第四章　玻尔兹曼机理论基础

深度信念网络是由一种叫作受限玻尔兹曼机（Restricted Boltzmann Machine, RBM）的能量模型组成的，研究深度信念网络的不变性自然应从研究受限玻尔兹曼机入手。本章我们将介绍玻尔兹曼机的概念，包括其使用的对比分歧（Contrastive Divergence, CD）训练算法。

4.1 能量模型概念

在机器学习领域，能量模型是一种具有与模型所有状态对应的能量函数的模型。这种模型的学习过程一般可概括为修改能量函数以使其达到某种期望的特性的过程，例如，我们可以把使得某些期望状态的能量最低作为优化代价函数的目标。能量概率模型则是在能量函数的基础上为模型定义了一种概率分布，如：

$$\mathrm{P}(x) = \frac{e^{-Energy(x)}}{Z} \tag{4-1}$$

，其中归一化常数 Z 在分析物理系统被时叫作配分函数，是整个输入空间相加之和，当 x 是连续值时，则是其积分。

$$\mathrm{Z} = \sum_x e^{-\mathrm{Energy}(x)} \tag{4-2}$$

在“专家乘积”（products of experts）公式[30]中，能量函数可看作是若干项之和，每项代表一个“专家”f_i：

$$\mathrm{Energy}(x) = \sum_i f_i(x) \tag{4-3}$$

此时，能量函数所定义的概率分布为：

$$\mathrm{P}(x) \propto= P_i(x) \propto \prod_i e^{-f_i(x)} \tag{4-4}$$

在式（4-4）中，每个专家$P_i(x)$可以看作是一个关于 x 的不合理状态的探测器，也就是说，是一个 x 的加强约束。例如，我们假设$f_i(x)$只能取两个值，当约束条件得到满足时，$f_i(x)$会取一个较小值，而当约束条件未满足时，则取一个较大的值。Hinton 等人在论文[30]中通过阐述“专家混合”（mixture of experts）的缺点解释了“专家乘积”的优点。所谓的“专家混合”是指将式（4-4）中，“专家”的相乘形式换成加权和的形式。简单的说，假设每个专家与一个约束条件相关联，在混合模型中，与一个专家相关联的约束是一个指示 x 是否属于某个区域的指征，且这种指征会排除 x 处于其他区域的可能。而“专家乘积”公式可以用$f_i(x)$的集合形成一种分布式的表示，即所有专家可以一起根据可能的状态配置来划分空间，而不是像混合模型那样一个专家负责一片区域。Hinton 等人在论文[30]中提出了一种估计式（4-4）中与专家相关联的参数的算法，即后文所述的对比分歧算法。

4.1.1 引入隐变量

在许多情况下，我们不需要观测 x 的所有元素，或者我们想引入一些非观测变量来增加模型的表达能力。此时，我们会同时考虑可观测部分 x 和隐含部分 h

$$\mathrm{P}(x,\mathrm{h})=\frac{e^{-Energy(x,h)}}{Z} \tag{4-5}$$

由于只有 x 是可观测的，我们只关心如下的边缘分布

$$\mathrm{P}(x)=\sum_h\frac{e^{-Energy(x,h)}}{Z} \tag{4-6}$$

在这种情况下，为了将此公式写成形如式(4-1)的形式，我们引入自由能（Free Energy）的概念，定义如下：

$$\mathrm{P}(x)=\frac{e^{-\mathrm{Free}Energy(x)}}{Z} \tag{4-7}$$

，其中$\mathrm{Z}=\sum_x e^{-\mathrm{Free}Energy(x)}$

举例来说，自由能可以有如下形式：

$$\mathrm{FreeEnergy}(x)=-\log\sum_h e^{-\mathrm{Energy}(x,\mathrm{h})} \tag{4-8}$$

由式（4-7），当定义θ为模型的参数时，数据的对数似然函数的梯度可以有如下的形式：

$$\frac{\partial\log P(x)}{\partial\theta}=-\frac{\partial FreeEnergy(x)}{\partial\theta}+\frac{1}{Z}\sum_{\tilde{x}}P(\tilde{x})\frac{\partial FreeEnergy(\tilde{x})}{\partial\theta} \tag{4-9}$$

则对数似然函数的平均梯度为：

$$E_{\hat{P}}\left[\frac{\partial\log P(x)}{\partial\theta}\right]=-E_{\hat{P}}\left[\frac{\partial\mathrm{FreeEnergy}(x)}{\partial\theta}\right]+E_P\left[\frac{\partial\mathrm{FreeEnergy}(x)}{\partial\theta}\right] \tag{4-10}$$

