

# Formulating Detection and Breaking of Parameter Symmetries

Kenny Chiu

December 8, 2019

# 1 Background

*Probabilistic programming* is a modern programming paradigm in which inference is automatically performed on probabilistic models that are specified by the user. The probabilistic models are often in the form of a generative Bayesian model and can be conveniently expressed by a graphical model [Rai17]. A key consideration of probabilistic programming is that the choice of the inference algorithm used for the given probabilistic model is left up to the underlying inference engine. By abstracting away the inference step, users have the flexibility to work with multiple models without needing to modify their inference procedure for each model.

However, existing inference algorithms do not work well in all models with one challenge being the possible existence of *symmetries* in the parameterization of a model [NMT13]. A symmetry can be informally described as a transformation to the model parameters that does not change the posterior distribution. If a symmetry is present in a model, this poses the problem of parameter nonidentifiability where the results of inference may be of poor quality and difficult to interpret. Another consequence of having parameter symmetries is that different samples of the model parameters may be strongly correlated as a result. In particular, this may reduce the performance of sampling-based inference algorithms.

The problems of parameter symmetries provide motivation for *detecting* and *breaking* them in probabilistic programs. Symmetry detection refers to determining if a symmetry exists in the parameterization of a model, and symmetry breaking refers to dealing with symmetries in some way such that the problems they cause are no longer of concern. The paper by Nishihara et al. [NMT13] introduces algorithms for automatically detecting various types of parameter symmetries in a probabilistic programming context. We directly build on their work by formulating two mathematical perspectives of symmetry detection and breaking that have potential for automation. We first provide an overview of the required concepts and notations in the remainder of this section.

## 1.1 Definitions and notations

We borrow the notation used by Nishihara et al. We represent probabilistic models using factor graphs  $(\theta, F)$  with variables  $\theta = (\theta_1, \dots, \theta_N)$  and factors  $F = (F_1, \dots, F_K)$ . Let  $\Theta$  denote the space of variable values.  $\theta \in \Theta$  contains all parameters, latent variables, and observations (which are fixed).  $F$  represents all functions, operations, constraints, priors and likelihoods that the posterior distribution may factorize into. Figure 1 shows an example factor graph from Nishihara et al.

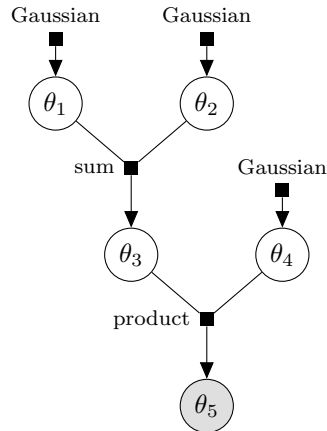


Figure 1: An example factor graph. A variable is represented by a circle and a factor is represented by a square. The shaded circle represents an observed variable.

Given data, the unnormalized posterior distribution can then be expressed as

$$\prod_k F_k(\theta)$$

where factor  $F_k$  may not necessarily depend on all of  $\theta$ . In the context of symmetries, priors are not of interest as it is assumed that the parameters of the prior are chosen by the user. The remainder of this report will refer to non-prior factors when considering  $F_k$ . We now formally define symmetries as provided in Nishihara et al.

**Definition 1.** A symmetry  $\sigma : \Theta \rightarrow \Theta$  is a measurable function with a measurable inverse that satisfies

$$\prod_k F_k(\theta) \propto \prod_k F_k(\sigma(\theta))$$

and where if  $\theta_n$  is some observed variable, then the symmetry keeps  $\theta_n$  fixed. In other words, a symmetry is a transformation on the variables that preserves the product likelihood up to a scaling constant. One may also consider a specific subset of symmetries, referred to as local symmetries.

**Definition 2.** A local symmetry is a symmetry  $\sigma$  that satisfies

$$F_k(\theta) \propto F_k(\sigma(\theta))$$

for all non-prior factors  $F_k$ . In contrast to a symmetry, a local symmetry is a transformation on the variables that preserves the likelihood at each factor up to a scaling constant. Nishihara et al. mainly focuses on detecting local symmetries as they tend to be easier to identify than non-local symmetries.

The work by Nishihara et al. discusses several types of symmetries and how they can be detected for a given factor graph. Their work is presented in the context of a probabilistic programming environment where it is assumed that all built-in factors have been annotated with constraints and/or labels that correspond to specific types of symmetries. In this report, we will assume that these annotations are intrinsic properties of the factors  $F_k$  and hence are known for a given factor graph. The next two subsections present two of the symmetries discussed in Nishihara et al. and review the approaches and annotations used in their algorithm to detect them.

## 1.2 Scaling symmetries

A local symmetry that describes the case where variables can be scaled without affecting the overall likelihood is known as a *scaling symmetry*.

**Definition 3.** A scaling symmetry is a local symmetry  $\sigma$  such that

$$\sigma : (\theta_1, \dots, \theta_N) \mapsto (r_1\theta_1, \dots, r_N\theta_N) = (e^{d_1}\theta_1, \dots, e^{d_N}\theta_N)$$

where  $r_n \in \mathbb{R}_+$  and  $d_n = \log r_n$ . Note that the definition of scaling symmetries only includes positive scaling. A symmetry that describes the case where the signs of variables can be flipped is a *sign-flip symmetry* which we only briefly allude to in this report.

