

# 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:

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

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  $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})$  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  $\mathbf{\Gamma}$  предлагается использовать произведение распределений Gumbel-Softmax ( $\mathcal{GS}$ ) [?]:

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

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:

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

где  $\mathbf{s}^{j,k} \in (0, \infty)^{K^{j,k}}$  — 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  $\gamma^{j,k}$  представима следующим образом:

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

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

2. Свойство округления:  $p(\gamma_{l_1} > \gamma_{l_2}, l_1 \neq l_2 | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \frac{s_{l_1}^{j,k}}{\sum_{l'} s_{l'}^{j,k}}$ .
3. При устремлении температуры к нулю реализация  $\hat{\gamma}^{j,k}$  случайной величины концентрируется на вершинах симплекса:

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

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

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

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

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

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

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:

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

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

вектора  $\gamma^{j,k}$ :

$$l' = \arg \max_{l \in \{1, \dots, K^{j,k}\}} \gamma_l^{j,k}.$$

Для  $l = l'$  предел выражения (1.3) при  $\lambda_{\text{temp}} \rightarrow \infty$  стремится к бесконечности. Для  $l \neq l'$  предел выражения (1.3) при  $\lambda_{\text{temp}} \rightarrow \infty$  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  $\gamma^{j,k}$ . 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. □

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

**Определение 1.** Случайную величину  $\psi$  с распределением  $q$  с параметрами  $\theta_\psi$  назовем репараметризованной через случайную величину  $\varepsilon$ , чье распределение не зависит от параметров  $\theta_\psi$ , если:

$$\psi = g(\varepsilon, \theta_\psi)$$

где  $g$  — некоторая непрерывная функция.

Идею репараметризации поясним на следующем примере.

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

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

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

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

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

$$\mathbf{w} = \mathbf{g}(\varepsilon, \theta_{\mathbf{w}}).$$

Тогда справедливо следующее выражение:

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



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

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

Таким образом, распределение, позволяющее произвести репараметризацию, является более удобным для вычисления интегральных оценок вида  $\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$ , 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  $\boldsymbol{\Gamma} \in \mathbb{I}$  равнозначны, выступает равномерное распределение. Выбор в качестве распределения структуры 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  $\lambda_1, \lambda_2$  получается более сильная или более слабая регуляризация [?]. Пример распределений  $\text{inv-gamma}(\lambda_1, \lambda_2)$  для разных значений метапараметров  $\lambda_1, \lambda_2$  изображен на Рис. 1.2. Оптимизации без регуляризации соответствует случай предельного распределения  $\lim_{\lambda_1, \lambda_2 \rightarrow 0} \text{inv-gamma}(\lambda_1, \lambda_2)$ .

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

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

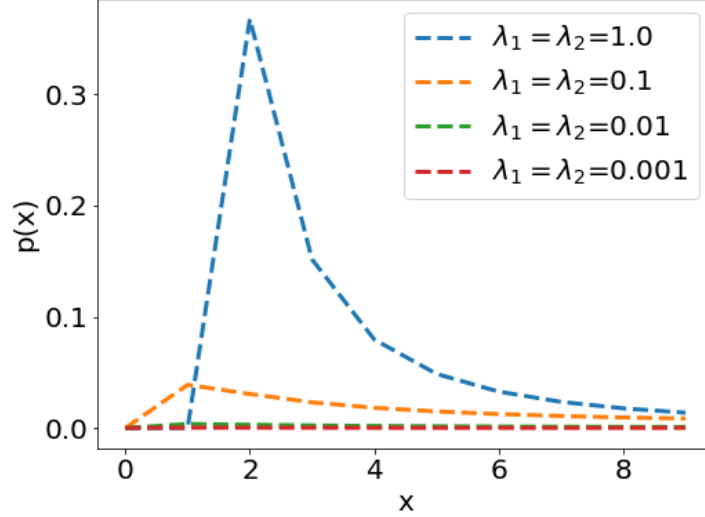


Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

2. Структура модели  $\Gamma$ , содержащая все структурные параметры  $\{\gamma^{j,k}, (j,k) \in E\}$ , 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.

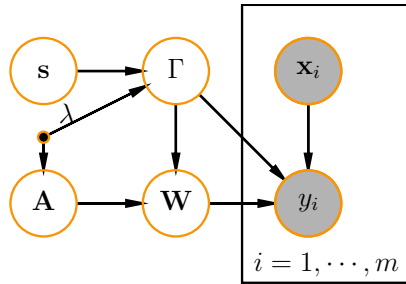


Рис. 1.3. TODO: сделать лямбду красивой (снова). График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

## 1.2. Вариационная оценка обоснованности вероятностной модели

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

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

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

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

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

Для вычисления обоснованности модели

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

из (1.4) предлагается использовать нижнюю вариационную оценку обоснованности.

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

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

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

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

Тогда справедлива следующая оценка:

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

$$\begin{aligned} & \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q_{\Gamma}(\Gamma | \boldsymbol{\theta}_{\Gamma}) || p(\Gamma | \mathbf{h}, \boldsymbol{\lambda})) - \\ & - D_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w} | \Gamma, \boldsymbol{\theta}_{\mathbf{w}}) || p(\mathbf{w} | \Gamma, \mathbf{h}, \boldsymbol{\lambda})), \end{aligned}$$

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

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

*Доказательство.* Перепишем обоснованность:

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) &= \log \int \int_{\Gamma, \mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) p(\Gamma|\mathbf{h}, \boldsymbol{\lambda}) d\Gamma d\mathbf{w} = \\
&= \log \int \int_{\Gamma, \mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda}) \frac{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})}{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} d\Gamma d\mathbf{w} = \\
&= \log \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})}.
\end{aligned}$$

Используя неравенство Йенсена получим

$$\begin{aligned}
\log \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} &\geq \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} = \\
&= -\mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}) || p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})).
\end{aligned}$$

Декомпозируем распределение  $q$  по свойству условной дивергенции:

$$\begin{aligned}
&D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}) || p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) = \\
&= D_{\text{KL}}(q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) || p(\Gamma|\mathbf{h}, \boldsymbol{\lambda})) + \mathbb{E}_{\Gamma \sim q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})} \mathbb{E}_{\mathbf{w} \sim q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \log \left( \frac{q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \right). \quad (1.6)
\end{aligned}$$

□

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

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

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

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

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

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:

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

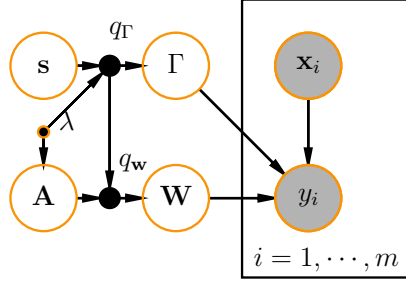


Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

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 [?]:

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

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

Обобщим понятие относительной вариационной плотности на случай произвольных непрерывных распределений.

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

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

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

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



Сформулируем и докажем теорему о связи относительной плотности и параметрической сложности модели. Перед доказательством теоремы докажем две вспомогательные леммы.