，其中$\hat{P}$代表训练集的经验分布，E_P代表模型的分布 P 的期望值。因此，如果我们从 P 做采样并计算其自由能，则可以用蒙特卡罗（Monte-Carlo）方法来获得对数似然函数梯度的随机估计量。

如果把能量函数写成隐含层节点相关项之和的如下形式：

$$\mathrm{Energy}(x,h)=-\beta(x)+\sum_i\gamma_i(x,h_i) \tag{4-11}$$

此时，似然函数的分子部分可以被准确的计算，即：

$$\mathrm{FreeEnergy}(x)=-\beta(x)-\sum_i\log\sum_{h_i}\mathrm{e}^{-\gamma_i(x,h_i)} \tag{4-12}$$

另外，在受限玻尔兹曼机条件下，式（4-5）可被写成如下形式：

$$\mathrm{P}(x)=\frac{1}{Z}e^{-FreeEnergy(x)}=\frac{1}{Z}\sum_h e^{-Energy(x,h)}=\frac{e^{\beta(x)}}{Z}\prod_i\sum_{h_i}e^{-\gamma_i(x,h_i)} \tag{4-13}$$

，其中，$\sum_h e^{-Energy(x,h)}$是一个对所有 h 可能取值的加和。当 h 是连续值时，加和应被积分替代。在大部分情况下，加和与积分都是非常容易计算的，因而通过此式计算自由能或似然函数都是非常方便的。

4.1.2 条件能量模型

一般情况下，计算配分方程 Z 是非常困难的。如果我们的最终目标是在给定变量 x 的情况下求 y 的值，则与其考虑(x,y)的所有情况，不如仅仅考虑在每个给定的 x 的情况下 y 的取值情况。一个常见的情况是 y 的取值范围只是一个很小的离散集，此时有：

$$\mathrm{P}(y|x)=\frac{e^{-Energy(x,y)}}{\sum_y e^{-Energy(x,y)}} \tag{4-14}$$

在这种情况下，能量函数参数的条件对数似然函数的梯度是可以被高效的计算的，相关方法已经应用于一系列的基于神经网络的概率语言模型中。

另外，能量模型不仅可以关于对数似然函数做优化，还可以在一种更一般的原则下进行优化，即选取那些可以使得能量在“正确”的信号下降低且在“错误”信号下增加的梯度。这个原则并不是训练一般概率模型所必需的，但它经常可以被用于训练给定 x 的情况下选择 y 的函数，而这种函数是许多应用所需要的。

4.2 波尔兹曼机

玻尔兹曼机（Boltzmann Machine, BM）是一个特殊形态的能量模型，而受限玻尔兹曼机是玻尔兹曼机的特殊形式。在一个玻尔兹曼机中，其能量函数是一个二阶多项式，形式如下：

$$\mathrm{Energy}(x,h)=-b'x-c'h-h'Wx-x'Ux-h'Vh \tag{4-15}$$

其中，$\mathrm{x}\in\{0,1\}^{d_x}$ 为显层节点，$\mathrm{h}\in\{0,1\}^{d_h}$ 为隐含层节点，$\theta=\{\mathrm{U},\mathrm{V},\mathrm{W},\mathrm{b},\mathrm{d}\}$ 为玻尔兹曼机模型的参数：显层节点间联系权值 U，隐层节点间联系权值 V，显层隐层间联系权值 W，显层节点的阈值 b 和隐层节点的阈值 c。矩阵 U 和 V 是对称的，对于大部分模型其对角线元素为 0，若其对角线元素不为 0，则会带来一些变体，如高斯玻尔兹曼机等。

由于隐含层节点 h 间存在联系，在式（4-13）中计算自由能的技巧无法在这里使用。然而，一种蒙特卡罗马尔科夫链（Monte Carlo Markov Chain, MCMC）采样过程可以被用于此处以便获得对于梯度的随机估计量。此时，对数似然函数的梯度可以写成如下形式：

$$\begin{aligned}\frac{\partial \log P(x)}{\partial\theta}&=-\frac{\partial \log\sum_h e^{-Energy(x,h)}}{\partial\theta}-\frac{\partial \log\sum_{x,h} e^{-Energy(x,h)}}{\partial\theta}\\&=-\sum_h P(h|x)\frac{\partial Energy(x,h)}{\partial\theta}+\sum_{x,h}P(x,h)\frac{\partial Energy(x,h)}{\partial\theta}\end{aligned} \tag{4-16}$$