Examples of the constraints that factors impose on the scaling of the variables are provided in Table 1. For example, a *sum factor* that sums inputs  $a + b = c$  preserves its integrity under scaling only if both  $a$ ,  $b$ , and  $c$  are scaled by the same amount. This is enforced by the constraint  $d_a = d_b = d_c$  that it imposes.

Let  $d = (d_1, \dots, d_N)$  be the vector of scaling exponents. Consider the matrix  $\mathcal{C}$  constructed by stacking all the constraints in a given factor graph, which we refer to as the *constraint matrix*. For example, the sum factor above adds the rows  $d_a - d_c = 0$  and  $d_b - d_c = 0$  to  $\mathcal{C}$ . Its null space  $\mathcal{N}(\mathcal{C})$  then describes the space of  $d$  that satisfies all the constraints. The scaling symmetries of the factor graph can then be expressed by the set  $\{\sigma_d : d \in \mathcal{N}(\mathcal{C})\}$ .

factor	constraints
$c = a + b$	$d_a = d_b = d_c$
$c = a \times b$	$d_c = d_a + d_b$
$x \geq 0$	none

Table 1: Example factors and the constraints they impose on potential scaling symmetries.

### 1.3 Permutation symmetries

The *permutation symmetry* describes the case where the variables can be permuted without changing the overall likelihood. Permutation symmetries are often present in models that have mixture components or latent features and are commonly characterized by the label-switching problem.

**Definition 4.** A permutation symmetry is a non-local symmetry  $\sigma$  that permutes the components of  $\theta$  and that satisfies

$$\prod_{\text{label}(k)=c} F_k(\sigma(\theta)) = \prod_{\text{label}(k)=c} F_k(\theta)$$

for all *factor labels*  $c$ . A factor label is an identifier of a factor that is shared only between factors of the same type. For example, all binary sum factors would have the same label but a binary sum factor and a ternary sum factor would not.

To detect permutation symmetries, the factors must have factor labels and their arguments (edges) must also be labeled such that two arguments share the same label if and only if the factor is symmetric with respect to those two arguments. For example, if the variables to the arguments  $a$  and  $b$  of the sum factor  $c = a + b$  are symmetric, the incoming edges of the sum factor would have the same label. Given the labels, Nishihara et al. then reduces the detection problem to a graph *automorphism* problem where the permutation symmetries of the factor graph are the automorphisms that preserve the factor and argument labels. In the context of graphs, an automorphism is a mapping of a graph onto itself that retains how the vertices and edges are connected. Detection of graph automorphism can be done in quasipolynomial time [Bab16] though in practice most graphs can be tested in linear time [BES80].

Nishihara et al. also note that this approach is dependent on how the factor graph is structured. For example, the approach will identify the symmetry across  $a$ ,  $b$  and  $c$  in the ternary sum factor  $a + b + c$  but will only identify the symmetry for  $a$ ,  $b$  and for  $a + b$ ,  $c$  in the layered binary sum factors  $(a + b) + c$ .

### 1.4 Other symmetries

We briefly introduce the other (local) symmetries discussed in Nishihara et al. without giving a formal definition. As referred to previously, a sign-flip symmetry describes the case where the signs of variables can be flipped. The symmetry is similar to the scaling symmetry in that each variable is multiplied by some  $r_n$  but with the restriction that  $r_n = (-1)^{s_n}$  where  $s_n \in \{0, 1\}$ . A *translation symmetry* describes the case where the variables can be shifted. The translation symmetry is different from the scaling and sign-flip symmetries in that the translation can depend on other variables, i.e.,

$$\sigma(\theta) = \theta + t(\theta)$$

where  $t(\theta)$  can depend on the variables not being translated. Detection of translation symmetries still involves aggregating factor constraints but with the difference that the system of equations is now nonlinear.

### 1.5 Symmetry breaking

We provide two examples of how symmetry breaking has been handled in the literature. Most discussions of symmetry breaking were in the context of specific models that have known symmetries. For example, the Rasch model used in political science and given by

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}(\gamma_{k,i}(\alpha_{j,i} - \beta_{k,i}))$$

has a scaling symmetry where parameter  $\gamma_k$  can be scaled by a constant and parameters  $\alpha_j$  and  $\beta_k$  divided by the same constant. Bafumi et al. [Baf+05] proposed two solutions to resolve the aliasing: constrain  $\alpha_j$  to have a fixed distribution (e.g. standard normal) or introduce hyperpriors for the parameters and recover the original parameters after estimation through normalization. Stephens et al. [Ste00] looked at permutation symmetries in a mixture model. They suggested a relabelling algorithm based on loss functions as an alternative to imposing ordering constraints on the parameters. In general, these solutions along with others that have been suggested in the literature tend to be specific to the model and context. We expect that the techniques used in these solutions are likely to be generalizable to deal with specific symmetries in different models.

## 1.6 Equivalence class

One formulation of symmetry detection and breaking that we will present in this report is based on the idea of *equivalence classes*. This section provides a brief review of the definitions and properties of equivalence classes that are relevant for our purposes.

