$, which is a symmetry of the $p_{X_s|X_{s-1},Z_{<s}}$ factors for $s \neq t$ and of the $p_{Z_s^i|X_s,Y_i}$ factors for $(s,i) \notin V_t$, where

$$V_t := \{(s,i) \mid s < t \leq b(i) \text{ or } b(i) < t \leq s\}.$$

Indeed, this is a maximal set of factors for which G_t can be a symmetry without being reduced to triviality. One can verify that an element of G_t is determined by $g \in \mathcal{G}$ that acts on w_s if $s \geq t$ and on w^i if $b(i) \geq t$; other components of $w \in W$ are left unchanged. The mixture components corresponding to landmark i use $G^i :=$

$\bigcap_t \big(H_t \cap \big(\bigcap_{j \neq i} H_t^j\big)\big)$, which is a symmetry of all the $p_{X_t|X_{t-1},Z_{<t}}$ factors and those $p_{Z_t^j|X_t,Y_j}$ factors with $j \neq i$; again, this is a maximal invariant set. The corresponding proposal kernel densities q_t and q^i are chosen to be proportional to $p_{X_t|X_{t-1},Z_{<t}}$ and $p_{Z_{b(i)}^i|X_{b(i)},Y_i}$, respectively, following Chapter 4. Procedure 3 shows the resulting algorithm[1]. Note that if the trajectory is stored in the tree structure of Fenwick (1994), modified to support non-commutative operations, the state update can be carried out in $O(\log T)$ time; the calculation of the acceptance probability then dominates, thus scaling with the number of factors whose values have changed.

PROCEDURE 3: Given $w \in W$ consisting of a trajectory $x_1, \ldots, x_T$ and landmarks $y_1, \ldots, y_N$, propose w':

(i) Sample either a time step t or a landmark i from a given discrete distribution with probabilities $a_t(w)$ and $a^i(w)$, respectively (i.e., $\sum_{t=1}^T a_t(w) + \sum_{i=1}^N a^i(w) = 1$).

(ii) If the previous step sampled time step t:

1. Set $x_t' \sim p_{X_t|X_{t-1},Z_{<t}}(\,\cdot \mid x_{t-1}, z_{<t})$.
2. Set $x_s' := x_t' {x_t}^{-1} x_s$ for $s > t$.
3. Set $y_i' := x_t' {x_t}^{-1} y_i$ for $b(i) \geq t$.
4. Calculate $\alpha := \dfrac{a_t(w')}{a_t(w)} \displaystyle\prod_{(s,i) \in V_t} \dfrac{p_{Z_s^i|X_s,Y_i}(z_s^i \mid x_s', y_i')}{p_{Z_s^i|X_s,Y_i}(z_s^i \mid x_s, y_i)}$.

(iii) Otherwise, if it sampled landmark i:

1. Set $y_i' \sim p_{Z_{b(i)}^i|X_{b(i)},Y_i}(z_{b(i)}^i \mid x_{b(i)}, \,\cdot\,)$.
2. Calculate $\alpha := \dfrac{a^i(w')}{a^i(w)} \displaystyle\prod_{t \neq b(i)} \dfrac{p_{Z_t^i|X_t,Y_i}(z_t^i \mid x_t, y_i')}{p_{Z_t^i|X_t,Y_i}(z_t^i \mid x_t, y_i)}$.

(iv) Accept new state w' with probability $\min\{1, \alpha\}$. All unmodified variables keep their original values.

1 We use the notation $x \sim p(\,\cdot\,)$ with the assumption that p is integrable and implying an appropriate normalizing constant.

6 EXPERIMENTS

We applied the MCMC-SLAM algorithm to two publicly available data sets (Djugash, 2010) from an autonomous robot with sensors that measure range to radio beacons. In the Plaza 1 data set, the robot traveled 1.9 km over 9,657 time steps and received 3,529 range observations of four landmarks. In the Plaza 2 data set, the robot traveled 1.3 km over 4,091 time steps and received 1,816 range measurements, also of four landmarks. Highly accurate ground truth trajectories were also recorded. We compare the algorithm to the Spectral SLAM algorithm (Boots and Gordon, 2013). We found that exploiting symmetries as outlined in Chapter 4 was crucial: the naive MCMC kernel that updated individual components of the trajectory or environment did not make any progress in a reasonable amount of time.