，$\frac{\partial Energy(x,h)}{\partial\theta}$是非常容易计算的。因而如果可以对$P(h|x)$和$P(x,h)$进行采样，我们即可以获得对于对数似然函数梯度的无偏随机估计量，Hinton 等人在论文[31]中阐述了这一过程：在正向阶段（positive phase），x 为可观测到的输入向量，接

着我们从给定的 x 中采样出 h；在反向阶段（negative phase），x 和 h 都被从模型本身采样出来。在此过程中，准确的采样是难以计算的，因而一般会使用近似的采样，例如，使用迭代过程来建立一个蒙特卡罗马尔科夫链。具体地说，蒙特卡罗马尔科夫链是基于吉布斯采样（Gibbs Sampling）的，对于 N 个随机变量$X_{1\ldots}X_N$的吉布斯采样是通过 N 次子采样序列得到的，其形式如下：

$$X_i \sim P(X_i|X_{-i} = x_{-i}) \tag{4-17}$$

，其中X_{-i}是 X 中除X_i外的其他 N-1 个随机变量。在这 N 个采样被得到后，蒙特卡罗马尔科夫链过程中的一个步骤即被完成，而当这种步骤的进行了无穷多次时，我们即可以获得收敛于分布$P(\mathrm{X})$的一个采样。

令$\mathrm{y} = (\mathrm{x}, \mathrm{h})$代表玻尔兹曼机中的所有节点，则$\mathrm{y}_{-i}$是除了第 i 个节点之外所有节点的可能取值的集合。若用向量 d 和矩阵 A 代表所有的参数，则玻尔兹曼机的能量函数可以被改写为

$$\mathrm{Energy}(\mathrm{y}) = -d'y - y'Ay \tag{4-18}$$

令d_{-i}表示排除了元素d_i的向量 d，A_{-i}表示不包括第 i 行和第 i 列的矩阵 A，A_i表示 A 的第 i 行或第 i 列向量，但不包括其中的第 i 个元素。$P(y_i|y_{-i})$可以很容易地被从玻尔兹曼机中计算和采样。例如，当$y_i \in \{0,1\}$时，有：

$$P(y_i = 1|y_{-i}) = \frac{\exp(d_i + d'_{-i}y_{-i} + \mathrm{A}'_i y_{-i} + y'_{-i}A_{-i}y_{-i})}{\exp(d_i + d'_{-i}y_{-i} + \mathrm{A}'_i y_{-i} + y'_{-i}A_{-i}y_{-i}) + exp(d'_{-i}y_{-i} + y'_{-i}A_{-i}y_{-i})}$$

$$= \frac{\exp(d_i + \mathrm{A}'_i y_{-i})}{\exp(d_i + \mathrm{A}'_i y_{-i}) + 1} = \frac{1}{1 + \exp(-d_i - \mathrm{A}'_i y_{-i})} = sigm(d_i + \mathrm{A}'_i y_{-i}) \tag{4-19}$$

此式即为玻尔兹曼机中根据其他神经元y_{-i}计算神经元输出的常用公式。

每个训练样本 x 需要两条蒙特卡罗马尔科夫链，一条用于正向阶段，另一条用于反向阶段，因而计算梯度的代价会非常高，进而使得训练过程变得非常长。这也是玻尔兹曼机在 80 年代后期被使用更有效的反向传播算法的神经网络所取代的原因。然而，近期的一些研究工作显示，在某些情况下，如受限玻尔兹曼机条件下，蒙特卡罗马尔科夫链可以非常短，而这也启发了后文所述的对比分歧算法。

4.3 受限波尔兹曼机

受限玻尔兹曼机是深度信念网络的基本组成部分，因为它可以独立构成深度信念网络中的任意一层，且在预训练阶段是单独训练的。在一个受限玻尔兹曼机中，式（4-15）中的 U=0 且 V=0，即其仅在隐含层节点和显层节点间具有联系，在隐含层的节点间和显层节点间没有任何联系。对于受限玻尔兹曼机模型的高效训练算法及该算法有效性的经验证明是在最近由 Hinton 等人提出的。由于缺少显层节点间的联系和隐层节点间的联系，受限玻尔兹曼机的能量函数是双线性的，