**Definition 5.** Let  $A$  be a set. For all  $a, b, c \in A$ , an *equivalence relation*  $\Xi$  on  $A$  is a binary relation that satisfies the following three properties:

1. **reflexivity:**  $a \Xi a$
2. **symmetry:** if  $a \Xi b$  then  $b \Xi a$
3. **transitivity:** if  $a \Xi b$  and  $b \Xi c$  then  $a \Xi c$

An equivalence relation  $\Xi$  partitions  $A$  into subsets called equivalence classes. For  $a, b \in A$ , if  $a \Xi b$  then  $a$  and  $b$  belong to the same equivalence class. We denote the equivalence class of  $a$  as  $[a]$ . The set of all equivalence classes in  $A$  with respect to  $\Xi$  is called the *quotient set* of  $A$  by  $\Xi$ , denoted  $A/\Xi$ .

**Definition 6.** A *section* is a function  $f : A/\Xi \rightarrow A$  that maps an equivalence class to one of its members. The member  $f([a])$  is called the *representative* of  $[a]$  with respect to  $f$ .

## 2 Formulation based on equivalence classes

The first formulation of symmetry detection and breaking that we introduce is based on the idea of equivalence classes. Symmetry detection can be framed as a problem of finding the set of variables that are symmetric (an equivalence class) for a given factor graph. Similarly, symmetry breaking can be framed as choosing a single member (representative) from the set of variables based on some desired criteria. We formalize these ideas mathematically.

**Definition 7.** Consider a factor graph  $(\theta, F)$ . Define  $\Xi$  to be the equivalence relation such that for  $\theta, \theta^* \in \Theta$ , we have  $\theta \Xi \theta^*$  if and only if there exists a local symmetry  $\sigma$  such that  $\sigma(\theta) = \theta^*$ .

We show that  $\Xi$  is a proper equivalence relation.

*Proof.* Consider  $\theta_1, \theta_2, \theta_3 \in \Theta$ .

1. **reflexivity:** let  $\sigma$  be the identity map. Then trivially,  $F_k(\theta_1) = F_k(\sigma(\theta_1))$ .
2. **symmetry:** suppose  $\theta_1 \Xi \theta_2$ . Then there exists some local symmetry  $\sigma$  such that  $\sigma(\theta_1) = \theta_2$  and  $F_k(\theta_1) \propto F_k(\theta_2)$ . By definition of symmetry,  $\sigma^{-1}$  exists and  $\sigma^{-1}(\theta_2) = \theta_1$ . Then

$$F_k(\theta_2) \propto F_k(\theta_1) = F_k(\sigma^{-1}(\theta_2))$$

and hence  $\theta_2 \Xi \theta_1$ .

3. **transitivity:** suppose  $\theta_1 \Xi \theta_2$  and  $\theta_2 \Xi \theta_3$ . Then there exists local symmetries  $\sigma_1, \sigma_2$  such that

$$\begin{aligned} \sigma_1(\theta_1) &= \theta_2 & F_k(\theta_1) &\propto F_k(\theta_2) \\ \sigma_2(\theta_2) &= \theta_3 & F_k(\theta_2) &\propto F_k(\theta_3) \end{aligned}$$

Let  $\sigma_3 = \sigma_2 \circ \sigma_1$  and so  $\sigma_3$  is also measurable with a measurable inverse. Then

$$\sigma_3(\theta_1) = \sigma_2(\sigma_1(\theta_1)) = \theta_3$$

and

$$F_k(\theta_1) \propto F_k(\theta_2) \propto F_k(\theta_3) = F_k(\sigma_3(\theta_1))$$

and hence  $\theta_1 \Xi \theta_3$ .

□

Notice that  $\Xi$  can only be defined in the context of a given factor graph  $(\theta, F)$ . It is therefore mathematically more convenient to consider factor graphs equipped with a  $\Xi$  that corresponds to a specific type of symmetry. We denote this as  $\mathcal{F} = (\theta, F, \Xi)$ . The constraint matrix  $\mathcal{C}$  of  $\mathcal{F}$  described in Section 1.2 is constructed from the constraints specified by  $\Xi$  on each factor  $F_k$ .

We now formally define symmetry detection and symmetry breaking in terms of equivalence classes.

**Definition 8.** A *symmetry detector*  $\Delta_{\mathcal{F}} : \Theta \rightarrow \Theta/\Xi$  is a measurable function that maps  $\theta$  to  $[\theta]$  with respect to  $\Xi$ .

**Definition 9.** A *symmetry breaker*  $\phi_{\mathcal{F}} : \Theta/\Xi \rightarrow \Theta$  is a measurable section that maps  $[\theta]$  to a representative  $\theta^* \in [\theta]$ .

**Definition 10.** An *automatic symmetry breaker*  $\Phi_{\mathcal{F}} = \phi_{\mathcal{F}} \circ \Delta_{\mathcal{F}} : \Theta \rightarrow \Theta$  is the composition of a symmetry breaker and its corresponding symmetry detector. It maps  $\theta$  to a representative  $\theta^*$  of its equivalence class.

The dependence of  $\Delta$ ,  $\phi$ , and  $\Phi$  on the factor graph  $\mathcal{F}$  is made clear by the subscript. When considering two functions of the same type, e.g.  $\Phi_{\mathcal{F}_1}$  and  $\Phi_{\mathcal{F}_2}$ , we will view them as being defined on the same factor graph  $(\theta, F)$  but with  $\Xi_1$  and  $\Xi_2$  corresponding to different types of symmetries.

