Глава 1
Выбор субоптимальной структуры модели

В данной главе рассматривается задача выбора структуры модели глубокого обучения. Предлагается ввести вероятностные предположения о распределении параметров и распределении структуры модели. Проводится градиентная оптимизация параметров и гиперпараметров модели на основе байесовского вариационного вывода. В качестве оптимизируемой функции для гиперпараметров модели предлагается обобщенная функция ее обоснованности. Показано, что данная функция оптимизирует ряд критериев выбора структуры модели: метод максимального правдоподобия, последовательное увеличение и снижению сложности модели, полный перебор структуры модели, а также получение максимума вариационной оценки обоснованности модели. Решается двухуровневая задача оптимизации: на первом уровне проводится оптимизация нижней оценки обоснованности модели по вариационным параметрам модели. На втором уровне проводится оптимизация гиперпараметров модели.

1.1. Вероятностная модель

Определим априорные распределения параметров и структуры модели следующим образом. Пусть для каждого ребра $(j,k) \in E$ и каждой базовой функции $\mathbf{g}_l^{j,k}$ параметры модели $\mathbf{w}_l^{j,k}$ распределены нормально с нулевым средним:

$$\mathbf{w}_l^{j,k} \sim \mathcal{N}\big(\mathbf{0}, \gamma_l^{j,k}(\mathbf{A}_l^{j,k})^{-1}\big),$$

где $(\mathbf{A}_l^{j,k})^{-1}$ — диагональная матрица, $l \in \{1,\dots,K^{j,k}\}$, где $K^{j,k}$ — количество базовых функций для ребра $K^{j,k}$. Априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h})$ параметров $\mathbf{w}_l^{j,k}$ зависит не только от гиперпараметров $\mathbf{A}_k^{j,k}$, но и от структурного параметра $\gamma_l^{j,k} \in (0,1)$.

В качестве априорного распределения для структуры $\boldsymbol{\Gamma}$ предлагается использовать произведение распределений Gumbel-Softmax ($\mathcal{GS}$) [**?**]:

$$p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}) = \prod_{(j,k)\in E} p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}}),$$

где для каждого структурного параметра $\boldsymbol{\gamma}^{j,k}$ с количеством базовых функций $K^{j,k}$ вероятность $p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}})$ определена следующим образом:

$$p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}}) = (K-1)!(\lambda_{\text{temp}})^{K-1} \prod_{l=1}^{K^{j,k}} s_l^{j,k}(\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}-1} \left(\sum_{l=1}^{K^{j,k}} s_l^{j,k}(\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}}\right)^{-K^{j,k}},$$

где $\mathbf{s}^{j,k} \in (0,\infty)^{Kj,k}$ — гиперпараметр, отвечающий за смещенность плотности распределения относительно точек симплекса на $K^{j,k}$ вершинах, $\lambda_{\text{temp}} > 0$ —

метапараметр температуры, отвечающий за концентрацию плотности вблизи вершин симплекса или в центре симплекса.

Перечислим свойства, которыми обладает распределение Gumbel-Softmax:

1. Компонента l случайной величины $\boldsymbol{\gamma}^{j,k}$ представима следующим образом:

$$\boldsymbol{\gamma}_l^{j,k} = \frac{\exp(\log s_l^{j,k} + g_l^{j,k})/\lambda_{\text{temp}}}{\sum_{l'=1}^{K^{j,k}} \exp(\log s_{l'}^{j,k} + g_{l'}^{j,k})/\lambda_{\text{temp}}},$$

   где $g^{j,k} \sim -\log\bigl(-\log \mathcal{U}(0,1)^{K^{j,k}}\bigr)$.

2. Свойство округления: $p(\boldsymbol{\gamma}_{l_1} > \boldsymbol{\gamma}_{l_2}, l_1 \neq l_2 | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \frac{s_l^{j,k}}{\sum_{l'} s_{l'}^{j,k}}$.

3. При устремлении температуры к нулю реализация $\hat{\boldsymbol{\gamma}}^{j,k}$ случайной величины концентрируется на вершинах симплекса:

$$p(\lim_{\lambda_{\text{temp}}\to 0} \hat{\gamma}_l^{j,k} = 1 | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \frac{s_l}{\sum_{l'} s_{l'}^{j,k}}.$$

4. При устремлении температуры к бесконечности плотность распределения концентрируется в центре симплекса:

$$\lim_{\lambda_{\text{temp}}\to\infty} p(\boldsymbol{\gamma}^{j,k} | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \begin{cases} \infty, \boldsymbol{\gamma}^{j,k} = \frac{1}{K^{j,k}}, l \in \{1, \dots, K^{j,k}\}, \\ 0, \text{ иначе.} \end{cases} \tag{1.1}$$

Доказательства первых трех утверждений приведены в [?]. Докажем утверждение 4.

Доказательство. Формула плотности с точностью до множителя записывается следующим образом :

$$p(\boldsymbol{\gamma}^{j,k} | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) \propto \frac{(\lambda_{\text{temp}})^{K^{j,k}-1}}{\left(\sum_{l=1}^{K^{j,k}} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\frac{K^{j,k}-1}{K}\lambda_{\text{temp}}} \prod_{l'=1}^{K^{j,k}} [l \neq l'] (\boldsymbol{\gamma}_l^{j,k})^{\frac{1}{K^{j,k}}\lambda_{\text{temp}}}\right)^{K^{j,k}}}. \tag{1.2}$$

Заметим, что числитель $(\lambda_{\text{temp}})^{K^{j,k}-1}$ имеет меньшую скорость сходимости, чем знаменатель, поэтому для вычисления предела достаточно проанализировать только знаменатель. Знаменатель под степенью $(-K^{j,k})$ представляется суммой слагаемых следующего вида:

$$\left(\frac{\prod_{l'\neq l} \gamma_{l'}^{\frac{1}{K^{j,k}}}}{\gamma_l^{\frac{K-1}{K^{j,k}}}}\right)^{\lambda_{\text{temp}}}. \tag{1.3}$$