**Лемма 1.** Пусть

1. Вариационное распределение  $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  и априорное распределение  $p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})$  являются абсолютно непрерывными и унимодальными на  $U_{\mathbf{h}}, U_{\boldsymbol{\theta}}$ . Их мода и матожидание совпадают:

$$\text{mode } p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w};$$

$$\text{mode } q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) = \mathbb{E}_{q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}.$$

2. Заданы компактные множества  $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \Theta_{\mathbf{w}}, U_{\boldsymbol{\theta}_{\Gamma}} \subset \Theta_{\Gamma}$ .
3. Мода априорного распределения  $p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})$  не зависит от гиперпараметров  $\mathbf{h}$  на  $U_{\mathbf{h}}$  и структуры  $\Gamma$  на  $U_{\boldsymbol{\theta}_{\Gamma}}$ :

$$\text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{m}$$

для любых  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \Gamma_1, \Gamma_2 \in U_{\boldsymbol{\theta}_{\Gamma}}$ .

4. Параметры модели  $\mathbf{w}$  имеют конечные вторые моменты по распределениям:

$$\int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) d\Gamma, \quad \int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) d\Gamma.$$

5. Вариационное распределение  $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  является липшецевым по  $\mathbf{w}$ .
6. Значение  $q_{\mathbf{w}}(\mathbf{m}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  не равно нулю при  $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$ .

Тогда

$$\begin{aligned} & \left| \mathbb{E}_{q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})} \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) - 1 \right| \leq \\ & \leq \frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}} \in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{m}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \iint_{\Gamma, \mathbf{w}} |\mathbf{w}| |q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})| q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) d\mathbf{w} d\Gamma, \end{aligned}$$

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

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

$$\left| \mathbb{E}_{q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})} \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) - 1 \right| =$$

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

представляя единицу как  $\frac{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}$

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

заменяя модуль на матожидание (по условию теоремы)

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

занося модуль под знак интеграла

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

используя липшецевость функции  $q_{\mathbf{w}}(\mathbf{w} | \Gamma, \boldsymbol{\theta}_{\mathbf{w}})$

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

расписывая матожидание через интеграл

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

□

**Лемма 2.** Пусть

1. Вариационное распределение  $q_{\mathbf{w}}(\mathbf{w} | \Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  и априорное распределение  $p(\mathbf{w} | \Gamma, \mathbf{h}, \boldsymbol{\lambda})$  являются абсолютно непрерывными.
2. Решение задачи

$$\mathbf{h}^* = \arg \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) \quad (1.7)$$

единственно для любого  $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$ .

3. Задана бесконечная последовательность векторов вариационных параметров  $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \dots, \boldsymbol{\theta}[i], \dots \in U_{\boldsymbol{\theta}}$ , такая что  $\lim_{i \rightarrow \infty} C_p(\boldsymbol{\theta}[i] | U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$ . Тогда следующее выражение стремится к нулю:

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

где  $\boldsymbol{\theta}[i] = [\boldsymbol{\theta}_{\mathbf{w}}[i], \boldsymbol{\theta}_{\Gamma}[i]]$ ,  $\mathbf{h}[i]$  — решение задачи (1.7) для  $\boldsymbol{\theta}[i]$ .

*Доказательство.* Воспользуемся неравенством Пинскера:

$$\|F_q((\boldsymbol{\theta}_{\mathbf{w}})_i) - F_p(\mathbf{h}_i)\|_{\text{TV}} \leq \sqrt{2 \log q_{\mathbf{w}}(\mathbf{w} | \Gamma, (\boldsymbol{\theta}_{\mathbf{w}})_i) - \log p(\mathbf{w} | \Gamma, \mathbf{h}_i, \boldsymbol{\lambda})} d\mathbf{w} = 0,$$

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

По условию дивергенция (1.6) стремится к нулю при  $i \rightarrow \infty$ . Она декомпозируется на два неотрицательных слагаемых, поэтому обе они стремятся к нулю. Рассмотрим второе слагаемое:

$$0 = \lim_{i \rightarrow \infty} \mathbb{E}_{\Gamma \sim q_{\Gamma}(\Gamma | \theta_{\Gamma}[i])} \mathbb{E}_{\mathbf{w} \sim q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i])} \log \left( \frac{q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i])}{p(\mathbf{w} | \Gamma, \mathbf{h}[i], \boldsymbol{\lambda})} \right) =$$

расписывая матожидание как интеграл

$$\lim_{i \rightarrow \infty} \left| \int_{\Gamma} \int_{\mathbf{w}} \log \left( \frac{q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i])}{p(\mathbf{w} | \Gamma, \mathbf{h}[i], \boldsymbol{\lambda})} \right) q_{\Gamma}(\Gamma | \theta_{\Gamma}[i]) q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i]) d\mathbf{w} d\Gamma \right| \geq$$

по неравенству Пинскера

$$\geq \lim_{i \rightarrow \infty} \int_{\Gamma} \|F_q(\theta_{\mathbf{w}}[i]) - F_p(\mathbf{h}_i)\|_{\text{TV}} q_{\Gamma}(\Gamma | \theta_{\Gamma}[i]) d\Gamma \geq 0.$$

Отсюда

$$\lim_{i \rightarrow \infty} \int_{\Gamma} \|F_q(\theta_{\mathbf{w}}[i]) - F_p(\mathbf{h}_i)\|_{\text{TV}} q_{\Gamma}(\Gamma | \theta_{\Gamma}[i]) d\Gamma = 0.$$

По теореме Шеффе данное выражение можно переписать как:

$$\lim_{i \rightarrow \infty} \frac{1}{2} \iint_{\Gamma, \mathbf{w}} |p(\mathbf{w} | \Gamma, \mathbf{h}[i], \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i])| q_{\Gamma}(\Gamma | \theta_{\Gamma}[i]) d\Gamma d\mathbf{w} = 0, \quad (1.8)$$

что и требовалось доказать.  $\square$

**Теорема 2.** Пусть выполнены условия Леммы 1 и Леммы 2. Тогда следующее выражение стремится к единице:

$$\lim_{i \rightarrow \infty} \mathbb{E}_{q_{\Gamma}(\theta_{\Gamma}[i] | \theta_{\Gamma})} \rho(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}[i], \mathbf{h}[i], \boldsymbol{\lambda}) = 1.$$

*Доказательство.* По Лемме 2

$$\begin{aligned} & \mathbb{E}_{q_{\Gamma}(\theta_{\Gamma} | \theta_{\Gamma})} \rho(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) \leq \\ & \leq \frac{C_l}{\min_{\theta_{\mathbf{w}} \in U_{\theta}} q_{\mathbf{w}}(\mathbf{m} | \Gamma, \theta_{\mathbf{w}})} \iint_{\Gamma, \mathbf{w}} |\mathbf{w}| |q_{\mathbf{w}}(\mathbf{w} | \Gamma, \theta_{\mathbf{w}}) - p(\mathbf{w} | \Gamma, \mathbf{h}, \boldsymbol{\lambda})| q_{\Gamma}(\Gamma | \theta_{\Gamma}) d\mathbf{w} d\Gamma. \end{aligned}$$