This formulation is most useful in the case where the main concern is parameter nonidentifiability in the presence of symmetries. If inference has been made on a posterior distribution but the question remains

whether the inferred value is the “correct” one, it may instead be more interpretable to work with the symmetry-broken posterior distribution

$$\prod_k F_k(\Phi_{\mathcal{F}_M} \circ \dots \circ \Phi_{\mathcal{F}_1}(\theta))$$

where  $\Phi_{\mathcal{F}_1}, \dots, \Phi_{\mathcal{F}_M}$  correspond to  $M$  different types of symmetries that are to be broken. The corresponding  $\phi_{\mathcal{F}_1}, \dots, \phi_{\mathcal{F}_M}$  are chosen to select representatives from the respective equivalence classes based on some desired criteria.

The following two subsections show examples of how this equivalence class formulation is used in the context of scaling symmetries and permutation symmetries.

## 2.1 Scaling symmetries

We describe the scaling symmetries in Section 1.2 under the formulation based on equivalence classes.

**Definition 11.** Let  $\Sigma_d$  be the diagonal matrix with  $(e^{d_1}, \dots, e^{d_N})$  on the diagonal. A scaling symmetry  $\sigma_d$  can be written as

$$\sigma_d(\theta) = \Sigma_d \theta$$

where the right-hand side is taken as a matrix-vector product. The matrix  $\Sigma_d$  is positive definite and so the inverse exists as required by the definition of symmetry. We then define the *scaling symmetry detector* to be

$$\Delta_{\mathcal{F}}(\theta) = \{\Sigma_d \theta : d \in \mathcal{N}(\mathcal{C})\} = [\theta]$$

where  $\mathcal{C}$  is the constraint matrix of  $\mathcal{F}$ .

We show that this detector correctly maps the variables  $\theta$  to their equivalence classes  $[\theta]$  with respect to  $\Xi$ . That is, the detector maps two variables  $\theta, \theta^* \in \Theta$  to the same equivalence class if and only if there exists a scaling symmetry  $\sigma_d$  such that  $\sigma_d(\theta) = \theta^*$ .

*Proof. Sufficiency:* Suppose that  $\theta^* = \Sigma_{d^*} \theta$  for some  $d^* \in \mathcal{N}(\mathcal{C})$ . Then

$$\begin{aligned} \Delta_{\mathcal{F}}(\theta^*) &= \{\Sigma_d \theta^* : d \in \mathcal{N}(\mathcal{C})\} \\ &= \{\Sigma_d \Sigma_{d^*} \theta : d \in \mathcal{N}(\mathcal{C})\} \\ &= \{\Sigma_{d+d^*} \theta : d \in \mathcal{N}(\mathcal{C})\} \\ &= \{\Sigma_d \theta : d \in \mathcal{N}(\mathcal{C})\} \\ &= \Delta_{\mathcal{F}}(\theta) \end{aligned}$$

where the above follows because  $\Sigma_d$  and  $\Sigma_{d^*}$  are diagonal and  $d + d^* \in \mathcal{N}(\mathcal{C})$ .

*Necessity:* We show this by the contrapositive. Suppose that  $\theta^* \neq \Sigma_d \theta$  for all  $d \in \mathcal{N}(\mathcal{C})$ . Then  $\theta^* \notin \Delta_{\mathcal{F}}(\theta)$  by definition of the symmetry detector. Also, by definition of equivalence class,  $\theta^* \in [\theta^*] = \Delta_{\mathcal{F}}(\theta^*)$ . Hence  $\Delta_{\mathcal{F}}(\theta) \neq \Delta_{\mathcal{F}}(\theta^*)$  and so the detector maps  $\theta, \theta^*$  to different equivalence classes.  $\square$

It is more difficult to provide a concrete example of a *scaling symmetry breaker*  $\phi_{\mathcal{F}}$  for two reasons. The first reason is that the equivalence class is described by using the input  $\theta$  as a reference point. If  $[\theta_1]$  and  $[\theta_2]$  describe the same equivalence class, the breaker must be able to map both to the representative  $\theta^*$ . The second reason is that the equivalence class is nonlinear in the space of diagonal matrices  $\Sigma_d$  (it is described by the linear space  $\mathcal{N}(\mathcal{C})$  rather). This means that for a given  $\theta$ , operations such as scaling may not return a member of the same equivalence class. These two reasons make it challenging to mathematically define a scaling symmetry breaker that correctly extracts the representative for a given equivalence class. An example of a possible breaker worth exploring is the one that returns the minimum norm, i.e.,

$$\phi_{\mathcal{F}}([\theta]) = \arg \min_{\theta \in [\theta]} \|\theta\|$$

However, this mathematically convenient notation hides the question of how to computationally solve for the minimum norm when  $[\theta]$  may be an uncountable nonlinear space. Consideration will also need to be given in how to deal with multiple members having the same norm.

## 2.2 Permutation symmetries

While not a local symmetry, the permutation symmetry discussed in Section 1.3 can also be described in terms of equivalence classes. In this case, the definition of  $\Xi$  is modified to consider the existence of a non-local symmetry between two members of an equivalence class.