其形式如下：

$$\mathrm{Energy}(x,h) = -b'x - c'h - h'Wx \tag{4-20}$$

将式(4-20)带入对于模型输入的自由能的分解式式(4-11)中，则$\beta(x) = b'x$，$\gamma_i(x,h_i) = h_i W_i x$，其中$W_i$是矩阵 W 的第 i 行行向量。将此式代入式（4-12）中，则受限玻尔兹曼机输入的自由能可以被表示如下：

$$\mathrm{FreeEnergy}(x) = -b'x - \sum_i \log \sum_{h_i} e^{h_i W_i x} \tag{4-21}$$

使用式（4-13）中的技巧，我们可以获得条件概率$P(h|x)$的表达式，其形式如下：

$$\begin{aligned} P(h|x) &= \frac{\exp(b'x+c'h+h'Wx)}{\sum_{\tilde{h}} \exp(b'x+c'\tilde{h}+\tilde{h}'Wx)} = \frac{\prod_i \exp(c_i h_i + h_i W_i x)}{\prod_i \sum_{\tilde{h_1}} \exp(c_i \tilde{h_1} + \tilde{h_1} W_i x)} \\ &= \prod_i \frac{\exp(c_i h_i + h_i W_i x)}{\sum_{\tilde{h_1}} \exp(c_i \tilde{h_1} + \tilde{h_1} W_i x)} = \prod_i P(h_i|x) \end{aligned} \tag{4-22}$$

在常见的$h_i \in \{0,1\}$情况下，给定输入时节点输出的一般公式为：

$$P(h_i = 1|x) = \frac{e^{c_i + W_i x}}{1 + e^{c_i + W_i x}} = \mathrm{sigm}(c_i + W_i x)$$

由于 x 和 h 在能量函数中是对称的，因而可以很容易地推导出$P(x|h)$的形式如下：

$$P(x|h) = \prod_i P(x_i|h) \tag{4-23}$$

，其二值形式为:

$$P(x_j = 1|h) = \mathrm{sigm}(b_j + W_j' h) \tag{4-24}$$

，其中W_j代表矩阵 W 的第 j 列列向量。

在 Hinton 等人的论文[1]中，灰度像素被当作一个二值事件的概率，以便使用二值输入节点为其编码。这种近似在手写数字识别应用上效果良好，但在其他应用上则不然。Le Roux 等人在论文[32]中的实验显示，当输入为连续值时，使用高斯输入节点可以获得比二值输入节点更好的效果，而 Welling 等人在论文[33]中指出了 x 和 h 可以使用任意一种指数家族的分布。

尽管受限玻尔兹曼机可能并不能有效表示某些非受限玻尔兹曼机可以良好表示的概率分布，但 Freund 和 Haussler 在论文[34]指出了，在有足够的隐含层节点的情况下，受限玻尔兹曼机可以表示出任意一种离散分布。另外，在受限玻尔兹曼机已经对训练集样本分布进行完美建模前，添加一个隐含层节点总是可以提高其对数似然函数。

每个受限玻尔兹曼机的隐含层节点的二值状态可视为对输入空间作了一个2划分，则隐含层节点的所有状态将输入空间划分成了若干个区域，而某个状态可以视为其中的一个区域。注意，并不是所有的隐含层节点状态都是输入空间的非空区域。

将指数数量的隐层节点状态相加求和可以被看作“混合专家”公式的一种形

式，这种公式具有指数数量的组件（取决于参数的数量），如下所示：

$$P(x) = \sum_h P(x|h)P(h) \quad (4\text{-}25)$$

，其中$P(x|h)$是对由 h 的状态索引的组件的建模。例如，如果$P(x|h)$是符合高斯分布的，则当 h 具有 n 比特时，式（4-25）是一个具有2^n个组件的混合高斯模型。当然，这2^n个组件并不能独立地被调整，因为他们依赖于一些共享的参数（受限玻尔兹曼机参数）。

对受限玻尔兹曼机进行采样是非常有用的。首先，这对于学习算法获得对于对数似然的梯度的估计量是非常有用的。其次，对从模型中采样生成的样本进行检查可以得知模型是否获取到了数据的分布。另外，由于深度信念网络是由受限玻尔兹曼机组成的，对受限玻尔兹曼机进行采样也就使得我们可以进一步对深度信念网络进行采样。

