# Formulating Detection and Breaking of Parameter Symmetries

Kenny Chiu

December 4, 2019

# 1 Background

[TODO context]

[NMT13]

We represent probabilistic models using factor graphs $(\theta, F)$ of variables $\theta = (\theta_1, ..., \theta_N)$ and factors $F = (F_1, ..., F_K)$. 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. 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 $F_k$.

A symmetry $\sigma : \Theta \to \Theta$ 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.

A local symmetry is a symmetry $\sigma$ that satisfies

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

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.

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 probabilistic programming, 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 following subsections provide examples of such annotations and how they are used to detect specific types of symmetries.

## 1.1 Scaling symmetries

A scaling symmetry is a local symmetry $\sigma$ such that

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

where $r_n \in \mathbb{R}_+$ and $d_n = \log r_n$. Examples of the constraints that factors impose on the scaling of the variables are provided in Table 1. For example, a 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 $d = (d_1, ..., d_N)$. Consider the matrix $\mathcal{C}$ constructed by stacking all the constraints in a given factor graph. For example, the sum factor above adds the rows $d_a - d_c = 0$ and $d_b - d_c = 0$ to $\mathcal{C}$. The 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 described 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.

## 1.2 Permutation symmetries

A permutation symmetry is a (non-local) symmetry $\sigma$ that permutes the components of $\theta$. Such a symmetry may be present in models that have mixture components or latent features, and is commonly characterized by the label-switching problem.

Nishihara et al. describe an approach to detect permutation symmetries that have the property

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

for all factor labels $c$. To do so, the factors have to be labeled with their type 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, the sum factor $c = a + b$ would be labeled as a binary sum factor and the arguments $a$ and $b$ would share the same label. Given the labels, Nishihara et al. then reduces the problem to a graph automorphism problem where the permutation symmetries of the factor graph are the automorphisms that preserve the factor and edge labels. While detecting the existence of graph automorphisms is not known to be solvable in polynomial time, most graphs can be tested in linear time.

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 $(a + b) + c$.

## 1.3 Equivalence class

One formulation of symmetry detection and breaking that we present in this report will be based on the concept of equivalence classes. We include a brief review of equivalence classes in this section.

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 sets 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 $A/\Xi$.

A section is a function $f : A/\Xi \to A$ 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

Let $(\theta, F)$ be a factor graph. Let $\Xi$ be an 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. 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 symmety, $\sigma^{-1}$ 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
$$\begin{aligned} \sigma_1(\theta_1) &= \theta_2 & \qquad F_k(\theta_1) &\propto F_k(\theta_2) \\ \sigma_2(\theta_2) &= \theta_3 & \qquad F_k(\theta_2) &\propto F_k(\theta_3) \end{aligned}$$
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 $\Xi$ is only describable 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 1.1 is constructed from the constraints specified by $\Xi$ on each factor $F_k$.

We now have all the definitions needed to formulate symmetry detection and symmetry breaking in terms of equivalence classes.

- A symmetry detector $\Delta_{\mathcal{F}} : \Theta \to \Theta/\Xi$ is a measurable function that maps $\theta$ to $[\theta]$ with respect to $\Xi$.
- A symmetry breaker $\phi_{\mathcal{F}} : \Theta/\Xi \to \Theta$ is a measurable section that maps $[\theta]$ to a representative $\theta_* \in [\theta]$.
- An automatic symmetry breaker $\phi_{\mathcal{F}} \circ \Delta_{\mathcal{F}} = \Phi_{\mathcal{F}} : \Theta \to \Theta$ 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 a 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 the nonidentifiability of the parameters 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
$$\prod_k F_k(\Phi_{\mathcal{F}_M} \circ \ldots \circ \Phi_{\mathcal{F}_1}(\theta))$$
where $\Phi_{\mathcal{F}_1}, \ldots, \Phi_{\mathcal{F}_M}$ correspond to $M$ different types of symmetries that are to be broken. The corresponding $\phi_{\mathcal{F}_1}, \ldots, \phi_{\mathcal{F}_M}$ are chosen to select representatives from the respective equivalence classes based on some desired criteria.