**Definition 12.** Let  $\Pi$  be a permutation matrix (a matrix where rows of the identity matrix are permuted). A permutation symmetry  $\sigma$  can be written as

$$\sigma(\theta) = \Pi\theta$$

where again the right-hand side is taken to be a matrix-vector product. Permutation matrices are orthogonal and so the inverse exists and is given by  $\Pi^T$ . We say that  $\Pi \in \mathcal{F}$  if the rows that correspond to non-permutable variables have 1 on the diagonal. The *permutation symmetry detector* is then defined to be

$$\Delta_{\mathcal{F}}(\theta) = \{\Pi\theta : \Pi \in \mathcal{F}\} = [\theta]$$

Under this definition, it is assumed that there are no restrictions on the permutations of the permutable variables. If variables are permutable with only certain others, multiple permutation symmetry detectors are needed where each detects the symmetries among a subset of permutable variables.

Notice that if  $\Pi, \Pi^* \in \mathcal{F}$ , then the matrix product  $\Pi\Pi^*$  is itself a permutation matrix by properties of permutation matrices. Furthermore,  $\Pi\Pi^* \in \mathcal{F}$  as both matrices have 1 on the diagonal in non-permutable rows and thus so does the product. We use this fact to again show that  $\Delta_{\mathcal{F}}(\theta) = \Delta_{\mathcal{F}}(\theta^*)$  if and only if there exists a permutation symmetry  $\sigma$  such that  $\sigma(\theta) = \theta^*$ .

*Proof. Sufficiency:* Suppose that  $\theta^* = \Pi^*\theta$  for some  $\Pi^* \in \mathcal{F}$ . Then

$$\begin{aligned} \Delta_{\mathcal{F}}(\theta^*) &= \{\Pi\theta^* : \Pi \in \mathcal{F}\} \\ &= \{\Pi\Pi^*\theta : \Pi \in \mathcal{F}\} \\ &= \{\Pi\theta : \Pi \in \mathcal{F}\} \\ &= \Delta_{\mathcal{F}}(\theta) \end{aligned}$$

where the above follows from the aforementioned fact.

*Necessity:* The proof for the converse follows the same argument as for the scaling symmetry. Hence this permutation symmetry detector correctly maps the variables to their equivalent classes.  $\square$

An example of a *permutation symmetry breaker* is easy to describe. For convenience, suppose that the permutable variables in  $\theta$  have distinct values. Then an example permutation breaker  $\phi_{\mathcal{F}}$  is one such that  $\phi_{\mathcal{F}}([\theta]) = \theta^*$  where if  $\theta_i$  and  $\theta_j$  are permutable, then  $\theta_i^* < \theta_j^*$ . In the context of probabilistic programming, this reduces to a sorting problem involving certain entries of  $\theta$ .

One distinction between permutation equivalence classes and scaling equivalence classes is that the permutation equivalence classes are always finite. A class can have at most  $N!$  members corresponding to the possible permutations of the variables. This makes the permutation symmetry breaker computationally simpler compared to the scaling symmetry breaker in that in the worst case, the breaker can just iterate over the equivalence class to pick a representative.

## 2.3 Other symmetries

We end this section with a brief comment about describing the other symmetries discussed in Nishihara et al. under the equivalence class formulation. The formulation for the sign-flip symmetry would likely closely resemble that of the scaling symmetry where the symmetries are given by the null space of the constraint matrix. An added benefit is that the space of symmetries is also finite due to having only two possible values for each exponent. On the other hand, we expect the translation symmetry to be difficult to describe under the equivalence formulation. The dependence on  $\theta$  in the symmetry itself makes it challenging to define detection as a function in general notation.



### 3 Formulation based on reparameterization

The formulation based on equivalence classes is useful as a post-processing step when the main concern of having symmetries is parameter nonidentifiability. However, when the presence of symmetries impact the performance of the inference algorithm, symmetry breaking should look to be done prior to inference. The equivalence class formulation cannot be used in this situation for two reasons. The first is that without having done inference, the equivalence classes are symbolic and generally difficult to work with. The second reason is that the symmetry breaker returns a specific representative of the equivalence class but does not actually remove the symmetry from the model. These reasons make the equivalence class formulation inappropriate for breaking symmetries pre-inference.

We present a second formulation of symmetry breaking based on reparameterizing the model which can be done before performing inference. The formulation directly builds on the symmetry detection approaches proposed in Nishihara et al. by changing the factor graph in such a way that the approaches detect no symmetries. We preface this section with a note that the ideas presented here are still relatively preliminary. In particular, further consideration of how this approach could be automated in a probabilistic program would be needed.

#### 3.1 Scaling symmetries

The approach proposed by Nishihara et al. for detecting scaling symmetries is based on constructing the constraint matrix and finding its null space. In other words, a scaling symmetry is present if the constraint matrix is underdetermined. Our approach under the reparameterization formulation is to modify the factor graph so that the constraint matrix becomes a determined system. We present this approach from two perspectives.

The first perspective looks at breaking the symmetry by reducing the number of variables in the factor graph. This corresponds to deleting columns of the constraint matrix and can be done by either removing or merging factors depending on the structure of the factor graph. For example, the ternary sum of common distributions from a single distribution family can typically be specified by another well-known distribution. If a scaling symmetry is introduced through a sum factor that takes in multiple Gaussian inputs, the sum and the Gaussian inputs may be replaced by a single Gaussian factor representing the sum. This breaks the scaling symmetry originating from the original sum factor in the factor graph. Figure 2 shows what this looks like graphically.