Table 1 shows the RMS distance of each robot pose from the ground truth for each data set. It is averaged over 50 independent runs of the MCMC algorithm, with the interval indicating one standard deviation. Since any SLAM solution is only specified up to the choice of origin, we apply the best-fit rigid transformation between the estimated and known maps (Boots and Gordon do the same).

The MCMC $(r + s)$ algorithms incrementally extend the SLAM posterior by introducing the factors coming from each time step, in turn. The chain takes r steps after each extension, and s steps at the end. At each time step, newly introduced variables are initialized by sampling from the corresponding proposal kernel. MCMC (10+1000) took approximately 13.8 s on Plaza 1 and 2.8 s on Plaza 2; MCMC (100+10000) took 131.1 s and 28.1 s, respectively. The larger number of steps is required to achieve good accuracy on Plaza 2 because it is more challenging: the robot consistently turns in one direction, making the control noise biased. In comparison, Spectral SLAM took 0.73 s and 0.51 s on a similar computer. The "Spectral + Opt." algorithm runs a final batch optimization pass and takes several thousands of seconds.

Thus, even though the MCMC algorithm is computationally somewhat more expensive, we see that it performs competitively with Spectral SLAM and all the other

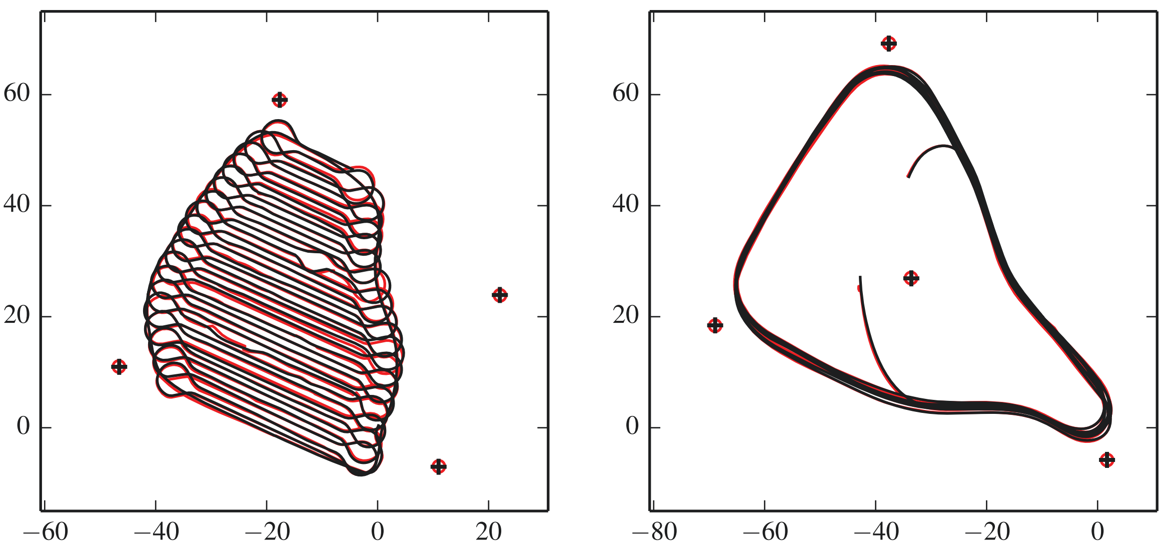

Figure 2: An MCMC sampled trajectory and map (black) overlaid over the ground truth (red) for the Plaza 1 (left) and Plaza 2 (right) data sets.

Table 1: Comparison of Trajectory RMS Errors.