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

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

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:

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

по неравенству треугольника и используя выражение  $\mathbf{w} = \mathbf{w} + \boldsymbol{\nu}(t) - \boldsymbol{\nu}(t)$

$$\begin{aligned} &\leq \iint_{\Gamma, \mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| |p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})| q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) d\mathbf{w} d\Gamma + \\ &+ \iint_{\Gamma, \mathbf{w}} |\boldsymbol{\nu}(t)| |q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})| q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) d\mathbf{w} d\Gamma. \end{aligned} \quad (1.9)$$

Рассмотрим первое слагаемое суммы (1.9). Т.к. вторые моменты  $E_{q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})} E_{q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}^2$ ,  $E_{q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})} E_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w}^2$  конечны, то случайная величина  $\mathbf{w}$  равномерно интегрируема как при маргинальном распределении  $\int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) d\Gamma$ , так и при маргинальном распределении  $\int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) d\Gamma$ . По определению равномерной интегрируемости для  $\mathbf{w}$  для любого числа  $\varepsilon$  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:

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

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

Тогда

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

так как модуль разностей меньше или равен сумме модулей

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

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

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

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

Рассмотрим второе слагаемое.

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

по ограниченности функции  $\boldsymbol{\nu}(t)$

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

Переходя к пределу в (1.9) получим:

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

добавим предел по  $t$ , от которого не зависит данное выражение

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

из выше написанных неравенств

$$\begin{aligned} & \lim_{t \rightarrow \infty} \lim_{i \rightarrow \infty} \iint_{\Gamma, \mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| |p(\mathbf{w}|\Gamma, \mathbf{h}[i], \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}[i])| d\mathbf{w} d\Gamma + \\ & + \iint_{\Gamma, \mathbf{w}} |\boldsymbol{\nu}(t)| |q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}[i]) - p(\mathbf{w}|\Gamma, \mathbf{h}[i], \boldsymbol{\lambda})| q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}[i]) d\mathbf{w} d\Gamma \leq \\ & \lim_{t \rightarrow \infty} 2\varepsilon(t) + \lim_{t \rightarrow \infty} \lim_{i \rightarrow \infty} t \iint_{\Gamma, \mathbf{w}} |q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}[i]) - p(\mathbf{w}|\Gamma, \mathbf{h}_i, \boldsymbol{\lambda})| q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}[i]) d\mathbf{w} d\Gamma = 0. \end{aligned}$$

Последнее равенство следует из Леммы 2. Таким образом выражение  $\left| \int_{\Gamma} \frac{q_{\mathbf{w}}(\text{mode}_{q_{\mathbf{w}}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode}_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})}(\mathbf{w})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) d\Gamma \right|$  стремится к единице, что и требовалось доказать.  $\square$

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

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

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

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

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, использующих вариационное распределение  $q(\mathbf{w}, \Gamma|\boldsymbol{\theta})$  для аппроксимации неизвестного апостериорного распределения параметров  $p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})$ .

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}). \quad (1.10)$$

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

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

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

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \quad (1.11)$$

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

Т.к. в рамках данной задачи (1.11) 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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log p(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) = p'|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \quad (1.12)$$

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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \quad (1.13)$$

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

где

$$\mathbf{h} = \arg \min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})), \quad (1.14)$$

$\lambda_{\text{prune}}$  — метапараметр алгоритма,  $U_{\mathbf{h}} \subset \mathbb{H}$  — область определения задачи по гиперпараметрам. Предложенное обобщение (1.13) применимо только в случае, если выражение (1.14) определено однозначно, т.е. существует единственный вектор гиперпараметров  $\mathbf{h} \in U_{\mathbf{h}}$ , доставляющий минимум дивергенции  $D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))$ .

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:

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

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

метапараметр  $\lambda_{\text{prune}}$  определен аналогично (1.14).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \quad (1.16)$$

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

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

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

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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \mathbf{y}_{\text{train}} \mathbf{X}_{\text{train}} + p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \rightarrow \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}}, \quad (1.17)$$

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

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 значениях метапараметров, будет называть *обобщающей*.

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

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

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

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

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  $K_1 > 0$ ,

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

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

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

выполняется неравенство

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



4. Критерий минимизации параметрической сложности: существует  $\lambda \in U_\lambda$  и  $K_2 > 0$ ,

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

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

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

параметрическая сложность первой модели меньше, чем второй:

$$C_p(\boldsymbol{\theta}^*(\mathbf{h}_1) | U_{\mathbf{h}}, \lambda) < C_p(\boldsymbol{\theta}^*(\mathbf{h}_2) | U_{\mathbf{h}}, \lambda).$$

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

$$Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}), \lambda) \propto \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta}'(\mathbf{h}))} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}'(\mathbf{h})) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)) + \log p(\mathbf{h} | \lambda)$$

для всех  $\mathbf{h} \in U_{\mathbf{h}}$ , где в качестве гиперпараметров  $\mathbf{h}$  рассматриваются все гиперпараметры модели:  $\mathbf{h} = [\mathbf{A}, \mathbf{s}]$ , где

$$\boldsymbol{\theta}'(\mathbf{h}) = \arg \max_{\boldsymbol{\theta} \in U_{\mathbf{h}}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)).$$

6. Критерий перебора оптимальных структур: существует константа  $K_3 > 0$ , такая что существует хотя бы одна пара гиперпараметров  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$  :

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

и набор метапараметров  $\lambda$ , такие что для локальных оптимумов  $\mathbf{h}_1, \mathbf{h}_2$  задачи оптимизации  $Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$ , полученных при метапараметрах  $\lambda$  и удовлетворяющих неравенствам

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

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

существует значение метапараметров  $\lambda' \neq \lambda$ , такие что

- (а) соответствие между вариационными параметрами  $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$  сохраняется при  $\lambda'$ ,
- (б) выполняется неравенство  $Q(\mathbf{h}_1 | \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) < Q(\mathbf{h}_2 | \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$  при  $\lambda'$ .

7. Критерий непрерывности: функции  $L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$  и  $Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$  непрерывны по метапараметрам  $\lambda \in U_\lambda$ .

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  $p(\Gamma|\mathbf{h}, \boldsymbol{\lambda})$  значимо высока. При этом соответствующие вариационные распределения  $q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma)$  могут оказаться достаточно близки. Возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения  $q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma)$  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 (1.10), (1.11), (1.12), (1.13), (1.15), (1.17) не являются обобщающими.

*Доказательство.* Задачи (1.10), (1.11), (1.12), (1.13), (1.15) не имеют гиперпараметров  $\mathbf{h}$ , подлежащих оптимизации, поэтому не могут оптимизировать вариационную оценку.