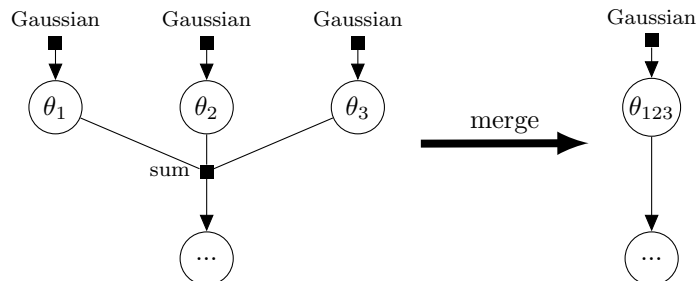


Figure 2: An example of merging factors to break a scaling symmetry.

The second perspective looks at breaking the symmetry by imposing additional constraints, which corresponds to adding new rows to the constraint matrix. The most obvious constraint that can be added is to make a variable observed. This fixes the variable so that it cannot be scaled, and is done in the constraint matrix by adding the row  $d_n = 0$ . The graphical representation of this is shown in Figure 3. Other constraints may be introduced by either adding new factors or imposing additional constraints on existing ones.

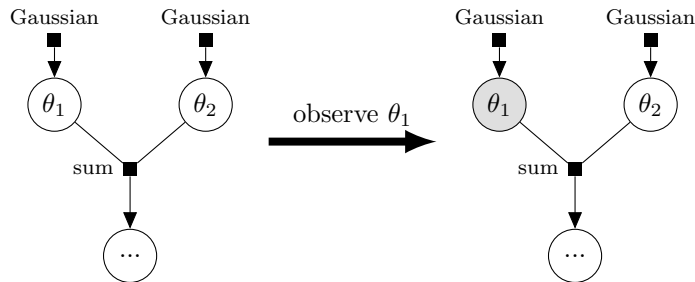


Figure 3: An example of imposing additional constraints to break scaling symmetry.

### 3.2 Permutation symmetries

Nishihara et al. reduces detection of permutation symmetries to a problem of finding the automorphism of a labeled graph. An automorphism may exist due to the symmetric arguments of a factor being annotated with the same label. The symmetry in the arguments can be removed however by imposing an additional constraint based on ordering. For example, the sum factor  $c = a + b$  is no longer symmetric with respect to  $a$  and  $b$  under the constraint  $a < b$ . This is equivalent to reparameterizing the factor graph such that the inputs to the factor are order statistics. Under this new parameterization, the inputs no longer receive the same label and so the only automorphism of the factor graph is the factor graph itself. Figure 4 shows what this reparameterization may look like graphically.

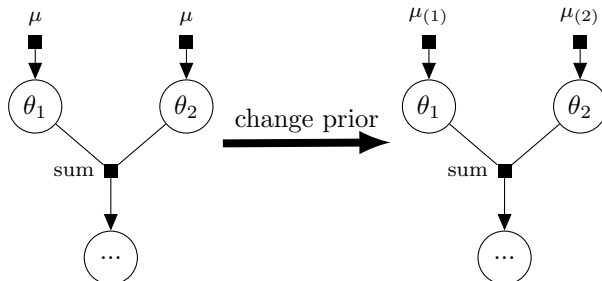


Figure 4: An example of changing the prior distribution  $\mu$  to the distributions  $\mu_{(1)}$  and  $\mu_{(2)}$  of the order statistics to break a permutation symmetry.

### 3.3 Remarks regarding reparameterization

Our formulation of symmetry breaking via reparameterization covers a broad class of techniques and closely resembles the symmetry breaking done in the literature. However, the flexibility that this formulation allows makes it difficult to automate for a general model in a probabilistic programming setting. In addition, changing the factor graph is also likely to change the posterior distribution or at least its assumptions. For example, there may be a practical reason to include multiple Gaussian inputs instead of a single combined Gaussian input. Fixing a variable for mathematical convenience is also something to question. An ideal solution to this problem would be to reparameterize the model in a way such that the original model could be recovered after inference.

## 4 Open questions and research directions

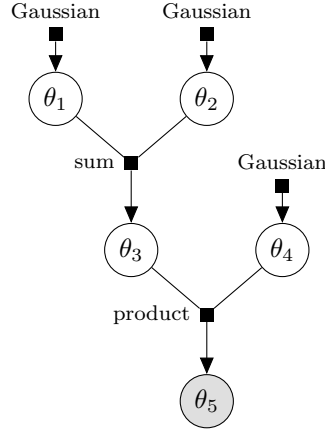
Parameter symmetries in models lead to problems of parameter nonidentifiability and reduced inference performance. Although these symmetries can often be dealt with manually as seen in the literature [Baf+05; Ste00], it is not as easy to handle them automatically in a probabilistic programming context. The paper by Nishihara et al. [NMT13] presents multiple algorithms for automatically detecting certain types of symmetries. Based on their work, we introduced two mathematical formulations of symmetry detection and breaking that have potential for automation. We conclude this report with a summary of both formulations and their possible research directions.

The equivalence class formulation frames the problem of detection and breaking as mapping the variables to a specific set of variables that they are symmetric to. The formulation is most useful as a post-processing step after inference as an intuitive solution to parameter nonidentifiability. Describing both detection and breaking as mathematical functions also naturally leads to direct translations of functions in a probabilistic programming language. The main challenges in the equivalence class formulation include specifying valid symmetry breakers and how they could be implemented to handle general classes that are nonlinear spaces. There has been some work in formalizing functions defined on equivalence classes [Pau06] that may be of relevance for making progress on these obstacles. Another challenge is the question of how the translation symmetries described by Nishihara et al. could be formulated as functions given their dependence on the input variables. Drawing on results from other areas of mathematics such as group theory may be necessary to solve these problems.