Рассмотрим два случая: когда вектор $\boldsymbol{\gamma}^{j,k}$ лежит не в центре симплекса, и когда $\boldsymbol{\gamma}^{j,k}$ лежит в центре симплекса. Пусть хотя бы для одной компоненты l выполнено: $\boldsymbol{\gamma}_l^{j,k} \neq \frac{1}{K^{j,k}}$. Пусть l' соответствует индексу максимальной компоненты вектора $\boldsymbol{\gamma}^{j,k}$:

$$l' = \underset{l \in \{1,\dots,K^{j,k}\}}{\arg\max}\ \boldsymbol{\gamma}_l^{j,k}.$$

Для $l = l'$ предел выражения (1.3) при $\lambda_{\text{temp}} \to \infty$ стремится к бесконечности. Для $l \neq l'$ предел выражения (1.3) при $\lambda_{\text{temp}} \to \infty$ стремится к нулю. Возводя сумму пределов в степень $(-K^{j,k})$ получаем предел плотности, равный нулю.

Рассмотрим второй случай. Пусть $\gamma_l^{j,k} = \frac{1}{K^{j,k}}$ для всех компонент вектора $\boldsymbol{\gamma}^{j,k}$. Тогда выражение (1.2) с точностью до множителя упрощается до $(\lambda_{\text{temp}})^{K^{j,k}-1}$. Предел данного выражения стремится к бесконечности. Таким образом, предел плотности Gumbel-Softmax равен выражению (1.1), что и требовалось доказать.

$\square$

Первое свойство Gumbel-Softmax распределения позволяет использовать репараметризацию при вычислении градиента в вариационном выводе (англ. reparametrization trick).

Определение 1. Случайную величину ψ с распределеним q с параметрами $\boldsymbol{\theta}_\psi$ назовем репараметризованной через случайную величину ε, чье распределение не зависит от параметров $\boldsymbol{\theta}_\psi$, если:

$$\psi = g(\varepsilon, \boldsymbol{\theta}_\psi)$$

где g — некоторая непрерывная функция.

Идею репараметризации поясним на следующем примере.

Пример 1. Пусть структура $\boldsymbol{\Gamma}$ зафиксирована для модели $\mathbf{f}$. Рассмотрим математическое ожидание логарифма правдоподобия выборки модели по некоторому непрерывному распределению $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$:

$$\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Продифференцируем данное выражение по параметрам $\boldsymbol{\theta}_{\mathbf{w}}$ вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$, полагая что оно удовлетворяет необходимым условиям для переноса оператора дифференцирования под знак интеграла:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Это выражение в общем виде не имеет аналитического решения. Пусть распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ для параметров $\mathbf{w}$ подлежит репараметризации через случайную величину $\boldsymbol{\varepsilon}$:

$$\mathbf{w} = \mathbf{g}(\boldsymbol{\varepsilon}, \boldsymbol{\theta}_{\mathbf{w}}).$$

Тогда справедливо следующее выражение:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{\varepsilon} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}) =$$

(а) (б) (в) (г)

Рис. 1.1. Пример распределения Gumbel-Softmax при различных значениях параметров: а) $\lambda_{\text{temp}} \to 0$, б) $\lambda_{\text{temp}} = 1, \mathbf{s} = [1,1,1]$, в) $\lambda_{\text{temp}} = 5, \mathbf{s} = [1,1,1]$, г) $\lambda_{\text{temp}} = 5, \mathbf{s} = [10, 0.1, 0.1]$.

$$= \int_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}) p(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon} = \mathsf{E}_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}).$$

Таким образом, распределение, позволяющее произвести репараметризацию, является более удобным для вычисления интегральных оценок вида $\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$, а также позволяет повысить точность приближенного вычисления значений таких функций [?]. Подробный анализ репараметризации для генеративных моделей глубокого обучения представлен в [?].

Пример распределения Gumbel-Softmax при различных параметрах представлен на Рис. 1.1. В качестве альтернативы для априорного распределения структуры выступает распределение Дирихле. В качестве предельного случая, когда все структуры $\boldsymbol{\Gamma} \in \mathbb{\Gamma}$ равнозначны, выступает равномерное распределение. Выбор в качестве распределения стрктуры произведения распределений Gumbel-Softmax обоснован выбором этого распределения в качестве вариационного.

Заметим, что предлагаемое априорное распределение неоднозначно: одно и то же распределение можно получить с различными значениями гиперпараметра $\mathbf{A}_l^{j,k}$ и структурного параметра $\gamma_l^{j,k}$. В качестве регуляризатора для матрицы $(\mathbf{A}_l^{j,k})^{-1}$ предлагается использовать обратное гамма-распределение:

$$(\mathbf{A}_l^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2),$$

где $\lambda_1, \lambda_2 \in \boldsymbol{\lambda}$ — метапараметры оптимизации. Использование обратного гамма-распределения в качестве распределения гиперпараметров можно найти в [?, ?]. В данной работе обратное распределение выступает как регуляризатор гиперпараметров. Варьированием метапараметров λ_1, λ_2 получается более сильная или более слабая регуляризация [?]. Пример распределений $\text{inv-gamma}(\lambda_1, \lambda_2)$ для разных значений метапараметров λ_1, λ_2 изображен на Рис. 1.2. Оптимизации без регуляризации соответствует случай предельного распределения $\lim_{\lambda_1, \lambda_2 \to 0} \text{inv-gamma}(\lambda_1, \lambda_2)$.

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

1. Параметры $\mathbf{w}$ модели, распределенные нормально.

