

# 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 обоснованности. Показано, что данная функция оптимизирует несколько 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. 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  $\gamma_l^{j,k}$ .

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

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

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

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

где  $\mathbf{s} \in (0, \infty)^K$  — гиперпараметр, отвечающий за смещенность плотности распределения относительно точек симплекса на  $K$  вершинах,  $\lambda_{\text{temp}}$  — метапараметр температуры, отвечающий за концентрацию плотности вблизи вершин симплекса или в центре симплекса.

- Перечислим свойства, которыми обладает распределение Gumbel-Softmax:
1. Реализация  $\hat{\gamma}_l$ , т.е.  $l$ -й компоненты случайной величины  $\gamma$  порождается следующим образом:

$$\hat{\gamma}_l = \frac{\exp(\log s_l + \hat{g}_l)/\lambda_{\text{temp}}}{\sum_{l'=1}^K \exp(\log s_{l'} + \hat{g}_{l'})/\lambda_{\text{temp}}},$$

где  $\hat{\gamma} \sim -\log(-\log \mathcal{U}(0, 1)^K)$ .

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

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

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

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

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

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

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

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

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

Рассмотрим два случая: когда вектор  $\gamma$  лежит не в центре симплекса, и когда  $\gamma$  лежит в центре симплекса. Пусть хотя бы для одной компоненты  $l$  выполнено:  $\gamma_l \neq \frac{1}{K}$ . Пусть  $l'$  соответствует индексу максимальной компоненты вектора  $\gamma$ . Для  $l = l'$  предел выражения (1.3) при  $\lambda_{\text{temp}}$  стремится к бесконечности. Для  $l \neq l'$  предел выражения (1.3) при  $\lambda_{\text{temp}}$  стремится к нулю. Возводя сумму пределов в степень  $(-K)$  получаем предел плотности, равный нулю.

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

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

**Определение 1.** Репараметризацией случайной величины  $\psi$ , распределенную по распределению  $q$  с параметрами  $\theta_\psi$  назовем представление величины с помощью другой случайной величины, имеющей распределение, не зависящее от параметров  $\theta$ :

$$\psi \sim q \iff \hat{\psi} \sim g(\epsilon, \theta_\psi),$$

где  $\epsilon$  — случайная величина, чье распределение не зависит от параметров  $\theta_\psi$ ,  $g$  — некоторая детерминированная функция,  $\hat{\psi}$  — реализация случайной величины  $\psi$ .

Идею репараметризации поясним на следующем примере.

**Пример 1.** Пусть структура  $\Gamma$  определена для модели  $\mathbf{f}$  однозначно. Рассмотрим математическое ожидание логарифма правдоподобия выборки модели по некоторому непрерывному распределению  $q$ :

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

Продифференцируем данное выражение по параметрам  $\theta_{\mathbf{w}}$  вариационного распределения  $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})$ :

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

Выражение общем виде не имеет аналитического решения. Пусть распределение  $q$  для параметров  $\mathbf{w}$  подлежит репараметризации:

$$\mathbf{w} \sim q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}}) \iff \hat{\mathbf{w}} \sim g(\epsilon, \theta_{\mathbf{w}}).$$

Тогда справедливо следующее выражение:

$$\begin{aligned} \nabla_{\theta_{\mathbf{w}}} \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) &= \nabla_{\theta_{\mathbf{w}}} \mathbb{E}_{\epsilon} \log p(\mathbf{y}|g(\epsilon, \theta_{\mathbf{w}}), \mathbf{X}, \mathbf{h}, \lambda) = \\ &= \int_{\epsilon} \nabla_{\theta_{\mathbf{w}}} \log p(\mathbf{y}|g(\epsilon, \theta), \mathbf{X}, \mathbf{h}, \lambda) p(\epsilon) d\epsilon = \mathbb{E}_{\epsilon} \nabla_{\theta} \log p(\mathbf{y}|g(\epsilon, \theta), \mathbf{X}, \mathbf{h}, \lambda). \end{aligned}$$

Таким образом, распределение, позволяющее произвести репараметризацию, является более удобным для вычисления интегральных оценок. Кроме того, данный подход позволяет значительно повысить точность вычисления градиента от функций, зависящих от случайных величин [?].