При использовании валидации на отложенной выборке (1.17) 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  $q_\Gamma$  — абсолютно непрерывное распределение с дифференцируемой плотностью, такой что:

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

Тогда задача (1.16) 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 соответствие  $\boldsymbol{\theta}^*(\mathbf{h}_1)$  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

$$\begin{aligned} & \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - \nabla_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma|\mathbf{h}_1, \boldsymbol{\lambda})) = \\ & = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - \nabla_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma|\mathbf{h}_1, \boldsymbol{\lambda}')). \end{aligned}$$

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\boldsymbol{\theta}} D_{\text{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\theta}} D_{\text{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}')),$$

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

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

Т.к. выражение  $\nabla_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)$  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  $\lambda_{\text{temp}}$  равен при разных значениях метапараметров:

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

Таким образом, метапараметры  $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$  отличаются лишь на метапараметры  $\lambda_1, \lambda_2$  регуляризации ковариационной матрицы  $\mathbf{A}^{-1}$ . Возьмем в качестве векторов гиперпараметров  $\mathbf{h}_1, \mathbf{h}_2$  гиперпараметры, отличающиеся только параметрами распределения структуры:

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

Метапараметры  $\lambda_1, \lambda_2$  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  $\lambda_1, \lambda_2$ . 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:

$$\begin{aligned} \mathbf{h}^* &= \arg \max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \\ &= \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\ &\quad - \lambda_{\text{prior}}^Q D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) - \\ &\quad - \sum_{p', \boldsymbol{\lambda} \in \mathbf{P}, \boldsymbol{\lambda}_{\text{struct}}^Q} \lambda D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*)||p') + \log p(\mathbf{h}|\boldsymbol{\lambda}), \\ \boldsymbol{\theta}^* &= \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \\ &= \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \lambda_{\text{prior}}^L D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})), \end{aligned} \tag{1.18}$$

где  $\mathbf{P}$  — непустое множество распределений на структуре  $\boldsymbol{\Gamma}$ ,  $\lambda_{\text{prior}}^Q, \lambda_{\text{prior}}^L, \lambda_{\text{struct}}^Q$  — некоторые числа. Множество распределений  $\mathbf{P}$  отвечает за перебор структур  $\boldsymbol{\Gamma}$  в процессе оптимизации модели. Более подробное объяснение данного множества дано ниже.

**Теорема 5.** Пусть:

- 1) задано непустое множество непрерывных по параметрам распределений на структуре  $\mathbf{P}$ , не принимающих нулевое значение, где хотя бы одно распределение  $p_1 \in \mathbf{H}$  является Gumbel-Softmax распределением и для любого значения  $\mathbf{s} \in U_{\mathbf{h}}$ ,  $\lambda_{\text{temp}} \in U_{\lambda}$  существует значение параметров распределения  $p_1$ , такое что  $p_1 = p(\mathbf{\Gamma}|\mathbf{h}, \lambda)$ .
- 2) Вариационное распределение  $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$  является абсолютно непрерывным, плотность которого непрерывна по метапараметрам  $\lambda$  и не принимает нулевое значение.
- 3) Задан компакт  $U = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\mathbf{r}}} \times U_{\mathbf{h}} \times U_{\lambda}$ , где параметры распределений  $\mathbf{P} \in U_{\lambda}$ .
- 4) Область определения каждого параметра  $w \in \mathbf{w}$ , гиперпараметра  $h \in \mathbf{h}$  и метапараметра  $\lambda \in \lambda$  не является пустым и не является точкой.
- 5) Для каждого значения гиперпараметров  $\mathbf{h} \in U_{\mathbf{h}}$  оптимальное решение нижней задачи оптимизации  $\boldsymbol{\theta}^*$  определено однозначно на  $U_{\boldsymbol{\theta}} = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\mathbf{r}}}$  при любых значениях метапараметров  $\lambda \in U_{\lambda}$ .
- 6) Область значений метапараметров  $\lambda_{\text{likelihood}}^{\text{Q}}, \lambda_{\text{prior}}^{\text{Q}}, \lambda_{\text{prior}}^{\text{L}}, \lambda_{\text{struct}}^{\text{Q}}$  включает отрезок от нуля до единицы.
- 7) Существует значение метапараметров  $\lambda_1 > 0, \lambda_2 > 0, \lambda_{\text{likelihood}}^{\text{Q}} > 0 \in U_{\lambda}$ , такое что

$$\max_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h}|\lambda) - \min_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h}|\lambda) < \max_{\mathbf{h} \in U_{\mathbf{h}}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) - \min_{\mathbf{h} \in U_{\mathbf{h}}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$$

при  $\lambda_{\text{struct}}^{\text{Q}} = 0, \lambda_{\text{prior}}^{\text{Q}} = 0$ .

- 8) Существует значение метапараметров  $\lambda_{\text{prior}}^{\text{Q}} > 0, \lambda_1 > 0, \lambda_2 > 0, \lambda_{\text{temp}} > 0 \in U_{\lambda}$ , такое что

$$\begin{aligned} & \max_{\mathbf{h} \in U_{\mathbf{h}}} \frac{1}{\lambda_{\text{prior}}^{\text{Q}}} \log p(\mathbf{h}|\lambda) - \min_{\mathbf{h} \in U_{\mathbf{h}}} \frac{1}{\lambda_{\text{prior}}^{\text{Q}}} \log p(\mathbf{h}|\lambda) + \\ & + \max_{\mathbf{h} \in U_{\mathbf{h}}} \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda)) - \\ & \min_{\mathbf{h} \in U_{\mathbf{h}}, \boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda)) + \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \\ & - \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) < \\ & < \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda)) - \\ & - \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda)). \end{aligned}$$

- 9) Существуют значения метапараметров  $\lambda_{\text{prior}}^Q > 0, \lambda_{\text{likelihood}}^Q > 0, \lambda_1 > 0, \lambda_2 > 0, \lambda_{\text{temp}} > 0 \in U_{\lambda}$ , такие что существуют гиперпараметры  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$ :

$$\begin{aligned} & D_{\text{KL}}(p(\mathbf{w}, \Gamma | \mathbf{h}_1, \lambda) || p(\mathbf{w}, \Gamma | \mathbf{h}_2, \lambda)) < \\ & < \frac{\max_{\mathbf{h}} Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta, \lambda) - \min_{\mathbf{h}} Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta, \lambda)}{\max_{\lambda_{\text{struct}}^Q} \lambda_{\text{struct}}^Q}, \\ & D_{\text{KL}}(p(\mathbf{w}, \Gamma | \mathbf{h}_2, \lambda) || p(\mathbf{w}, \Gamma | \mathbf{h}_1, \lambda)) < \\ & < \frac{\max_{\mathbf{h}} Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta, \lambda) - \min_{\mathbf{h}} Q(\mathbf{h} | \mathbf{y}, \mathbf{X}, \theta, \lambda)}{\max_{\lambda_{\text{struct}}^Q} \lambda_{\text{struct}}^Q}. \end{aligned}$$

при  $\lambda_{\text{struct}}^Q = \mathbf{0}$ . TODO - как написать?

Тогда задача (1.18) является обобщающей на  $U$ .

*Доказательство.* Для доказательства теоремы требуется доказать критерии 1-7 из определения обобщающей задачи. Выполнение критериев 1 и 2 следует из условий задачи.