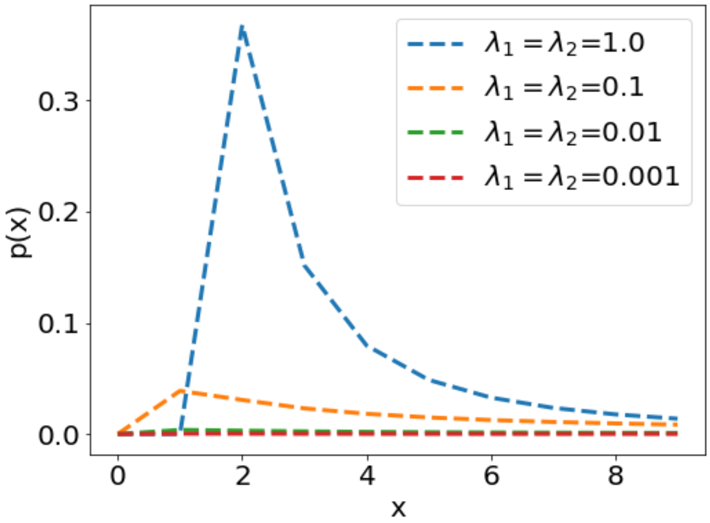

Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

2. Структура модели $\boldsymbol{\Gamma}$, содержащая все структурные параметры $\{\boldsymbol{\gamma}^{j,k}, (j,k) \in E\}$, распределенные по распределению Gumbel-Softmax.
3. Гиперпараметры $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$, где $\mathbf{A}$ — конкатенация матриц $\mathbf{A}^{j,k}, (j,k) \in E$, $\mathbf{s}$ — конкатенация параметров Gumbel-Softmax распределений $\mathbf{s}^{j,k}, (j,k) \in E$, где E — множество ребер, соответствующих графу рассматриваемого параметрического семейства моделей $\mathfrak{F}$.
4. Метапараметры: $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_{\text{temp}}]$. Эти параметры не подлежат оптимизации и задаются экспертно.

График вероятностной модели в формате плоских нотаций представлен на Рис. 1.3.

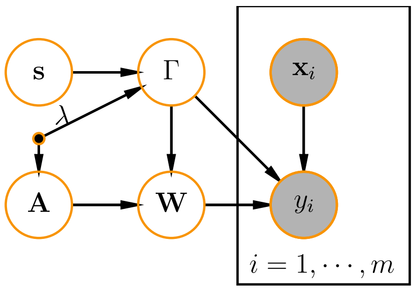

Рис. 1.3. TODO: сделать любмду красивой (снова). График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

1.2. Вариационная оценка обоснованности вероятностной модели

Задача выбора структуры $\mathbf{\Gamma}$ и параметров $\mathbf{w}$ заключается в получении оценок на апостериорное распределение $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$. Оно зависит от гиперпараметров $\mathbf{h}$. В качестве критерия выбора гиперпараметров предлагается использовать апостериорную вероятность гиперпараметров:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h}\in\mathbb{H}}. \tag{1.4}$$

Структура модели и параметры модели выбираются на основе полученных значений гиперпараметров:

$$\mathbf{w}^*, \mathbf{\Gamma}^* = \underset{\mathbf{w}in\mathbb{W}, \mathbf{\Gamma}\in\mathbb{\Gamma}}{\arg\max}\, p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}^*, \boldsymbol{\lambda}),$$

где $\mathbf{h}^*$ — решение задачи оптимизации (1.4).

Для вычисления обоснованности модели

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \iint_{\mathbf{\Gamma},\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})d\mathbf{\Gamma}d\mathbf{w}$$

из (1.4) предлагается использовать нижнюю вариационную оценку обоснованности.

Теорема 1. Пусть $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) = q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ — вариационное распределение с параметрами $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\mathbf{\Gamma}}]$, аппроксимирующее апостериорное распределение структуры и параметров:

$$q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) \approx p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}),$$

$$q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) \approx p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}),$$
$$q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) \approx p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}).$$

Тогда справедлива следующая оценка:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \geq \tag{1.5}$$

$$\mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - D_{\mathrm{KL}}\big(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})||p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)-$$
$$-D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\big),$$

где $D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\big)$ вычисляется по формуле условной дивергенции [**?**]:

$$D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\big) = \mathsf{E}_{\mathbf{\Gamma}\sim q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{\mathbf{w}\sim q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})}\right).$$

Доказательство. Перепишем обоснованность:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \log \iint_{\boldsymbol{\Gamma}, \mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) d\boldsymbol{\Gamma} d\mathbf{w} =$$

$$= \log \iint_{\boldsymbol{\Gamma}, \mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \frac{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})}{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} d\boldsymbol{\Gamma} d\mathbf{w} =$$

$$= \log \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})}.$$

Используя неравенство Йенсена получим

$$\log \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \geq \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} =$$

$$-\mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

Декомпозируем распределение q по свойству условной дивергенции:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) =$$

$$= D_{\mathrm{KL}}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) || p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) + \mathsf{E}_{\boldsymbol{\Gamma} \sim q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})} \mathsf{E}_{\mathbf{w} \sim q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} \log \left(\frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})} \right). \quad (1.6)$$

□

В качестве вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ предлагается использовать нормальное распределение, не зависящее от структуры модели $\boldsymbol{\Gamma}$:

$$q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q),$$

где $\mathbf{A}_q$ — диагональная матрица с диагональю $\boldsymbol{\alpha}_q$.

В качестве вариационного распределения $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ предлагается использовать произведение распределений Gumbel-Softmax. Конкатенацию параметров концентрации распределений обозначим $\mathbf{s}_q$. Его температуру, общую для всех структурных параметров $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, обозначим θ_{temp}. Вариационными параметрами распределения $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ являются параметры распределений $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}), q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$:

$$\boldsymbol{\theta} = [\boldsymbol{\mu}_q, \boldsymbol{\alpha}_q, \mathbf{s}_q, \theta_{\mathrm{temp}}].$$

График вероятностной вариационной модели в формате плоских нотаций представлен на Рис. 1.4.