Algorithm	**Plaza 1**	**Plaza 2**
Spectral	0.79 m	0.35 m
Spectral + Opt.	0.69 m	0.30 m
MCMC (10+1000)	0.32 ± 0.02 m	0.54 ± 0.06 m
MCMC (100+10000)	0.33 ± 0.04 m	0.36 ± 0.03 m

methods tested by Boots and Gordon (2013). In addition, it has the advantage of easily handling missing observations, without a process of imputing them as is done by Spectral SLAM. Finally, being a Bayesian algorithm, it produces the SLAM posterior distribution rather than just a solution; we expect it to perform better if the robot noise characteristics are faithfully modeled. Indeed, because the variance of the RMS error *increases* when the chain is allowed to run longer on Plaza 1, we conclude that it is converging to an inaccurate SLAM posterior, which is probably a symptom of poorly chosen priors.

It remains to specify the probabilistic robot models used in the experiments, as well as the landmark "anchoring" function b and the mixture coefficients $a_t(w)$ and $a_i(w)$ employed in the specification of Procedure 3. We adopted an extension of the "velocity motion model" of Thrun et al. (2005). The control input is given as a commanded velocity $\hat{v}$ and steering rate $\hat{\omega}$. The model samples $v \sim \mathcal{N}(\hat{v}, \sigma_v)$ and $\omega \sim \mathcal{N}(\hat{\omega}, \sigma_\omega)$ as being normally distributed. The robot is assumed to travel in a circular arc with length $d := v \cdot \Delta T$ and central angle $\alpha := \omega \cdot \Delta T$ (unless $\omega = 0$, in which case it moves in a straight line). The final heading of the robot changes by $(\omega + \rho) \cdot \Delta T$, where $\rho \sim \mathcal{N}(0, \sigma_\rho)$ is a *slippage angle*. Explicitly, if the robot's pose is given by $x = y = \theta = 0$ (i.e., it is at the origin, facing in the +X direction), its pose at the next step is (if $\omega \neq 0$)

$$\begin{aligned}
x' &= d \sin(\alpha)/\alpha \\
y' &= d(\cos(\alpha) - 1)/\alpha \\
\theta' &= \alpha + \rho \cdot \Delta T.
\end{aligned}$$

If $\omega = 0$ then $x' = d$, $y' = 0$, and $\theta' = \rho \cdot \Delta T$. For any other initial pose, the effect of the same control input can be found by a rigid transformation (which is a symmetry of the model). The standard deviations were chosen as follows: $\sigma_v := 0.1|\hat{v}|$, $\sigma_\omega := (1.0\,\mathrm{deg/m})|\hat{v}| + 0.1|\hat{\omega}|$, and $\sigma_\rho := (0.1\,\mathrm{deg/m})|\hat{v}| + \sqrt{0.001}|\hat{\omega}|$.

The observation model uses the sensor range reading $\hat{z}$, samples $z \sim \mathcal{N}(\hat{z}, \sigma_z)$ and $\theta \sim \mathcal{U}([0, 2\pi))$, and produces the landmark location $x := z\cos(\theta)$ and $y := z\sin(\theta)$ in the robot's local coordinate system; we used $\sigma_z = 1.0\,\mathrm{m}$.

The function b was chosen as $i \mapsto \arg\min_t z_t^i$, so each landmark would be "anchored" to the time step at which the robot observed the closest range to it. The mixture coefficients were chosen to be $a_t(w) := \kappa_3(p_{X_t|X_{t-1},Z_{<t}})$ and $a^i(w) := \kappa_2(p_{Z^i_{b(i)}|X_{b(i)},Y_i})$, where $\kappa_\gamma(p) = \gamma p^{-1/\gamma}$. The γ parameter is intended to represent the degrees of freedom in the robot poses and landmark positions, respectively.

The variables are initialized as they are introduced into the state space, using the maximum-likelihood estimates of X_t from the corresponding motion model or of Y_i from the first observation. With range-only SLAM, the maximum likelihood estimate of a landmark's position is not unique: points equidistant from the robot have equal likelihood: one of these points is chosen randomly following the uniform distribution on the circle.