Докажем критерий 3. Пусть  $\lambda_{\text{prior}}^Q = 0, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Пусть  $\lambda_1, \lambda_2, \lambda_{\text{likelihood}}^Q$  удовлетворяют седьмому условию теоремы. Возьмем в качестве  $K_1$  следующее выражение:

$$K_1 = \max_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h} | \lambda) - \min_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h} | \lambda).$$

Пусть  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, Q(\mathbf{h}_1 | \mathbf{y}, \mathbf{X}, \theta, \lambda) - Q(\mathbf{h}_2 | \mathbf{y}, \mathbf{X}, \theta, \lambda) > K_1$ . Тогда

$$\begin{aligned} Q(\mathbf{h}_1 | \mathbf{y}, \mathbf{X}, \theta, \lambda) - Q(\mathbf{h}_2 | \mathbf{y}, \mathbf{X}, \theta, \lambda) &= \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta^*(\mathbf{h}_1))} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \\ &- \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta^*(\mathbf{h}_2))} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) + \log p(\mathbf{h}_1 | \lambda) - \log p(\mathbf{h}_2 | \lambda) > K_1. \end{aligned}$$

Отсюда следует выполнение критерия 3:

$$\lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta_1)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta_2)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) > 0.$$

Докажем критерий 4. Пусть  $\lambda$  удовлетворяют восьмому условию теоремы и  $\lambda_{\text{likelihood}}^Q = 0, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Пусть

$$\begin{aligned} K_2 &= \max_{\mathbf{h} \in U_{\mathbf{h}}} \frac{1}{\lambda_{\text{prior}}^Q} \log p(\mathbf{h} | \lambda) - \frac{1}{\lambda_{\text{prior}}^Q} \min_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h} | \lambda) + \\ &+ \max_{\mathbf{h} \in U_{\mathbf{h}}} \min_{\theta \in U_{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)) - \\ &\min_{\mathbf{h} \in U_{\mathbf{h}}, \theta \in U_{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)) + \max_{\theta \in U_{\theta}} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \\ &\min_{\mathbf{h} \in U_{\mathbf{h}}} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma). \end{aligned}$$

Пусть  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$ ,  $Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) > K_2$ . Рассмотрим разность параметрических сложностей двух векторов:

$$C_p(\boldsymbol{\theta}_2) - C_p(\boldsymbol{\theta}_1) = \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - \\ - \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) \geq$$

оценим снизу, а также добавим и вычтем  $D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda}))$

$$\geq \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1)||p(\mathbf{w}, \Gamma|\mathbf{h}_1, \boldsymbol{\lambda})) +$$

$$+ D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda})) =$$

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

$$= Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \frac{1}{\lambda_{\text{prior}}^Q} \log p(\mathbf{h}_1|\boldsymbol{\lambda}) + \frac{1}{\lambda_{\text{prior}}^Q} \log p(\mathbf{h}_2|\boldsymbol{\lambda}) +$$

$$+ \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda})) >$$

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

$$> K_2 - \frac{1}{\lambda_{\text{prior}}^Q} \log p(\mathbf{h}_1|\boldsymbol{\lambda}) + \frac{1}{\lambda_{\text{prior}}^Q} \log p(\mathbf{h}_2|\boldsymbol{\lambda}) + \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_1)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) \\ - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda})).$$

Рассмотрим разность:

$$\min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}_2, \boldsymbol{\lambda})) =$$

т.к.  $\boldsymbol{\theta}_2$  — решение нижней задачи оптимизации:

$$\min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) +$$

$$\max_{\boldsymbol{\theta}} \left( \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{h}_2, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) \right) \geq$$

получим оценку снизу:

$$\geq \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_2)||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) - \max_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) +$$

$$\max_{\boldsymbol{\theta}} \left( \min_{\boldsymbol{\theta}'} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta}')} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{h}_2, \Gamma|\mathbf{h}, \boldsymbol{\lambda})) \right) \geq$$

оценим первое слагаемое

$$\begin{aligned} &\geq \min_{\boldsymbol{\theta}, \mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) - \max_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) + \\ &\min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{h}_2, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) \geq \end{aligned}$$

оценим последнее слагаемое

$$\begin{aligned} &\geq \min_{\boldsymbol{\theta}, \mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) - \max_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \\ &+ \min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{prior}}^{\text{L}}} \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \max_{\mathbf{h}} \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

Складывая полученную оценку с  $K_2 - \log \frac{1}{\lambda_{\text{prior}}^{\text{Q}}} p(\mathbf{h}_2|\boldsymbol{\lambda}) + \log \frac{1}{\lambda_{\text{prior}}^{\text{Q}}} p(\mathbf{h}_2|\boldsymbol{\lambda})$  получаем разность параметрических сложностей больше нуля, что и требовалось доказать.

Докажем критерий 5. Пусть  $\lambda_{\text{prior}}^{\text{Q}} = \lambda_{\text{prior}}^{\text{L}} = \lambda_{\text{likelihood}}^{\text{Q}} = 1$ ,  $\boldsymbol{\lambda}_{\text{struct}}^{\text{Q}} = \mathbf{0}$ . Тогда функции  $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  и  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$  можно записать как:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})),$$

$$\begin{aligned} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) + \\ &+ \log p(\mathbf{h}|\boldsymbol{\lambda}). \end{aligned}$$

Двухуровневая задача оптимизации совпадает с оптимизацией вариационной оценки обоснованности, что и требовалось доказать.

Докажем критерий 6. Пусть задан вектор метапараметров  $\boldsymbol{\lambda}$ , удовлетворяющий девятому условию теоремы и  $\boldsymbol{\lambda}_{\text{struct}}^{\text{Q}} = \mathbf{0}$ . Пусть заданы векторы гиперпараметров  $\mathbf{h}_1, \mathbf{h}_2$ , такие что  $Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) > 0$ . Возьмем в качестве  $K_4$  следующее выражение:

$$K_4 = \frac{\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})}{\max_{\boldsymbol{\lambda}_{\text{struct}}^{\text{Q}}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})}.$$

Пусть вектор метапараметров  $\boldsymbol{\lambda}'$  отличается от  $\boldsymbol{\lambda}$  лишь метапараметром  $\lambda_{\text{struct}}^{\text{Q}}$ . Для обоих векторов метапараметров нижняя задача оптимизации  $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  совпадает, поэтому выполняется первое условие критерия.

По условию теоремы во множество  $\mathbf{P}$  входит хотя бы одно распределение Gumbel-Softmax:

$$p_1 \sim \mathcal{GS}, p \in \mathbf{P}.$$

Положим для  $\lambda'$  метапараметр перед данным распределением  $\lambda_{\text{struct}}^Q \in \lambda_{\text{struct}}^Q$  равным максимальному значению. Положим также значение параметров данного распределения равным параметрам распределения  $p(\mathbf{h}_1, \Gamma|\mathbf{h}, \lambda)$  :

$$p_1 = p(\mathbf{h}_1, \Gamma|\mathbf{h}, \lambda).$$

Для остальных распределений  $p' \in \mathbf{P}$  положим коэффициент  $\lambda_{\text{struct}}^Q \in \lambda_{\text{struct}}^Q$  равным нулю. Тогда справедливо следующее неравенство:

$$\begin{aligned} & Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda') - Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda') = \\ & = Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) - Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) + \max_{\lambda_{\text{struct}}^Q} \lambda_{\text{struct}}^Q D_{\text{KL}}(p(\mathbf{h}_2, \Gamma|\mathbf{h}, \lambda) || p(\mathbf{h}_1, \Gamma|\mathbf{h}, \lambda)) = \\ & = Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) - Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda) + \max_{\lambda_{\text{struct}}^Q} \lambda_{\text{struct}}^Q K_4 > 0. \end{aligned}$$

что и требовалось доказать.