Для анализа сложности полученной модели введем понятие *параметрической сложности.*

Определение 2. Параметрической сложностью $C_p(\boldsymbol{\theta}|U_{\mathbf{h}}, \boldsymbol{\lambda})$ модели с вариационными параметрами $\boldsymbol{\theta}$ на компакте $U_{\mathbf{h}} \subset \mathbb{H}$ назовем минимальную дивергенцию между вариационным и априорным распределением:

$$C_p(\boldsymbol{\theta}|U_{\mathbf{h}}, \boldsymbol{\lambda}) = \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

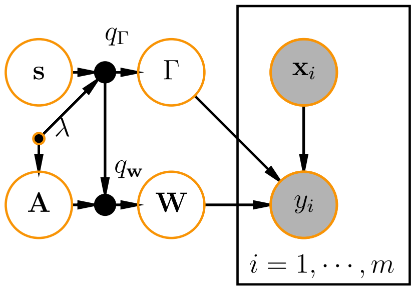

Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

Параметрическая сложность модели соответствует минимальной по $\mathbf{h} \in U_\mathbf{h}$ ожидаемой длине описания параметров модели при условии заданного параметрического априорного распределения [**?**].

Одним из критериев удаления неинформативных параметров в вероятностных моделях является отношение вариационной плотности параметров в моде распределения к вариационной плотности параметра в нуле [**?**]:

$$\frac{q_\mathbf{w}(w = \mu_q|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(w = 0|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} = \exp\left(-\frac{2\alpha_q^2}{\mu_q^2}\right),$$

где параметру модели w соответствуют вариационные параметры μ_q, α_q: $q_\mathbf{w}(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) \sim \mathcal{N}(\mu_q, \alpha_q)$.

Обобщим понятие относительной вариационной плотности на случай произвольных непрерывных распределений.

Определение 3. Относительной вариационной плотностью параметра $w \in \mathbf{w}$ при условии структуры $\mathbf{\Gamma}$ и гиперпараметров $\mathbf{h}$ назовем отношение вариационной плотности в моде вариационного распределения параметра к вариационной плотности в моде априорного распределения параметра:

$$\rho(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \lambda) = \frac{q_\mathbf{w}(\text{mode } q_\mathbf{w}(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(\text{mode } p(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})}.$$

Относительной вариационной плотностью вектора параметров $\mathbf{w}$ назовем следующее выражение:

$$\rho(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) = \prod_{w \in \mathbf{w}} \rho(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}).$$

Сформулируем и докажеми теорему о связи относительной плотности и параметрической сложности модели:

Теорема 2. Пусть

1. Заданы компактные множества $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \mathbb{\Theta}_{\mathbf{w}}, U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \subset \mathbb{\Theta}_{\boldsymbol{\Gamma}}$.
2. Мода априорного распределения $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ не зависит от гиперпараметров $\mathbf{h}$ на $U_{\mathbf{h}}$ и структуры $\boldsymbol{\Gamma}$ на $U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}}$:
$$\text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}_1, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{m}$$
для любых $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in U_{\boldsymbol{\Gamma}}$.
3. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ и априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ являются абсолютно непрерывными и унимодальными на $U_{\mathbf{h}}, U_{\boldsymbol{\theta}}$.
4. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ является липшецевым по $\mathbf{w}$.
5. Значение $q_{\mathbf{w}}(\mathbf{M}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ не равно нулю при $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.
6. Решение задачи
$$\mathbf{h}^* = \underset{\mathbf{h} \in U_{\mathbf{h}}}{\arg\min}\, D_{\text{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \tag{1.7}$$
единственно для любого $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.
7. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по распределениям:
$$\int_{\boldsymbol{\Gamma}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\boldsymbol{\Gamma}, \quad \int_{\boldsymbol{\Gamma}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) d\boldsymbol{\Gamma}.$$
8. Мода и матожидание вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ и априорного распределения $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ совпадают:
$$\text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w};$$
$$\text{mode } q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) = \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}.$$
9. Задана бесконечная последовательность векторов вариационных параметров $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_i, \dots \in U_{\boldsymbol{\theta}}$, такая что $\lim_{i \to \infty} C_p(\boldsymbol{\theta}_i|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$.

Тогда следующее выражение стремится к единице:
$$\mathsf{E}_{q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})} \rho(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) \to 1.$$

Доказательство. Обозначим за $\mathbf{h}_i$ — решение задачи (1.7) для вектора вариационных параметров $\boldsymbol{\theta}_i = [(\boldsymbol{\theta}_{\mathbf{w}})_i, (\boldsymbol{\theta}_{\boldsymbol{\Gamma}})_i]$.

Воспользуемся неравенством Пинскера:
$$||F_q((\boldsymbol{\theta}_{\mathbf{w}})_i) - F_p(\mathbf{h}_i)||_{\text{TV}}, \leq \sqrt{2 \log q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i) - \log p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda}) d\mathbf{w}},$$

где $||\cdot||_{\text{TV}}$ — расстояние по вариации, F_q, F_p — функции распределения $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}), p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$. Правая часть неравенства под корнем соответствует дивергенции Кульбака-Лейблера между распределениями $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i), p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda})$ при фиксированном значении структуры $\mathbf{\Gamma}$.

По условию дивергенция (1.6) стремится к нулю при $i \to \infty$. Она состоит из двух неотрицательных величин, поэтому обе они стремятся к нулю. Рассмотрим вторую величину:

$$0 = \lim_{i\to\infty} \mathsf{E}_{\mathbf{\Gamma}\sim q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i)} \mathsf{E}_{\mathbf{w}\sim q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},(\boldsymbol{\theta}_{\mathbf{w}})_i)} \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i)}{p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda})}\right) =$$

$$\lim_{i\to\infty}\left|\int_{\mathbf{\Gamma}}\int_{\mathbf{w}} \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i)}{p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda})}\right) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i) q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i) d\mathbf{w}d\mathbf{w}\right| \geq$$

$$\geq \lim_{i\to\infty}\int_{\mathbf{\Gamma}} ||F_q((\boldsymbol{\theta}_{\mathbf{w}})_i) - F_p(\mathbf{h}_i)||_{\text{TV}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i) d\mathbf{\Gamma}.$$

Отсюда $\lim_{i\to\infty}\int_{\mathbf{\Gamma}} ||F_q((\boldsymbol{\theta}_{\mathbf{w}})_i) - F_p(\mathbf{h}_i)||_{\text{TV}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i) d\mathbf{\Gamma} = 0$. По теореме Шеффе данное выражение можно переписать как:

$$\lim_{i\to\infty}\frac{1}{2}\iint_{\mathbf{w},\mathbf{\Gamma}} |p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, (\boldsymbol{\theta}_{\mathbf{w}})_i)| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i) d\mathbf{\Gamma}d\mathbf{w} = 0.$$