7 CONCLUSIONS AND FUTURE WORK

The Metropolis-Hastings (MH) algorithm is a widely used technique to implement approximate probabilistic inference, but its "textbook version" is quite limited. To build potentially faster mixing chains, in this thesis we explore the possibility of proposals where the next state is based on transforming the current one using a randomly chosen transformation. The main contribution of the thesis is a formula that shows how the acceptance function can be calculated in closed form in this case. This is shown both for a single kernel, and when a mixture kernel is used. The strength of the approach is its generality: We derive the results without any differentiability requirements, making them applicable to both continuous and discrete domains. While the increased generality made the thesis more technical, to enhance clarity, we used the SLAM problem to illustrate the ideas. On a challenging domain, we obtained strong experimental evidence in favor of our new approach. While it remains for future work to demonstrate the approach on a wider range of problems, we believe that the approach proposed in the thesis, due to its generality and flexibility, will have a profound impact on how AI systems perform approximate inference.

BIBLIOGRAPHY

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Byron Boots and Geoffrey J. Gordon. A spectral learning approach to range-only SLAM. In *ICML*, pages 19–26, 2013.

Nicolas Bourbaki. *Integration II, Chapters 7–9*. Elements of Mathematics. Springer, 2004.

Henri Cartan. Sur la mesure de Haar. *Comptes Rendus de l'Académie des Sciences de Paris*, 211:759–762, 1940.

J.T Chang and D. Pollard. Conditioning as disintegration. *Statistica Nederlandica*, 51(3):287–317, 1997.

Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

Persi Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, California, 1988.

Joseph Djugash. *Geolocation with Range: Robustness, Efficiency and Scalability*. PhD thesis, Carnegie Mellon University, 2010.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1989.

Peter M. Fenwick. A new data structure for cumulative frequency tables. *Software: Practice and Experience*, 24(3):327–336, 1994.

João Filipe Ferreira and Jorge Miranda Dias. *Probabilistic Approaches to Robotic Perception*. Springer, 2014.

Alfréd Haar. Der Massbegriff in der Theorie der kontinuierlichen Gruppen. *Annals of Mathematics*, 34(1):147–169, 1933.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Risi Kondor. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008.

Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence*. Chapman and Hall/CRC, 2003.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, 2004.

Jun S. Liu and Chiara Sabatti. Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, 87(2):353–369, 2000.

Jun S. Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274, 1999.

Christopher Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Leopoldo Nachbin. *The Haar Integral*. The University Series in Higher Mathematics. Van Nostrand, Princeton, NJ, 1965.

Mathias Niepert. Markov chains on orbits of permutation groups. In *UAI*, 2012a.

Mathias Niepert. Lifted probabilistic inference: An MCMC perspective. In *International Workshop on Statistical Relational AI*, 2012b.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

David L. Poole and Alan K. Mackworth. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, 2010.

Simon J.D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.

G. Roberts and J. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2009.

Roshan Shariff, András György, and Csaba Szepesvári. Exploiting symmetries to construct efficient MCMC algorithms with an application to SLAM. In *AISTATS*, volume 38 of *JMLR: W&CP*, 2015. Forthcoming.

Alex J. Smola and Risi Kondor. Kernels and regularization on graphs. In *COLT*, pages 144–158, 2003.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.

Luke Tierney. Markov chains for exploring posterior distributions (with discussions). *The Annals of Statistics*, 22:1701–1762, 1994.

Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998.

Péter Torma, András György, and Csaba Szepesvári. A Markov-chain Monte Carlo approach to simultaneous localization and mapping. In *AISTATS*, pages 852–859, 2010.

André Weil. *L'intégration dans les groupes topologiques et ses applications*, volume 869 of *Actualités Scientifiques et Industrielles*. Hermann Cie., 1940.

Robert A. Wijsman. *Invariant Measures on Groups and Their Use in Statistics*, volume 14 of *Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, California, 1990.

Yang Xiang. *Probabilistic Reasoning in Multiagent Systems: A Graphical Models Approach*. Cambridge University Press, 2002.