Докажем критерий 7. Достаточным условием непрерывности функций  $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$ ,  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$  является непрерывность входящих в нее слагаемых.

Слагаемое  $E_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma)$  не зависит от метапараметров  $\lambda$ . Слагаемое  $\log p(\mathbf{h}|\lambda)$  непрерывно по метапараметрам по свойству обратного гамма-распределения.

Достаточным условием непрерывности функций вида  $D_{\text{KL}}(p_1 || p_2)$  является непрерывность по метапараметрам функций  $p_1(\log p_1 - \log p_2)$  почти всюду и ограниченность интегрируемой функцией. Априорные распределения задаются непрерывными функциями плотности  $p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda)$ ,  $p(\Gamma|\mathbf{h}, \lambda)$ , не принимающими нулевое значение, и являющимися непрерывными по метапараметрам. Функция  $q(\mathbf{w}, \Gamma|\boldsymbol{\theta})$  принимает нулевое значение лишь в конечном количестве точек, поэтому функция  $q(\mathbf{w}, \Gamma|\boldsymbol{\theta})(\log q(\mathbf{w}, \Gamma|\boldsymbol{\theta}) - \log p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda))$  почти всюду непрерывна по метапараметрам. Она ограничена на компакте  $U_{\lambda}$ , поэтому слагаемое  $D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta}) || p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda))$  является непрерывным по метапараметрам. Выражения вида  $p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)(\log p(\Gamma|\mathbf{h}, \lambda) - \log p)$ ,  $p \in \mathbf{P}$  также являются непрерывными по метапараметрам и ограниченными, поэтому слагаемые вида  $D_{\text{KL}}(p(\Gamma|\mathbf{h}, \lambda) || p)$  являются непрерывными. Поэтому функции  $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$ ,  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \lambda)$  являются непрерывными по метапараметрам, что и требовалось доказать.  $\square$

Метапараметрами данной задачи (1.18) являются коэффициенты  $\lambda_{\text{prior}}^L$ ,  $\lambda_{\text{prior}}^Q$ , отвечающие за регуляризацию верхней и нижней задачи оптимизации, коэффициент  $\lambda_{\text{likelihood}}^Q$  отвечает за максимизацию правдоподобия, а также параметры распределений  $\mathbf{P}$  и вектор коэффициентов перед ними  $\lambda_{\text{struct}}^Q$ .

Условия 7-9 теоремы задают вид области  $U$ , на которой представленная оптимизационная задача является обобщающей. Условие 7 выполняется при



небольшом разбросе значений  $\log p(\mathbf{h}|\boldsymbol{\lambda})$  в зависимости от  $\lambda_1, \lambda_2$ . Т.к. эти метапараметры выполняют роль регуляризатора, для области гиперпараметров  $U_{\mathbf{h}}$ , выбранной адекватно, данное условие выполняется.

В случае, если  $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$  — нормальное распределение, а  $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$  — распределение Gumbel-softmax, такие что для любого  $\mathbf{h}$  существует  $\boldsymbol{\theta}$ :

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

а также полагая что  $\log p(\mathbf{h}|\boldsymbol{\lambda})$  приблизительно равен для всех  $\mathbf{h} \in U_{\mathbf{h}}$ , восьмое условие можно представить в следующем виде:

$$\begin{aligned} & \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\ & - \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda_{\text{prior}}^L} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) < \\ & < \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) - \\ & - \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

Это условие говорит, что существует набор метапараметров  $\boldsymbol{\lambda}$ , такой что максимальная разница дивергенций на  $U$  больше, чем максимальная разница между усредненными по  $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$  логарифмами правдоподобия выборки, поделенными на  $\lambda_{\text{likelihood}}^Q$ . Данное условие будет выполняться при достаточно больших  $\lambda_{\text{likelihood}}^Q$ .

Условие 9 выполняется при достаточно больших значениях метапараметра  $\lambda_{\text{struct}}^Q$ .

В предельном случае, когда температура  $\lambda_{\text{temp}}$  близка к нулю, а множество  $\mathbf{P}$  состоит из распределений, близких к дискретным, а соответствующим всем возможным структурам, калибровка  $\lambda_{\text{struct}}^Q$  порождает последовательность задач оптимизаций, схожую с перебором структур. Рассмотрим следующий пример.

**Пример 2.** Рассмотрим вырожденный случай поведения функции  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ , когда  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^Q = 0$ . Пусть модель использует один структурный параметр, в качестве априорного распределения на структуре задано распределение Gumbel-Softmax с  $\lambda_{\text{temp}}$ . Пусть в качестве множества распределений  $\mathbf{P}$  используется два распределения Gumbel-Softmax, сконцентрированных близко к вершинам симплекса:

$$\mathbf{P} = [\mathcal{GS}([0.95, 0.05, 0.05]^T, 1.0), \mathcal{GS}([0.95, 0.05, 0.05]^T, 1.0)].$$

Из определения распределения Gumbel-Softmax следует, что достаточно рассмотреть только значения параметра  $\mathbf{s}$ , находящиеся внутри симплекса. На рис. 1.5 изображены значения функции  $Q$  в зависимости от метапараметров  $\lambda_{\text{struct}}^Q$  и значений гиперпараметра  $\mathbf{s}$  распределения на структуре. Видно, что варьируя коэффициенты метапараметров получается последовательность оптимизаций, схожая с полным перебором структуры.

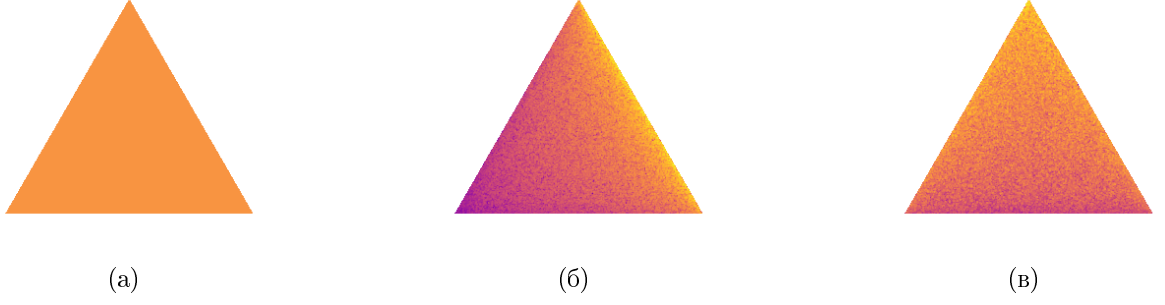


Рис. 1.5. Пример зависимости функции  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$  от гиперпараметра  $\mathbf{s}$  при различных значениях метапараметров  $\boldsymbol{\lambda}_{\text{struct}}^Q$ . Темные точки на графике соответствуют наименее предпочтительным значениям гиперпараметра. а)  $\boldsymbol{\lambda}_{\text{struct}}^Q = [0, 0]$ , б)  $\boldsymbol{\lambda}_{\text{struct}}^Q = [1, 0]$ , в)  $\boldsymbol{\lambda}_{\text{struct}}^Q = [1, 1]$ .