Для произвольного $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\mathbf{\Gamma}}]$ рассмотрим выражение:

$$\left|\int_{\mathbf{\Gamma}}\left(\frac{q_{\mathbf{w}}(\text{mode } q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}\right) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma} - 1\right| =$$

$$= \left|\int_{\mathbf{\Gamma}}\left(\frac{q_{\mathbf{w}}(\text{mode } q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})]}\right) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma}\right| =$$

$$= \left|\int_{\mathbf{\Gamma}}\left(\frac{q_{\mathbf{w}}(\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}\right) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma}\right| \leq$$

$$\leq \int_{\mathbf{\Gamma}}\left|\frac{q_{\mathbf{w}}(\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma}\right| \leq$$

$$\frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}\int_{\mathbf{\Gamma}} |\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w} - \mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma} \leq$$

$$\leq \frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{M}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{w}d\mathbf{\Gamma},$$

где C_l — максимальная константа Липшица для $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ на $U_{\boldsymbol{\theta}}$.

Определим случайную величину $\boldsymbol{\nu}(t), t \geq 0$ следующим образом:

$$\boldsymbol{\nu}(t) = \mathbf{max}(-t\cdot\mathbf{1}, \mathbf{min}(t\cdot\mathbf{1}, \mathbf{w})).$$

Данная величина совпадает с $\mathbf{w}$ при $|\mathbf{w}| < t$ и принимает значение t или $-t$ при $|\mathbf{w}| \geq t$, Тогда для любого $t > 0$ справедливо:

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq$$

$$\leq \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}+$$

$$+ \iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|d\mathbf{w} \leq$$

$$\leq \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|d\mathbf{w}d\mathbf{\Gamma}+ \tag{1.8}$$

$$+ \iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}.$$

Рассмотрим первое слагаемое суммы (1.8). Т.к. вторые моменты $\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w}^2, \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}^2$ конечны, то случайная величина $\mathbf{w}$ равномерно интегрируема как при маргинальном распределении $\int_{\mathbf{\Gamma}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})d\mathbf{\Gamma}$, так и при маргинальном распределении $\int_{\mathbf{\Gamma}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})d\mathbf{\Gamma}$. По определению равномерной интегрируемости для $\mathbf{w}$ для любого числа ε существует число t_0, такое что для любого $t \geq t_0$, любого $\mathbf{h} \in U_{\mathbf{h}}, \boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$, справедливо выражение:

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}|\mathbf{w} - \boldsymbol{\nu}(t)| = \iint_{\mathbf{w},\mathbf{\Gamma}} |\mathbf{w} - \boldsymbol{\nu}(t)|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq \varepsilon,$$

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}|\mathbf{w} - \boldsymbol{\nu}(t)| = \iint_{\mathbf{w},\mathbf{\Gamma}} |\mathbf{w} - \boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq \varepsilon.$$

Тогда

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|d\mathbf{w}d\mathbf{\Gamma} \leq$$

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) + \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})d\mathbf{\Gamma}d\mathbf{w} < 2\varepsilon$$

для любого $t \geq t_0$. Обозначим за $\varepsilon(t)$ минимальное число ε, удовлетворяющее предыдущим неравенствам. Тогда

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|d\mathbf{w}d\mathbf{\Gamma} \leq 2\varepsilon(t),$$

где $\lim_{t\to\infty} \varepsilon(t) = 0$.

Рассмотрим второе слагаемое.

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})d\mathbf{w}d\mathbf{\Gamma} \leq$$

$$\leq t \iint_{\mathbf{\Gamma},\mathbf{w}} |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}.$$

Переходя к пределу в (1.8) получим:

$$\lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}_i,\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i)d\mathbf{w}d\mathbf{\Gamma} =$$

$$= \lim_{t\to\infty}\lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},(\boldsymbol{\theta}_{\mathbf{w}})_i) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}_i,\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i)d\mathbf{w}d\mathbf{\Gamma} \leq$$

$$\lim_{t\to\infty}\lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)||p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}_i,\boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},(\boldsymbol{\theta}_{\mathbf{w}})_i)|d\mathbf{w}d\mathbf{\Gamma}+$$

$$+ \iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},(\boldsymbol{\theta}_{\mathbf{w}})_i) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}_i,\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i)d\mathbf{w}d\mathbf{\Gamma} \leq$$

$$\lim_{t\to\infty} 2\varepsilon(t) + \lim_{t\to\infty}\lim_{i\to\infty} t \iint_{\mathbf{\Gamma},\mathbf{w}} |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},(\boldsymbol{\theta}_{\mathbf{w}})_i) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}_i,\boldsymbol{\lambda})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|(\boldsymbol{\theta}_{\mathbf{\Gamma}})_i) = 0.$$

Таким образом выражение $\left|\int_{\mathbf{\Gamma}} \frac{q_{\mathbf{w}}(\text{mode} q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode} p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{\Gamma}\right|$ стремится к единице, что и требовалось доказать. □

Теорема утверждает, что при устремлении параметрической сложности модели к нулю, все параметры модели подлежат удалению в среднем по всем возможным значениям структуры $\mathbf{\Gamma}$ модели. Заметим, что теорема применима для случая, когда последовательность вариационных распределений q не имеет предела. Так, в случае, если структура $\mathbf{\Gamma}$ определена однозначно, последовательность $\boldsymbol{\theta}_i$ может являться последовательностью нормальных распределений, чье матожидание стремится к нулю:

$$\boldsymbol{\theta}_i \sim \mathcal{N}((\boldsymbol{\mu}_q)_i, (\mathbf{A}_q^{-1})_i), (\boldsymbol{\mu}_q)_i \to \mathbf{0}.$$

Априорным распределением $p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})$ при этом может являться семейство нормальных распределений с нулевым средним:

$$p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{0},\mathbf{A}^{-1}).$$

При этом сама последовательность распределений $\boldsymbol{\theta}_i$ не обязана иметь предел.