We show an example to demonstrate this idea.

## 2.1 Scaling symmetries

We describe the scaling symmetries in Section 1.1 under the formulation based on equivalence classes. Let $\Sigma_d$ be the diagonal matrix with $(e^{d_1}, ..., e^{d_N})$ on the diagonal. Then a scaling symmetry $\sigma_d$ can be written as

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

We define the scaling symmetry detector as

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

where $\mathcal{C}$ is the constraint matrix of $\mathcal{F}$. We show that this detector correctly maps the variables $\theta$ to their equivalence classes $[\theta]$ induced by $\Xi$. 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 $\sigma$ such that $\sigma(\theta) = \theta^*$.

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 $d + d^* \in \mathcal{N}(\mathcal{C})$.

We show the converse 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. 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.

A scaling symmetry breaker $\phi_{\mathcal{F}}$ is more difficult to formalize 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 $\mathcal{N}(\mathcal{C})$). 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 correctly identify the representative for the 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, consideration will need to be given in how to deal with multiple members having the same minimum norm and also in how this could be solved computationally for an uncountable $[\theta]$.

Under the assumption that the representative $\theta^*$ is known for all descriptions $[\cdot]$ of any equivalence class in $\Theta/\Xi$, it is straightforward to transform any $\theta$ to the representative of its class. Note that

$$\theta = \Sigma_{d^*} \theta^*$$

for some $d^* \in \mathcal{N}(\mathcal{C})$. Then because $\Sigma_{d^*}$ is positive definite, the inverse exists and hence

$$\Sigma_{d^*}^{-1} \theta = \theta^*$$

## 2.2 Permutation symmetries

While not a local symmetry, the permutation symmetry discussed in Section 1.2 can also be formulated 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. Let $\Pi$ be a permutation matrix (a matrix where rows of the identity matrix are permuted). Then a permutation symmetry $\sigma$ can be written as

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

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 can then be defined as

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

Note 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 for non-permutable rows and thus so does the product.

This fact directly shows that $\Delta_{\mathcal{F}}(\theta) = \Delta_{\mathcal{F}}(\theta^*)$ if there exists a permutation symmetry $\sigma$ such that $\sigma(\theta) = \Pi^*\theta = \theta^*$, $\Pi^* \in \mathcal{F}$, as

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

The proof for the converse follows the same argument for the scaling symmetry. Hence this permutation symmetry detector correctly maps the variables to their equivalent classes.

An example of a permutation symmetry breaker is easy to formulate. 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 $\theta_i^* < \theta_j^*$. This breaker is well-defined as there are only at most $N!$ possible permutations of $\theta$. In the context of a probabilistic programming, this reduces to applying a sorting algorithm to certain entries of $\theta$.

# 3 Formulation based on reparameterization

The formulation based on equivalence classes is useful when the main concern of having symmetries is parameter nonidentifiability. In this case, it is assumed that the inference algorithm is able to run without issues and that the symmetries can be broken post-inference. However, there may also be cases where the presence of symmetries impact the performance of the inference algorithm. For example, sampling algorithms may converge at a slower rate when variables are strongly correlated due to the existence of a symmetry. In this situation, the symmetry should be broken before running inference. We present a formulation of symmetry breaking based on reparameterizing the model. 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.

## 3.1 Scaling symmetries

Nishihara et al. discussed an approach for detecting scaling symmetries by constructing and finding the null space to the constraint matrix $\mathcal{C}$. In other words, a scaling symmetry is present if the constraint matrix is underdetermined. We look to change the factor graph such that the constraint matrix becomes a determined system. This can be done from two perspectives: by merging factors to reduce the number of variables or by adding factors to impose additional constraints.

Merging

# 4 Open questions and research directions

# A Exercises

# References

[NMT13] R. Nishihara, T. Minka, and D. Tarlow. *Detecting Parameter Symmetries in Probabilistic Models*. 2013. arXiv: 1312.5386 [stat.ML].