The reparameterization formulation directly works off of the algorithms by Nishihara et al. and aims to break symmetries by changing the factor graph. This formulation allows for symmetries to be broken pre-inference and is appropriate in situations where the performance of the inference algorithm is expected to be affected by the presence of symmetries. However, our formulation of reparameterization is highly dependent on the structure and context of the factor graph. This makes it difficult to implement for a general model in its current state. Further development of this formulation is needed to determine how feasible it is to automate this approach. A research direction worth exploring is understanding the class of factor graph transformations that do not change the model or at least allow for the original model to be recovered after inference. Such transformations would mitigate the drawback of implicitly changing the model when the factor graph is changed. The paper by Gelman [Gel04] discusses some reparameterization techniques for improving computability of Bayesian models. We expect that some of these techniques can be repurposed to break parameter symmetries in a model.

## A Exercises

1. This exercise is a worked extended example from Nishihara et al. [NMT13]. The factor graph from Figure 1 is shown again here.



Some factors and the scaling constraints that they impose is provided in the following table.

factor	constraints
$c = a + b$	$d_a = d_b = d_c$
$c = a \times b$	$d_c = d_a + d_b$
$c$ observed	$d_c = 0$

- (a) Find the scaling constraint matrix  $\mathcal{C}$  for this factor graph.

*Solution.* The sum factor adds the rows  $d_1 - d_3 = 0$  and  $d_2 - d_3 = 0$ .

The product factor adds the row  $d_3 + d_4 - d_5 = 0$ .

The observed factor adds the row  $d_5 = 0$ .

Aggregating the rows produces the following constraint matrix.

$$\mathcal{C} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- (b) Find the null space of  $\mathcal{C}$ .

*Solution.* The null space of  $\mathcal{C}$  is the solution to the following system of equations:

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving for the solution gives the null space

$$\mathcal{N}(\mathcal{C}) = \left\{ c \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ 0 \end{bmatrix} : c \in \mathbb{R} \right\}$$

- (c) Under the equivalence class formulation, give the definition of the scaling symmetry detector for this factor graph  $\mathcal{F} = (\theta, F, \Xi)$  where  $\Xi$  is the equivalence relation for the scaling symmetry.

*Solution.* The scaling symmetry detector is defined as

$$\Delta_{\mathcal{F}}(\theta) = \{\Sigma_d \theta : d \in \mathcal{N}(\mathcal{C})\}$$

where  $\Sigma_d$  is the diagonal matrix with  $(e^{d_1}, \dots, e^{d_5})$  on the diagonal and where  $\mathcal{N}(\mathcal{C})$  is as found in the previous question.

- (d) Given  $\theta^* = (\theta_1^*, \dots, \theta_5^*)^T$ , what is  $[\theta^*]$ ?

*Solution.* The equivalence class of  $\theta^*$  can be found using the symmetry detector defined in the previous question.

$$[\theta^*] = \Delta_{\mathcal{F}}(\theta^*) = \{\Sigma_d \theta^* : d \in \mathcal{N}(\mathcal{C})\} = \left\{ \begin{bmatrix} e^c \theta_1^* \\ e^c \theta_2^* \\ e^c \theta_3^* \\ e^{-c} \theta_4^* \\ \theta_5^* \end{bmatrix} : c \in \mathbb{R} \right\}$$

2. Consider the equivalence class formulation. Suppose that a factor graph  $\mathcal{F}$  has a scaling symmetry such that every equivalence class  $[\theta]$  has some  $\theta^*$  with Euclidean norm  $\|\theta^*\| = 1$ . Define

$$\phi_{\mathcal{F}}([\theta]) = \{\theta : \theta \in [\theta], \|\theta\| = 1\}$$

Show that  $\phi_{\mathcal{F}}([\theta])$  is a singleton for any  $[\theta]$ , i.e., that  $\phi_{\mathcal{F}}$  is a scaling symmetry breaker. (*Remark: here we use a singleton to denote the representative to allow this  $\phi_{\mathcal{F}}$  to be defined in convenient notation.*)

*Solution.* We show this by contradiction.

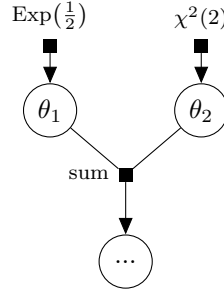
Suppose that for some equivalence class  $[\theta]$ ,  $\phi_{\mathcal{F}}([\theta])$  returns a set containing two members  $\theta^*$  and  $\theta$ . By definition of  $[\theta]$ , we must have  $\theta^* = \Sigma_{d^*} \theta$  for some  $d^* \in \mathcal{N}(\mathcal{C})$ . Then

$$\|\theta^*\| = \|\Sigma_{d^*} \theta\| = 1 = \|\theta\| \iff \Sigma_{d^*}^T \Sigma_{d^*} = \mathcal{I}$$

where  $\mathcal{I}$  is the identity matrix. Hence  $\Sigma_{d^*}$  must be orthogonal. But because  $\Sigma_{d^*}$  is diagonal and positive definite,  $\Sigma_{d^*}$  must be the identity matrix and so  $\theta^* = \mathcal{I} \theta = \theta$ . Contradiction and thus  $\phi_{\mathcal{F}}$  returns a singleton for any  $[\theta]$ .