1.3. Обобщающая задача

В данном разделе проводится анализ основных критериев выбора моделей, а также предлагается их обобщение на случай моделей, испольюзующих вариационное распределение $q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})$ для аппроксимации неизвестного апостериорного распределения параметров $p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})$.

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \to \max_{\mathbf{w}\in U_{\mathbf{w}}, \mathbf{\Gamma}\in U_{\mathbf{\Gamma}}}.$$

   Для использования данного критерия в качестве задачи выбора модели предлагается следующее обобщение:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}). \tag{1.9}$$

   Данное обобщение (1.9) эквивалентно критерию правдоподобия при выборе в качестве $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ эмпирического распределения параметров и структуры. Метод не предполагает оптимизации гиперпараметров $\mathbf{h}$. Для формального соответствия данной задачи задаче выбора модели (**??**), т.е. двухуровневой задачи оптимизации, положим $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \to \max_{\boldsymbol{\theta}\in U_{\boldsymbol{\theta}}},$$

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \to \max_{\mathbf{h}\in U_{\mathbf{h}}},$$

2. Метод максимальной апостериорной вероятности.

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \to \max_{\mathbf{w}\in U_{\mathbf{w}}, \mathbf{\Gamma}\in U_{\mathbf{\Gamma}}}.$$

   Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{1.10}$$

$$= \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})}\big(\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) + \log p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

   Т.к. в рамках данной задачи (1.10) не предполагается оптимизации гиперпараметров $\mathbf{h}$, положим параметры распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ фиксированными:

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_{\text{temp}}, \mathbf{s}, \text{diag}(\mathbf{A})].$$

3. Полный перебор структуры:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log p(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) = p'|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \tag{1.11}$$

   где p' — некоторое распределение на структуре $\mathbf{\Gamma}$, выступающее в качестве метапараметра.

4. Критерий Акаике:

$$\text{AIC} = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) + |\mathbb{W}|.$$

Т.к. все рассматриваемые модели принадлежат одному параметрическому семейству моделей $\mathfrak{F}$, то количество параметров у всех рассматриваемых моделей совпадает. Тогда критерий Акаике совпадает с критерием максимального правдоподобия. Для использования критерия Акаике для сравнения моделей, принадлежащих одному параметрическому семейству $\mathfrak{F}$ предлагается следующая переформулировка:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \tag{1.12}$$

$$-|\{w : D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) < \lambda_{\text{prune}}\}|,$$

где

$$\mathbf{h} = \mathop{\arg\min}_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big), \tag{1.13}$$

λ_{prune} — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение (1.12) применимо только в случае, если выражение (1.13) определено однозначно, т.е. существует единственный вектор гиперпараметров $\mathbf{h} \in U_{\mathbf{h}}$, доставляющий минимум дивергенции $D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big)$.

5. Информационный критерий Шварца:

$$\text{BIC} = \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - 0.5\log(m)|\mathbb{W}|.$$

Переформулируем данный критерий аналогично критерию AIC:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \tag{1.14}$$

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \log m|\{w : D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) < \lambda_{\text{prune}}\}|,$$

метапараметр λ_{prune} определен аналогично (1.13).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \tag{1.15}$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) + p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}},$$

$$Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) =$$

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) + p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h} \in U_{\mathbf{h}}},$$

В рамках данной задачи функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ и $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ совпадают, все гиперпараметры $\mathbf{h}$ подлежат оптимизации.

7. Валидация на отложенной выборке:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})\mathbf{y}_{\text{train}}\mathbf{X}_{\text{train}} + p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}}, \tag{1.16}$$

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})\mathbf{y}_{\text{test}}\mathbf{y}_{\text{test}} \to \max_{\mathbf{h}\in U_{\mathbf{h}}},$$

где $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}), (\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$ — разбиение выборки на обучающую и контрольную подвыборку. В рамках данной задачи, все гиперпараметры $\mathbf{h}$ подлежат оптимизации.

Каждый из рассмотренных критериев удовлетворяет хотя бы одному из перечисленных свойств:
1) модель, оптимизируемая согласно критерию, доставляет максимум правдоподобия выборки;
2) модель, оптимизируемая согласно критерию, доставляет максимум оценки обоснованности;
3) для моделей, доставляющих сопоставимые значения правдоподобия выборки, выбирается модель с меньшим количеством информативных параметров.
4) критерий позволяет производить перебор структур для отбора наилучших.

Формализуем рассмотренные критерии. Оптимизационную задачу, которая удовлетворяет всем перечисленным свойствам при некоторых значинях метапараметров, будет называть *обобщающей.*

Определение 4. Двухуровневую задачу оптимизации будем называть *обобщающей* на компакте

$$U = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \mathbb{\Theta}_{\mathbf{w}} \times \mathbb{\Theta}_{\boldsymbol{\Gamma}} \times \mathbb{H} \times \lambda$$

, если она удовлетворяет следующим критериям.

1. Область определения каждого параметра $w \in \mathbf{w}$, гиперпараметра $h \in \mathbf{h}$ и метапараметра $\lambda \in \boldsymbol{\lambda}$ не является пустым множеством и не является точкой.
2. Для каждого значения гиперпараметров $\mathbf{h}$ оптимальное решение нижней (**??**) задачи оптимизации

$$\boldsymbol{\theta}^*(\mathbf{h}) = \arg\max_{\boldsymbol{\theta}\in\mathbb{\Theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$$

   определено однозначно при любых значениях метапараметров $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.
3. Критерий максимизации правдоподобия выборки: существует $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ и

$$K_1 > 0, \quad K_1 < \max_{\mathbf{h}_1,\mathbf{h}_2\in U_{\mathbf{h}}} Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}),$$

   такие что для любых векторов гиперпараметров, удовлетворяющих неравенству

$$\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}) > K_1,$$

   выполняется неравенство

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_1))} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) > \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_2))} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}).$$

4. Критерий минимизации параметрической сложности: существует $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ и

$$K_2 > 0, \quad K_2 < \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}} Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}),$$

такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$, удовлетворяющих неравенству

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}) > K_2,$$

параметрическая сложность первой модели меньше, чем второй:

$$C_p(\boldsymbol{\theta}^*(\mathbf{h}_1)|U_{\mathbf{h}}, \boldsymbol{\lambda}) < C_p(\boldsymbol{\theta}^*(h_2)|U_{\mathbf{h}}, \boldsymbol{\lambda}).$$

5. Критерий приближения оценки обоснованности: существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что значение функций потерь $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ и валидации $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ пропорционален вариационной оценки обоснованности модели:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \propto Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \propto$$

$$\propto \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) + \log p(\mathbf{h}|\boldsymbol{\lambda})$$

для всех $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}$, где в качестве гиперпараметров $\mathbf{h}$ рассматриваются все гиперпараметры модели: $\mathbf{h} = [\mathbf{A}, \mathbf{s}]$.
6. Критерий перебора оптимальных структур: существует набор метапараметров $\boldsymbol{\lambda}$ и константа $K_3 > 0$:

$$K_3 < \max_{\mathbf{h}_1, \mathbf{h}_2} \min\left(D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big), D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big)\right),$$

такие что для локальных оптимумов задачи оптимизации $\mathbf{h}_1, \mathbf{h}_2$, полученных при метапараметрах $\boldsymbol{\lambda}$ и удовлетворяющих неравенствам

$$D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) > K_3, D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big) > K_3,$$

,

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) > Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}),$$

существует значение метапараметров $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$, такие что
   (a) соответствие между вариационными параметрами $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$ сохраняется при $\boldsymbol{\lambda}'$,
   (b) выполняется неравенство $Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) < Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ при $\boldsymbol{\lambda}'$.
7. Критерий нерперывности: функции $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ и $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ непрерывны по метапараметрам $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.

Первый критерий является техническим и используется для исключения из рассмотрения вырожденных задач оптимизации. Второй критерий говорит о том, что решение первого и второго уровня должны быть согласованы и определены однозначно. Критерии 3-5 определяют возможные критерии оптимизации, которые должны приближаться обобщающей задачей. Критерий 6 говорит о возможности перехода между различными структурами модели. Данный критерий говорит о том, что мы можем перейти от одного набора гиперпараметров $\mathbf{h}_1$ к другим $\mathbf{h}_2$, если они соответствуют локальным оптимумам задачи оптимизации, и дивергенция соответствующих априорных распределений на структурах $p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ значимо высока. При этом соответствующие вариационные распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ могут оказаться достаточно близки. Возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ различаются значимо. Последний критерий говорит о том, что обобщающая задача должна позволять производить переход между различными методами выбора параметров и структуры модели непрерывно.

Теорема 3. Рассмотренные задачи (1.9),(1.10),(1.11),(1.12),(1.14),(1.16) не являются обобщающими.

Доказательство. Задачи (1.9),(1.10),(1.11),(1.12),(1.14) не имеют гиперпараметров $\mathbf{h}$, подлежащих оптимизации, поэтому не могут оптимизировать вариационную оценку.

При использовании валидации на отложенной выборки (1.16) в функцию валидации $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ не входит ни один метапараметр, поэтому критерий перебора структур 6 для нее также не выполняется.

□

Теорема 4. Пусть $q_{\mathbf{\Gamma}}$ — абсолютно непрерывное распределение с дифференцируемой плотностью, такой что:

1. Градиент плотности $\nabla_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ является нулевым не более чем счетное количество раз.
2. Выражение $\nabla_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) \log p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ ограничено на $U_{\boldsymbol{\theta}}$ некоторой случайной величиной с конечным первым моментом.

Тогда задача (1.15) не является обобщающей.

Доказательство. Пусть выполнены условия критерия 6 о переборе структур, и $\mathbf{h}_1, \mathbf{h}_2$ — локальные оптимумы функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ при метапараметрах $\boldsymbol{\lambda}$. По условию критерия соответстcovtие $\boldsymbol{\theta}^*(\mathbf{h}_1)$ и $\boldsymbol{\theta}^*(\mathbf{h}_2)$ должны сохраняться, т.е. для некоторого $\boldsymbol{\lambda}'$ решение нижней задачи оптимизации $\boldsymbol{\theta}^*(\mathbf{h}_1)$ должно совпадать с решением $\boldsymbol{\theta}^*(\mathbf{h}_1)$ при метапараметрах $\boldsymbol{\lambda}$. Тогда

$$\nabla_{\boldsymbol{\theta}} \mathsf{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_1)} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})) =$$

$$= \nabla_{\boldsymbol{\theta}} \mathsf{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_1)} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')).$$

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}')),$$

Из второго условия теоремы следует, что по теореме Лебега о мажорируемой сходимости осуществим переход дифференцирования под знак интеграла:

$$\int_{\boldsymbol{\Gamma}\in\mathbb{\Gamma}} \nabla_{\boldsymbol{\theta}_\Gamma} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)(\log\, p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}) - \log\, p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}'))d\boldsymbol{\Gamma} = 0.$$

Т.к. выражение $\nabla_{\boldsymbol{\theta}_\Gamma} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)$ принимает нулевое значение в счетном количестве точек, то выражение $\log\, p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}) - \log\, p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}')$ равно нулю почти всюду, что означает что метапараметр температуры λ_{temp} равен при разных значениях метапараметров:

$$\lambda_{\mathrm{temp}} = \lambda'_{\mathrm{temp}}, \quad \lambda_{\mathrm{temp}} \in \boldsymbol{\lambda}, \lambda'_{\mathrm{temp}} \in \boldsymbol{\lambda}'.$$

Таким образом, метапараметры $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$ отличаются лишь на метапараметры λ_1, λ_2 регуляризации ковариационной матрицы $\mathbf{A}^{-1}$. Возьмем в качестве векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2$ гиперпараметры, отличающиеся только параметрами распределения структуры:

$$\mathbf{h}_1 = [\mathbf{s}_1, \mathrm{diag}(\mathbf{A}_1)], \mathbf{h}_2 = [\mathbf{s}_2, \mathrm{diag}(\mathbf{A}_2)], \quad \mathbf{s}_1 \neq \mathbf{s}_2, \mathbf{A}_1 = \mathbf{A}_2.$$

Метапараметры λ_1, λ_2 не влияют на значение функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ при гиперпараметрах, отличающихся только параметрами распределения структуры, поэтому значение функции Q для них будет неизменно при любых значениях λ_1, λ_2. Приходим к противоречию: значение $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ не меняется при изменении метапараметров $\boldsymbol{\lambda}$.