#### 1.4. Анализ обобщающей задачи

В данном разделе рассматриваются свойства предложенной задачи при различных значениях метапараметров, а также характер асимптотического поведения задач.

**Теорема 6.** Пусть  $m \gg 0$ ,  $\lambda_{\text{prior}}^L > 0$ ,  $\frac{m}{\lambda_{\text{prior}}^L} \in \mathbb{N}$ ,  $\frac{m}{\lambda_{\text{prior}}^L} \gg 0$ . Тогда оптимизация функции

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - \lambda_{\text{prior}}^L D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda}))$$

эквивалентна оптимизации вариационной оценки обоснованности

$$\mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda}))$$

для произвольной случайной подвыборки  $\hat{\mathbf{y}}, \hat{\mathbf{X}}$  мощности  $\frac{m}{\lambda_{\text{prior}}^L}$  из генеральной совокупности.

*Доказательство.* Рассмотрим величину  $\frac{1}{m}L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ :

$$\frac{1}{m}L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{1}{m}\mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - \frac{\lambda_{\text{prior}}^L}{m} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})). \quad (1.19)$$

При  $m \gg 0$  по усиленному закону больших чисел данная функция эквивалентна:

$$\begin{aligned} \frac{1}{m}L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) &\approx \mathbb{E}_{y, \mathbf{x}} \mathbb{E}_{q(\mathbf{w}, \Gamma|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) \\ &\quad - \frac{\lambda_{\text{prior}}^L}{m} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})||p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

Аналогично рассмотрим вариационную оценку обоснованности для произвольной выборки мощностью  $m_0 = \frac{m}{\lambda_{\text{prior}}^L}$ , усредненную на мощность выборки:

$$\begin{aligned} \frac{1}{m_0} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \frac{1}{m_0} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) &\approx \quad (1.20) \\ &\approx \mathbb{E}_{y, \mathbf{x}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \frac{1}{m_0} D_{\text{KL}}(p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda}) || q(\mathbf{w}, \Gamma | \boldsymbol{\theta})) = \\ &= \mathbb{E}_{y, \mathbf{x}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \frac{\lambda_{\text{prior}}^L}{m} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

Таким образом, задачи оптимизации функций (1.19), (1.20) совпадают, что и требовалось доказать.  $\square$

Теорема показывает, что для достаточно большого  $m$  и  $\lambda_{\text{prior}}^L > 0$ ,  $\lambda_{\text{prior}}^L \neq 1$  оптимизация параметров и гиперпараметров эквивалентна нахождению оценки обоснованности для выборки другой мощности: чем выше значение  $\lambda_{\text{prior}}^L$ , тем выше мощность выборки, для которой проводится оптимизация.

Следующие теоремы говорят о соответствии предлагаемой обобщающей задачи вероятностной модели. В частности, задача оптимизации параметров и гиперпараметров соответствует двухуровневому байесовскому выводу.

**Теорема 7.** Пусть  $\lambda_{\text{prior}}^Q = \lambda_{\text{prior}}^L = \lambda_{\text{likelihood}}^Q = 1$ ,  $\boldsymbol{\lambda}_{\text{struct}}^Q = \mathbf{0}$ . Тогда:

1. Задача оптимизации (1.18) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) + \\ + \log p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda}) \rightarrow \max_{\mathbf{h}}. \end{aligned}$$

2. Вариационное распределение  $q(\mathbf{w}, \Gamma | \boldsymbol{\theta})$  приближает апостериорное распределение  $p(\mathbf{w}, \Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  наилучшим образом:

$$D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})) \rightarrow \min_{\boldsymbol{\theta}}.$$

3. Если существуют такие значения параметров  $\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\Gamma}$ , что  $p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \Gamma, \mathbf{h}, \boldsymbol{\lambda}) = q_{\mathbf{w}}(\mathbf{w} | \Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ ,  $p(\Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = q_{\Gamma}(\Gamma | \boldsymbol{\theta}_{\Gamma})$ , то решение задачи оптимизации  $L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  доставляет эти значения вариационных параметров.

*Доказательство.* Так как параметры  $\boldsymbol{\theta}$  не зависят от слагаемых при коэффициентах  $\boldsymbol{\lambda}_{\text{struct}}^Q$ , а также от  $\log p(\mathbf{h} | \boldsymbol{\lambda})$ , то при  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^L = 1$  как верхняя, так и нижняя задачи оптимизации (1.18) эквивалентны оптимизации вариационной оценки обоснованности, поэтому первое утверждение выполняется.

Докажем второе утверждение. Рассмотрим логарифм обоснованности модели:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log \frac{p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})}{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} + D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) = \\ &= \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) + \\ &\quad + D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})).\end{aligned}$$

Из данного равенства следует:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})) = \\ \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})),\end{aligned}$$

где правая часть равенства соответствует вариационной оценки обоснованности. Выражение  $\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  не зависит от вариационного распределения  $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ , поэтому максимизации вариационной оценки эквивалентна минимизации дивергенции  $D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}))$ .

Докажем третье утверждение. Т.к. вариационное распределение  $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$  декомпозируется на  $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ ,  $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ , апостериорное распределение  $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  декомпозируется на  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ ,  $p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ , поэтому достижимо значение нулевое значение дивергенции:  $D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})) = 0$ . Она представима в следующем виде (1.6). Отсюда следует что соответствующие вариационные и апостериорные распределения совпадают.  $\square$

Таким образом, предлагаемая обобщающая задача позволяет производить оптимизацию вариационной оценки обоснованности, а также оптимизацию обоснованности для выбор с другим эффективным размером. Чем больше размер выборки, тем больше влияние априорного распределения, которое выступает в качестве регуляризатора. Можно регулировать сложность модели следующим образом:

1. Варьируя сложность на верхнем уровне оптимизации оптимизации с использованием коэффициента  $\lambda_{\text{prior}}^Q$ ;
2. Варьируя сложность на нижнем уровне оптимизации оптимизации с использованием коэффициента  $\lambda_{\text{prior}}^L$ ;
3. Варьируя сложность на обоих уровнях оптимизации.

Рассмотрим различие вариантов 1-3 на примере.

**Пример 3.** Пусть  $\lambda_{\text{struct}}^Q = 0$ . Пусть требуется уменьшить вклад априорного распределения в итоговую оптимизацию. При варьировании нижней задачи оптимизации ( $\lambda_{\text{prior}}^L \rightarrow 0$ ) оптимизационная задача становится эквивалента методу максимального правдоподобия:

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

При этом верхняя задача  $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \rightarrow \max_{\mathbf{h}}$  не имеет смысла, т.к. параметры  $\boldsymbol{\theta}$  не зависят от гиперпараметров  $\mathbf{h}$ .