We note that this  $\phi_{\mathcal{F}}$  is only well-defined under the assumption that every equivalence class has some  $\theta$  with unit norm. This is generally not true in all factor graphs. For example, a factor graph that has no scaling symmetries has every  $\theta$  in its own equivalence class but  $\theta$  may not have a unit norm.

3. Consider the following factor graph that has an exponential prior with rate  $\beta = \frac{1}{2}$  and a chi-squared prior with degrees of freedom  $d = 2$ :



This factor graph has a permutation symmetry. Explain why and suggest one way to break the symmetry under the reparameterization formulation.

Let  $\lambda$  denote the Lebesgue measure. You may find the following useful:

$$\begin{aligned}\mathbb{P}(\theta_1 \in dx) &= \lambda(dx) \beta e^{-\beta x} & x \in \mathbb{R}_+ \\ \mathbb{P}(\theta_2 \in dx) &= \lambda(dx) \frac{1}{2^{\frac{d}{2}} \Gamma(\frac{d}{2})} x^{\frac{d}{2}-1} e^{-\frac{x}{2}} & x \in \mathbb{R}_+\end{aligned}$$

*Solution.* Notice that

$$\mathbb{P}(\theta_1 \in dx) = \lambda(dx) \frac{1}{2} e^{-\frac{1}{2}x} \quad x \in \mathbb{R}_+$$

and

$$\mathbb{P}(\theta_2 \in dx) = \lambda(dx) \frac{1}{2^{\frac{2}{2}} \Gamma(\frac{2}{2})} x^{\frac{2}{2}-1} e^{-\frac{x}{2}} = \lambda(dx) \frac{1}{2} e^{-\frac{x}{2}} \quad x \in \mathbb{R}_+$$

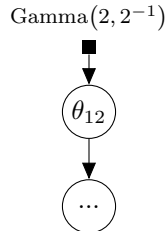
and so  $\theta_1$  and  $\theta_2$  have the same distribution. Hence the permutation symmetry exists because the inputs to the sum factor are symmetric. In this case, the symmetry can be broken by merging the sum factor with the priors as the sum is specified by a gamma distribution with shape  $\alpha = 2$  and rate  $\beta = \frac{1}{2}$ . We show this using the Laplace transform. For  $r \in \mathbb{R}_+$ , we have

$$\mathbb{E}[e^{-r\theta_1}] = \mathbb{E}[e^{-r\theta_2}] = \frac{2^{-1}}{2^{-1} + r}$$

Then by independence of  $\theta_1$  and  $\theta_2$ ,

$$\mathbb{E}[e^{-r(\theta_1 + \theta_2)}] = \mathbb{E}[e^{-r\theta_1}] \mathbb{E}[e^{-r\theta_2}] = \left( \frac{2^{-1}}{2^{-1} + r} \right)^2$$

which corresponds to the Laplace transform of the mentioned gamma distribution. The new factor graph obtained from merging is



## References

- [Bab16] L. Babai. “Graph Isomorphism in Quasipolynomial Time”. In: *CoRR* abs/1512.03547 (2016). arXiv: [1512.03547](https://arxiv.org/abs/1512.03547). URL: <http://arxiv.org/abs/1512.03547>.
- [BES80] L. Babai, P. Erdős, and S. M. Selkow. “Random Graph Isomorphism”. In: *SIAM Journal on Computing* 9.3 (1980), pp. 628–635. DOI: [10.1137/0209047](https://doi.org/10.1137/0209047). eprint: <https://doi.org/10.1137/0209047>. URL: <https://doi.org/10.1137/0209047>.
- [Baf+05] J. Bafumi et al. “Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation”. In: *Political Analysis* 13.2 (2005), pp. 171–187. DOI: [10.1093/pan/mpi010](https://doi.org/10.1093/pan/mpi010).
- [Gel04] A. Gelman. “Parameterization and Bayesian Modeling”. In: *Journal of the American Statistical Association* 99.466 (2004), pp. 537–545. DOI: [10.1198/016214504000000458](https://doi.org/10.1198/016214504000000458). eprint: <https://doi.org/10.1198/016214504000000458>. URL: <https://doi.org/10.1198/016214504000000458>.
- [NMT13] R. Nishihara, T. Minka, and D. Tarlow. *Detecting Parameter Symmetries in Probabilistic Models*. 2013. arXiv: [1312.5386](https://arxiv.org/abs/1312.5386) [stat.ML].
- [Pau06] L. C. Paulson. “Defining Functions on Equivalence Classes”. In: *ACM Trans. Comput. Logic* 7.4 (2006), pp. 658–675. ISSN: 1529-3785. DOI: [10.1145/1183278.1183280](https://doi.org/10.1145/1183278.1183280). URL: <http://doi.acm.org/10.1145/1183278.1183280>.
- [Rai17] T. Rainforth. “Automating Inference, Learning, and Design using Probabilistic Programming”. PhD thesis. University of Oxford, 2017. URL: <http://www.robots.ox.ac.uk/~twgr/assets/pdf/rainforth2017thesis.pdf>.
- [Ste00] M. Stephens. “Dealing With Label-Switching in Mixture Models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62 (July 2000). DOI: [10.1111/1467-9868.00265](https://doi.org/10.1111/1467-9868.00265).