□

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{1.17}$$

$$= \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) -$$

$$-\lambda^{\mathrm{Q}}_{\mathrm{prior}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- \sum_{p', \lambda \in \mathbf{P}, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}} \lambda D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*||p'\big) + p(\mathbf{h}|\boldsymbol{\lambda}),$$

$$q^* = \arg\max_q L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f}) \tag{L^*}$$

$$-\lambda^{\mathrm{prior}}_{\mathrm{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w}, \boldsymbol{\Gamma})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f})\big),$$

где $\mathbf{P}$ — непустое множество распределений на структуре $\boldsymbol{\Gamma}$, $\lambda^{\mathrm{prior}}_{\mathrm{Q}}, \lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \boldsymbol{\lambda}^{\mathrm{struct}}_{\mathrm{Q}}$ — некоторые числа. Множество распределений $\mathbf{P}$ отвечает за перебор структур $\boldsymbol{\Gamma}$ в процессе оптимизации модели. Подробное объяснение данного множества дано ниже.

Список основных обозначений

$\mathbf{x}_i \in \mathbf{X}$ — вектор признакового описания i-го объекта
$y_i \in \mathbf{y}$ — метка i-го объекта
$\mathfrak{D}$ — выборка
$\mathbf{X} \subset \mathbb{X}$ — матрица, содержащая признаковое описание объектов выборки
$\mathbf{y} \subset \mathbb{Y}$ — вектор меток объектов выборки
m — количество объектов в выборке
n — количество признаков в признаковом описании объекта
$\mathbb{X} =?\mathbb{R}^m$ — признаковое пространство объектов
$\mathbb{Y}$ — множество меток объектов
R — множество классов в задаче классификации
(V, E) — граф со множеством вершин V и множеством ребер E
$\mathbf{g}^{j,k}$ — вектор базовых функций для ребра (j, k)
$K^{j,k}$ — мощность вектора базовых функций для ребра (j, k)
$\mathbf{agg}_v$ — функция аггрегации для вершины v
$\boldsymbol{\gamma}^{j,k}$ — структурный параметр для ребра (j, k)
$\mathfrak{F}$ — параметрическое семейство моделей
U — область определения оптимизационной задачи
$\mathbf{w} \in \mathbb{W}$ — параметры модели
$\mathbb{W}$ — пространство параметров модели
$U_{\mathbf{w}} \subset \mathbb{W}$ — область определения параметров модели
$\boldsymbol{\Gamma} \in \mathbb{\Gamma}$ — структура модели
$\mathbb{\Gamma}$ — множество значений структуры модели
$U_{\boldsymbol{\Gamma}} \subset \mathbb{\Gamma}$ — область определения параметров модели
$\mathbf{h} \in \mathbb{H}$ — гиперпараметры модели
$\mathbb{H}$ — пространство гиперпараметров модели
$U_{\mathbf{h}} \subset \mathbb{H}$ — область определения гиперпараметров
$\boldsymbol{\theta} \in \mathbb{\Theta}$ — параметры вариационного распределения
$\mathbb{\Theta}$ — пространство параметров вариационного распределения
$U_{\boldsymbol{\theta}} \subset \mathbb{\Theta}$ — область определения вариационных параметров модели
$\boldsymbol{\theta}_{\mathbf{w}} \in \mathbb{\Theta}_{\mathbf{w}}$ — параметры вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$\mathbb{\Theta}_{\mathbf{w}}$ — пространство параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \mathbb{\Theta}_{\mathbf{w}}$ — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$\boldsymbol{\theta}_{\boldsymbol{\Gamma}} \in \mathbb{\Theta}_{\boldsymbol{\Gamma}}$ — параметры вариационного распределения, аппроксимирующего апостериорное распределение структуры модели
$\mathbb{\Theta}_{\boldsymbol{\Gamma}}$ —пространство параметров вариационного распределения, аппроксимирующего апостериорное распределение структуры модели
$U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \subset \mathbb{\Theta}_{\boldsymbol{\Gamma}}$ — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение структуры модели

$\boldsymbol{\lambda} \in \lambda$ — вектор метапараметров
λ — пространство метапараметров
$U_{\boldsymbol{\lambda}} \subset \lambda$ — область определения метапараметров
$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})$ — правдоподобие выборки
$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение параметров и структуры модели
$p(\mathbf{h}|\boldsymbol{\lambda})$ — распределение гиперпараметров модели
$p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение структуры модели
$p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение параметров модели
$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение параметров и структуры модели
$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение структуры модели
$p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение структуры модели
$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda})$ — апостериорное распределение гиперпараметров
$p(y, \mathbf{w}, \mathbf{\Gamma}|\mathbf{x}, \mathbf{h})$ — вероятностная модель глубокого обучения
$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — обоснованность модели
$q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ — вариационное распределение параметров и структуры модели
$q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ — вариационное распределение структуры модели
$q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ — вариационное распределение параметров модели
$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — функция потерь
$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ — валидационная функция
$T(\boldsymbol{\theta}|L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}))$ — оператор оптимизации
$\mathfrak{Q}$ — семейство вариационные распределений
S — энтропия распределения
M — множество моделей без общей параметризации
$D_{\mathrm{KL}}\big(p_1||p_2\big)$ — дивергенция Кульбака-Лейблера между распределениями p_1 и p_2
$\mathbf{A}^{-1}$ — матрица коварияций параметров модели
$\mathbf{s}$ — конкатенация параметров концентрации на структуре модели