对于一个完整的玻尔兹曼机进行采样是非常慢的，因为正向阶段（将观测输入向量赋予 x）及反向阶段（从模型中采样生成 x 和 h）均需要进行采样。另一个原因是吉布斯链中的子步骤的数量会与网络中节点的数量一样多。然而，受限玻尔兹曼机的因式分解形式有效地解决了这些问题：首先，我们不需要在正向阶段进行采样，因为自由能及其梯度具有解析形式。其次，变量的采样仅需吉布斯链中的两个子步骤即可，即先在给定 x 的情况下对 h 进行采样，再在给定 h 的情况下对 x 进行采样。在一个普通的“专家乘积”模型中，一种对吉布斯采样的替代方法是混合蒙特卡罗（hybrid Monte-Carlo），即一种在马尔科夫链的每步中加入一些计算自由能梯度的子步骤的蒙特卡罗马尔科夫链方法。因而受限玻尔兹曼机结构是“专家乘积”模型的一种特殊情况：式（4-21）中$\log \sum_{h_i} e^{W_i x h_i}$的第 i 项可以被当作一个专家，即每个隐层节点具有一个专家，每个输入偏差也具有一个专家。通过这种特殊的结构，受限玻尔兹曼机可以进行一种高效的吉布斯采样。以 k 步吉布斯采样为例，其形式如下：

$$\begin{aligned} x_0 &\sim \hat{P}(x) \\ h_0 &\sim P(h|x_0) \\ x_1 &\sim P(x|h_0) \\ h_1 &\sim P(h|x_1) \\ &\cdots \\ x_k &\sim P(x|h_{k-1}) \end{aligned}$$

4.4 对比分歧算法

对比分歧算法是一种对于对数似然函数梯度的近似估计，已被证明是训练受限玻尔兹曼机的成功算法。该算法的具体内容如下所示：

令x_1是从训练集分布中得到的采样，ϵ是对比分歧算法中随机梯度下降的学

习速率，W 是受限玻尔兹曼机的权值矩阵，其维度为（隐含层节点数量，输入节点的数量），b 是受限玻尔兹曼机隐层节点的偏差向量，c 是受限玻尔兹曼机输入节点的偏差矩阵，则节点为二值的受限玻尔兹曼机的对比分歧算法如下：

1. 对于所有隐含层节点 i，计算$\mathrm{Q}(h_{1i}=1|x_1)=\mathrm{sigm}\left(b_i+\sum_j W_{ij}x_{1j}\right)$，并从中采样出$h_{1i}$
2. 对于所有显层节点 j，计算$\mathrm{P}\left(x_{2j}=1|h_1\right)=\mathrm{sigm}\left(c_j+\sum_i W_{ij}h_{1i}\right)$，并从中采样出$x_{2j}$
3. 对于所有隐含层节点 i，计算$\mathrm{Q}(h_{2i}=1|x_2)=\mathrm{sigm}\left(b_i+\sum_j W_{ij}x_{2j}\right)$
4. 按以下几式进行更新：

$$\mathrm{W}\leftarrow \mathrm{W}+\epsilon(h_1x_1'-\mathrm{Q}(h_{2i}=1|x_2)x_2')$$
$$\mathrm{b}\leftarrow \mathrm{b}+\epsilon\left(h_1-\mathrm{Q}(h_2=1|x_2)\right)$$
$$\mathrm{c}\leftarrow \mathrm{c}+\epsilon(x_1-x_2)$$

由上可见，对比分歧算法是通过对比真实训练样本输入和模型采样输入的差距来进行的。我们可以把这种无监督训练过程看成寻找一个将高概率区域（即有许多可观测的训练样本的区域）与其他的区域分开的决策面（decision surface）。当模型产生的样本处于错误的一边时，我们将对其作出某种惩罚，而决定这种决策面的一个好办法即对比训练样本与模型生成的样本。

为了实现该算法，我们需要做的第一个近似估计是将式（4-10）中的第二项，即所有可能输入的平均值，替换为一个单独的采样。由于我们总是在更新参数的值（如为一个训练样本做的随机梯度更新或为几个训练样本做的小批梯度更新），在此过程中已经进行了一些求平均的步骤，而在连续的参数更新中，由一个或一些蒙特卡洛马尔科夫链采样替代整体加和求均值所引入的额外方差可以被在线梯度更新（online gradient update）过程部分的抵消。不过无论如何，对比分歧算法的这种对梯度的近似估计总会带来一些额外的方差。

运行一个长的蒙特卡罗马尔科夫链过程在计算代价上是非常昂贵的。因而，k 步对比分歧算法引入了第二种近似估计过程，而这个近似估计与对数似然的梯度相比会有一些偏差，其具体过程如下：从观测样本 x 开始运行 k 步蒙特卡罗马尔科夫链，则对参数的第 k 步的更新式为：

$$\Delta\,\theta=-\frac{\partial FreeEnergy(x)}{\partial\theta}+\frac{\partial FreeEnergy(\tilde{x})}{\partial\theta}$$