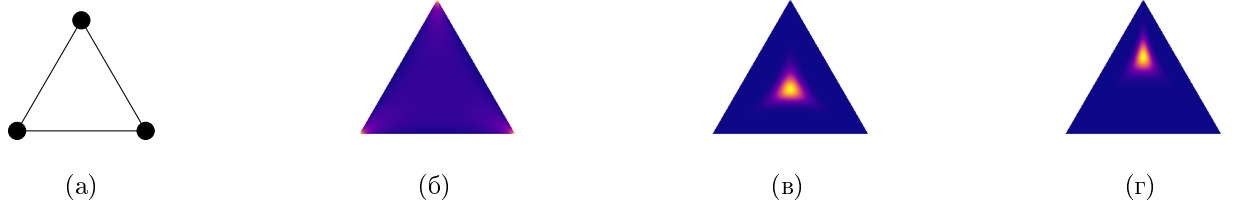


Рис. 1.1. Пример распределения Gumbel-Softmax при различных значениях параметров: а)  $\lambda_{\text{temp}} \rightarrow 0$ , б)  $\lambda_{\text{temp}} = 1, \mathbf{s} = [1, 1, 1]$ , в)  $\lambda_{\text{temp}} = 5, \mathbf{s} = [1, 1, 1]$ , г)  $\lambda_{\text{temp}} = 5, \mathbf{s} = [10, 0.1, 0.1]$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

Пример распределения Gumbel-Softmax при различных параметрах представлен на Рис. 1.1. В качестве альтернативы для априорного распределения на структуре выступает распределение Дирихле. В качестве предельного случая, когда все структуры равнозначны, выступает равномерное распределение. Выбор в качестве распределения на структуре произведения Gumbel-Softmax распределения обоснован выбором этого же 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. Варьируя метапарамы  $\lambda_1, \lambda_2$  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.
2. Структура модели  $\boldsymbol{\Gamma}$ , содержащая все структурные параметры  $\{\gamma^{j,k}, (j, k) \in E\}$  распределены 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.
4. Метапараметры:  $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_{\text{temp}}]$ . Эти параметры не подлежат оптимизации и задаются экспертно.

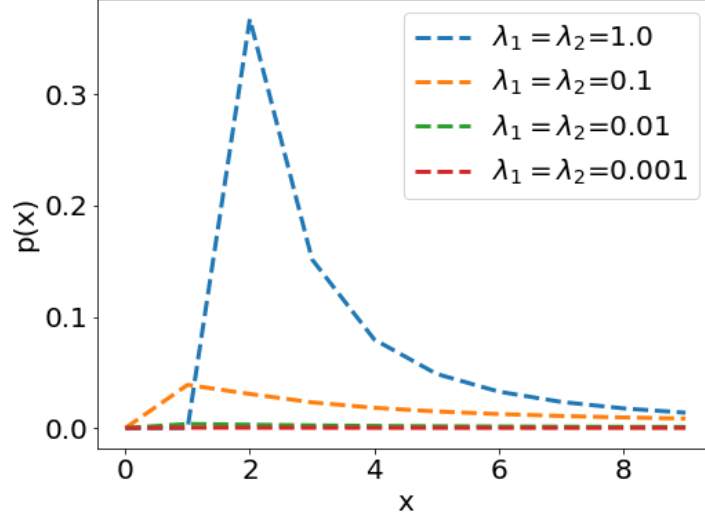


Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

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

## 1.2. Вариационная оценка для обоснованности вероятностной модели

В качестве критерия выбора структуры модели предлагается использовать апостериорную вероятность гиперпараметров:

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

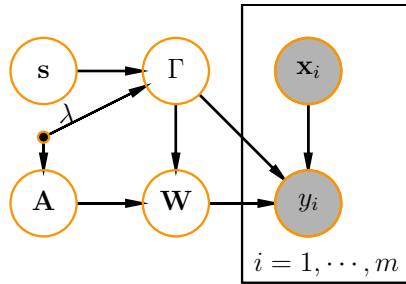


Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

где структура модели и параметры модели выбираются на основе полученных значений гиперпараметров:

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

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

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

Для вычисления обоснованности

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

из (1.4) предлагается использовать 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:

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

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

$$q_{\Gamma}(\Gamma | \boldsymbol{\theta}_{\Gamma}) \approx q_{\Gamma}(\Gamma | \boldsymbol{\theta}_{\Gamma}).$$

Тогда справедлива следующая оценка:

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

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

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.* Рассмотрим обоснованность:

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

Используя неравенство Йенсена получим

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

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

Декомпозируем распределение  $q$  по свойству условной дивергенции:

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

□

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. ??.

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:

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

Параметрическая сложность модели соответствует 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 [?]:

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

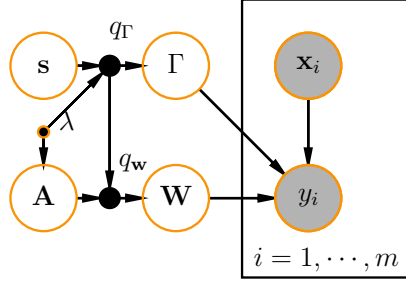


Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

где  $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})(w|\boldsymbol{\theta}_{\mathbf{w}}) \sim \mathcal{N}(\mu_q, \alpha_q)$ .