При варьировании только верхней задачи оптимизации ( $\lambda_{\text{prior}}^Q \rightarrow 0, \lambda_{\text{prior}}^L = \lambda_{\text{likelihood}}^Q = 1$ ), на нижнем уровне задача  $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$  совпадает с задачей поиска обоснованных параметров при фиксированном  $\mathbf{h}$ :

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})).$$

При этом на верхнем уровне оптимизации мы ищем те гиперпараметры  $\mathbf{h}$ , при которых параметры будут давать максимум правдоподобия:

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

Таким образом, варьирование сложности на верхнем оптимизации приводит к оптимизации параметров и гиперпараметров с большей регуляризацией, чем варьирование сложности на нижнем уровне оптимизации.

Следующая теорема анализирует оптимизацию при  $\frac{\lambda_{\text{prior}}^Q}{\lambda_{\text{likelihood}}^Q} = \lambda_{\text{prior}}^L$ . В частности, если  $\lambda_{\text{likelihood}}^Q = 1$ , то такая оптимизация соответствует оптимизации вариационной оценки обоснованности на обоих уровнях оптимизации для выборки размера  $\lceil \frac{m}{\lambda_{\text{prior}}^L} \rceil$ , о чем говорилось в Теореме 1.4.

**Теорема 8.** Пусть  $\frac{\lambda_{\text{prior}}^Q}{\lambda_{\text{likelihood}}^Q} = \lambda_{\text{prior}}^L$ . Тогда задача оптимизации (1.18) представима в виде одноуровневой задачи оптимизации:

$$\begin{aligned} & \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \lambda_{\text{prior}}^Q D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) - \\ & - \sum_{p', \lambda \in \mathbf{P}, \lambda_{\text{struct}}^Q} D_{\text{KL}}(p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||p') - \log p(\mathbf{h}|\boldsymbol{\lambda}) \rightarrow \max_{\mathbf{h}, \boldsymbol{\theta}}. \end{aligned}$$

*Доказательство.* Т.к. выполнено равенство  $\frac{\lambda_{\text{prior}}^Q}{\lambda_{\text{likelihood}}^Q} = \lambda_{\text{prior}}^L$ , то нижняя задача оптимизации эквивалентна следующей задаче:

$$\begin{aligned} & \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\ & - \lambda_{\text{prior}}^Q D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

Параметры  $\boldsymbol{\theta}$  вариационного распределения  $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$  не зависят от слагаемых вида  $\log p(\mathbf{h}|\boldsymbol{\lambda})$  и  $D_{\text{KL}}(p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||p'), p' \in \mathbf{P}$ , поэтому нижняя задача оптимизации эквивалентна следующей задаче:

$$\begin{aligned} & \lambda_{\text{likelihood}}^Q \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\ & - \lambda_{\text{prior}}^Q D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})). \end{aligned}$$

$$- \sum_{p', \lambda \in \mathbf{P}, \boldsymbol{\lambda}_{\text{struct}}^Q} D_{\text{KL}}(p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) || p') + \log p(\mathbf{h}|\boldsymbol{\lambda}) \rightarrow \max_{\boldsymbol{\theta}}$$

для любого вектора  $\boldsymbol{\lambda}_{\text{struct}}^Q$ .

а значит верхняя и нижняя задачи совпадают:

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

где

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

Из свойства

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

следует доказательство теоремы. □

## Список основных обозначений

$\mathbf{x}_i \in \mathbf{X}$  — вектор признакового описания  $i$ -го объекта  
 $y_i \in \mathbf{y}$  — метка  $i$ -го объекта  
 $\mathcal{D}$  — 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  $(j, k)$   
 $\text{agg}_v$  — функция агрегации для вершины  $v$   
 $\gamma^{j,k}$  — 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  
 $\mathbf{\Gamma} \in \mathbb{\Gamma}$  — структура модели  
 $\mathbb{\Gamma}$  — множество значений структуры модели  
 $U_{\mathbf{\Gamma}} \subset \mathbb{\Gamma}$  — 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  
 $\boldsymbol{\theta} \in \Theta$  — параметры вариационного распределения  
 $\Theta$  — пространство параметров вариационного распределения  
 $U_{\boldsymbol{\theta}} \subset \Theta$  — область определения вариационных параметров модели  
 $\boldsymbol{\theta}_{\mathbf{w}} \in \Theta_{\mathbf{w}}$  — параметры вариационного распределения, аппроксимирующего апостериорное распределение параметров модели  
 $\Theta_{\mathbf{w}}$  — пространство параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели  
 $U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \Theta_{\mathbf{w}}$  — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели  
 $\boldsymbol{\theta}_{\mathbf{\Gamma}} \in \Theta_{\mathbf{\Gamma}}$  — параметры вариационного распределения, аппроксимирующего апостериорное распределение структуры модели  
 $\Theta_{\mathbf{\Gamma}}$  — пространство параметров вариационного распределения, аппроксимирующего апостериорное распределение структуры модели  
 $U_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} \subset \Theta_{\mathbf{\Gamma}}$  — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение структуры модели

$\lambda \in \Lambda$  — вектор метапараметров

$\Lambda$  — пространство метапараметров

$U_\lambda \subset \Lambda$  — область определения метапараметров

$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma)$  — правдоподобие выборки

$p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)$  — априорное распределение параметров и структуры модели

$p(\mathbf{h}|\lambda)$  — распределение гиперпараметров модели

$p(\Gamma|\mathbf{h}, \lambda)$  — априорное распределение структуры модели

$p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda)$  — априорное распределение параметров модели

$p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$  — апостериорное распределение параметров и структуры модели

$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \Gamma, \mathbf{h}, \lambda)$  — апостериорное распределение структуры модели

$p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$  — апостериорное распределение структуры модели

$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \lambda)$  — апостериорное распределение гиперпараметров

$p(y, \mathbf{w}, \Gamma|\mathbf{x}, \mathbf{h})$  — вероятностная модель глубокого обучения

$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda)$  — обоснованность модели

$q(\mathbf{w}, \Gamma|\boldsymbol{\theta})$  — вариационное распределение параметров и структуры модели

$q_{\mathbf{w}}(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  — вариационное распределение структуры модели

$q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})$  — вариационное распределение параметров модели

$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$  — функция потерь

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

$T(\boldsymbol{\theta}|L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda))$  — оператор оптимизации

$\mathfrak{Q}$  — семейство вариационных распределений

$S$  — энтропия распределения

$M$  — множество моделей без общей параметризации

$D_{\text{KL}}(p_1||p_2)$  — дивергенция Кульбака-Лейблера между распределениями  $p_1$  и  $p_2$

$\mathbf{A}^{-1}$  — матрица ковариаций параметров модели

$\mathbf{s}$  — конкатенация параметров концентрации на структуре модели