其中$\tilde{x}$是我们的马尔科夫链运行到第 k 步后的采样样本。我们知道当 k 趋于无穷时，偏差即会消失，同时，当模型的分布与实际分布非常接近时，若我们从实际分布的一个采样开始运行马尔科夫链，则这条蒙特卡罗马尔科夫链实际上已经收敛了，而我们仅需要一个步骤来从模型分布中获得无偏样本。

实验结果惊人地显示，即使 k=1，对比分歧算法仍可以有非常好的效果。论

文[35]给出了 k 步对比分歧算法与准确的对数似然梯度的试验比较。在那些实验中，虽然当 k 大于 1 时会有更精确的结果，但是 k=1 时已经能得到非常好的近似结果。

一种对于对比分歧算法的理论解释认为其是在训练点x_1附近对于对数似然梯度进行近似估计的。具体地说，对于能量模型的训练算法来说，最重要的即是使观测输入的能量（自由能是能量函数对隐含层节点做边缘化得到的）变小，进而将能量传递到他处，特别是其周围邻居。而随机重建的x_{k+1}的分布在某种意义上是以x_1为中心的，而且随着 k 的增加这种中心性逐渐减弱，直到变为模型的分布。k 步对比分歧算法将减少训练点x_1的自由能（当其他的自由能不变时，这会增加似然），进而增加其邻居x_{k+1}的自由能。注意，x_{k+1}虽然是x_1的邻居，但其更倾向于出现在模型具有高概率的区域（尤其是 k 较大时）。

第五章 通过信息几何度量受限玻尔兹曼机的不变性

在这部分中，我们将首先给出一种在概率流形上度量不变性的方法。接着，本论文将关注于两种特殊的玻尔兹曼机：

1. 无隐含层的单层玻尔兹曼机（single layer Boltzmann Machine without hidden units, SBM），其能量函数如下：

$$Energy_{\mathrm{SBM}}(x) = -\frac{1}{2}x'U\mathrm{x} - b'x \tag{5-1}$$

2. 受限玻尔兹曼机（Restricted Boltzmann Machine, RBM），即只有显层和隐层联系的玻尔兹曼机，其能量函数如下：

$$Energy_{\mathrm{RBM}}(x) = -\frac{1}{2}x'W\mathrm{x} - b'x - d'h \tag{5-2}$$

我们将先使用前述的度量方法来研究单层玻尔兹曼机的不变性，继而给出单层玻尔兹曼机和受限玻尔兹曼机的关系。

在本论文余下的部分中，我们将数据及其经过某种变换得到的变体数据称为相关数据。另一方面，两组数据被称为无关意味着他们在某种任务中不能被归结到同一事物。以人脸识别任务为例，一个人的图片以及此人不同穿着、发型的其他图片是相关的，因为他们是同一个人，而此人与其他人的图片则是无关的。

5.1 通过信息几何方法度量概率流形上的不变性

考虑数据 X，其背景分布为概率流形上 S 的一个点Q_x。从信息几何角度看，某模型 GM 对 X 的估计问题可看作如下过程：GM 将Q_x从流形 S 投影到 GM 所给出的流形 M 上的某点P_x。另外，在费舍尔信息度量下，点P_x是流形 M 上距离点Q_x最近的一个点。

给出 X 的相关数据$\mathrm{X}' = \mathrm{T}(X)$，其中 T 是某种变换，再给出 X 的无关数据 Y，他们的背景分布分别为流形 S 上的点Q_x'和Q_Y。模型 GM 对于点Q_x'和Q_Y的估计分布分别为流形 M 上的点P_x'和P_Y。在本论文其余部分，我们将相关数据在流形上的点称为相关点，而不相关数据在流形上的点称为无关点。

我们可以定义流形 S 上的两点被映射到流形 M 上后其距离的变化，即两点映射前后的相对距离，其形式如下：

$$\mathrm{R}(X,\mathrm{Y}) = D(P_X, P_Y)/\mathrm{D}(Q_X, Q_Y) \tag{5-3}$$

其中 D 为流形上的费舍尔信息距离。当$\mathrm{R}(X,\mathrm{Y}) < 1$时，X 和 Y 在映射到流形 M 上后变得相对更近了，也就是说，模型 GM 使其输出比其输入更相似了。

使用相对距离的概念，我们可以进一步定义度量模型 GM 的不变性程度的度