Обобщим понятие относительной вариационной плотности на случай произвольных непрерывных распределений.

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

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

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

**Теорема 2.** Пусть

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

$$\text{mode } p(\mathbf{w}|\mathbf{\Gamma}_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\mathbf{\Gamma}_2, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{M} \forall \mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in U_{\boldsymbol{\theta}_{\mathbf{\Gamma}}}.$$

3. вариационное распределение  $q_{\mathbf{w}}$  и априорное распределение  $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})$  являются абсолютно непрерывными и унимодальными на  $U_{\mathbf{h}}, U_{\boldsymbol{\theta}}$ .
4. Решение задачи вида

$$\mathbf{h} = \arg \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h})) \quad (1.6)$$

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



5. Параметры модели  $\mathbf{w}$  имеют конечные вторые моменты по маргинальным распределениям  $q(\mathbf{w}|\boldsymbol{\theta}_w), p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda})$ :

$$\mathbb{E}_q \mathbf{w}^2 = \mathbb{E}_{q_\Gamma} \mathbb{E}_{q_w} \mathbf{w}^2 < \infty;$$

$$\mathbb{E}_{p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})} \mathbf{w}^2 = \mathbb{E}_{p(\Gamma|\mathbf{h}, \boldsymbol{\lambda})} \mathbb{E}_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w}^2 < \infty,$$

где

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

6. мода и матожидание вариационного распределения  $q_w$  и априорного распределения  $p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})$  совпадают:

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

$$\text{mode } q(\mathbf{w}|\Gamma, \boldsymbol{\theta}_w) = \mathbb{E}_{q_w(\mathbf{w}|\Gamma, \boldsymbol{\theta}_w)} \mathbf{w};$$

7. задана бесконечная последовательность векторов вариационных параметров  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_i \in U_{\boldsymbol{\theta}}$ , такая что  $\lim_{i \rightarrow \infty} C_p(\boldsymbol{\theta}_i|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$ .

Тогда следующее выражение стремится к единице:

$$\mathbb{E}_{q_\Gamma} \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_w, \mathbf{h}, \boldsymbol{\lambda})^{-1} \rightarrow 1.$$

*Доказательство.* Воспользуемся неравенством Пинскера:

$$\|F_q(\boldsymbol{\theta}) - F_p(\mathbf{h})\|_{\text{TV}} \leq \sqrt{2D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}))},$$

где  $\|\cdot\|_{\text{TV}}$  — расстояние по вариации,  $F_q, F_p$  — функции распределения  $q(\mathbf{w}, \Gamma|\boldsymbol{\theta})$  и  $p(\mathbf{w}, \Gamma|\mathbf{h}, \boldsymbol{\lambda})$ . Т.к. дивергенция состоит из двух неотрицательных величин, то обе они стремятся к нулю. Рассмотрим вторую величину:

$$\lim 0 = D_{\text{KL}}(q_w|p(\mathbf{w})) = \int_{\Gamma} D_{\text{KL}}(q_w|p(\mathbf{w})q(\Gamma)) \geq \int_{\Gamma} \|F_q - F_p\|_{\text{TV}}.$$

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

$$\lim \frac{1}{2} \iint_{\mathbf{w}, \Gamma} |p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}) - q(\mathbf{w}|\boldsymbol{\theta})| q(\Gamma) d\Gamma d\mathbf{w} = 0.$$

Для произвольного  $\boldsymbol{\theta}_j$  рассмотрим выражение:

$$\begin{aligned} \left| \int_{\Gamma} \frac{q_j(\mathbb{E}_q \mathbf{w})}{q(\mathbf{M})} - \frac{q_j(\mathbb{E}_p \mathbf{w})}{q(\mathbf{M})} \right| &\leq \int_{\Gamma} \left| \frac{q_j(\mathbb{E}_q \mathbf{w})}{q(\mathbf{M})} - \frac{q_j(\mathbb{E}_p \mathbf{w})}{q(\mathbf{M})} \right| \leq \frac{\max L}{\min q(\mathbf{M})} \int_{\Gamma} |\mathbb{E}_q \mathbf{w} - \mathbb{E}_p \mathbf{w}| \leq \\ &\leq L \int_{\Gamma} |\mathbf{w}| |q - p|. \end{aligned}$$

Т.к. вторые моменты  $\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}^2, \mathbb{E}_{p(\mathbf{w}|\mathbf{h})} \mathbf{w}^2$  конечны, то случайная величина  $\mathbf{w}$  равномерно интегрируема как при маргинальном распределении  $q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})$ , так и при маргинальном распределении  $p(\mathbf{w}|\mathbf{h})$ . Определим случайную величину  $\boldsymbol{\nu}(t), t \geq 0$  следующим образом:

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

Данная величина совпадает с  $\mathbf{w}$  при  $|\mathbf{w}| < t$  и принимает значение  $t$  или  $-t$  при  $|\mathbf{w}| \geq t$ ,

По определению равномерной интегрируемости для  $\mathbf{w}$  для любого числа  $\varepsilon$  существует число  $t_0$ , такое что для любого  $t \geq t_0$  справедливо выражение:

$$\mathbb{E}|\mathbf{w} - \boldsymbol{\nu}(t)| \leq \varepsilon.$$

где матожидание берется по распределениям  $q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})$  и  $p(\mathbf{w}|\mathbf{h})$ . Тогда

$$\begin{aligned} \int_{\Gamma} |\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w} - \mathbb{E}_{p(\mathbf{w}|\mathbf{h})} \mathbf{w}| &\leq \int_{\Gamma} \int_{\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| |p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} + \\ &+ \int_{\Gamma} \int_{\mathbf{w}} |\boldsymbol{\nu}(t)| |p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} \leq \\ &\leq \int_{\Gamma} \int_{\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| p(\mathbf{w}|\mathbf{h}) d\mathbf{w} + \int_{\Gamma} \int_{\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w} + \\ &+ \int_{\Gamma} \int_{\mathbf{w}} |\boldsymbol{\nu}(t)| |p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w}. \end{aligned}$$

Обозначим за  $\mathbf{h}_i$  — решение задачи (??) для вектора  $\boldsymbol{\theta}_i$ . Т.к.  $|\boldsymbol{\nu}(t)|$  — ограничена, то

$$\begin{aligned} \int_{\Gamma} \int_{\mathbf{w}} |\boldsymbol{\nu}(t)| |p(\mathbf{w}|\mathbf{h}_i) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} &\leq t \int_{\Gamma} \int_{\mathbf{w}} |p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} = \\ &= t \int_{\Gamma} \int_{\mathbf{w}} |p(\mathbf{w}, \Gamma|\mathbf{h}_i, \boldsymbol{\lambda}) - q(\mathbf{w}, \Gamma|\boldsymbol{\theta})| d\Gamma d\mathbf{w} \rightarrow_{i \rightarrow \infty} 0. \end{aligned}$$

$$\lim_{i \rightarrow \infty} \int_{\Gamma} |\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w} - \mathbb{E}_{p(\mathbf{w}|\mathbf{h}_i)} \mathbf{w}| \leq \int_{\Gamma} \int_{\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| |p(\mathbf{w}|\mathbf{h}_i) - q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w}$$

для любого  $t$ . Устремляя  $t$  к бесконечности, получим  $\lim_{i \rightarrow \infty} \int_{\Gamma} |\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w} - \mathbb{E}_{p(\mathbf{w}|\mathbf{h}_i)} \mathbf{w}| = 0$ . Таким образом, мода  $q(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$  стремится в среднем к моде априорного распределения  $\mathbf{M}$ .

□

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  $\mathbf{\Gamma}$  определена однозначно, последовательность  $q_i$  может являться последовательностью нормальных распределений, чье матожидание стремится к нулю:

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

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

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

При этом сама последовательность распределений  $q_i$  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  $q$  для аппроксимации неизвестного апостериорного распределения параметров  $p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ .

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

$$L = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \mathbf{\Gamma}). \quad (1.7)$$

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

$$L = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) \rightarrow \max_{\boldsymbol{\theta}},$$

$$Q = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \mathbf{\Gamma}).$$

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L = Q = \mathbb{E}_q(\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) + \log p(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\lambda})). \quad (1.8)$$

Т.к. в рамках данной задачи (??) 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. Перебор структуры:

$$L = Q = \mathbb{E}_q \log p(\mathbf{y}, \mathbf{w}|\mathbf{X})[q_{\mathbf{\Gamma}} = p'] \quad (1.9)$$

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:

$$L = Q = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - |\{w : D_{\text{KL}}(\theta, \mathbf{h}) < \lambda_{\text{prune}}\}|, \quad (1.10)$$

где

$$\mathbf{h} = \arg \min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}', \boldsymbol{\lambda})), \quad (1.11)$$

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

### 5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

$$L = Q = \text{BIC}_{\lambda} = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \log(m) |\{w : D_{\text{KL}}(\theta, \mathbf{h}) < \lambda_{\text{prune}}\}|, \quad (1.12)$$

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

6. Метод вариационной оценки обоснованности:

$$L = \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\Gamma, \mathbf{w}|\mathbf{h}, \boldsymbol{\lambda})) + \log p(\mathbf{h}|\boldsymbol{\lambda}) \rightarrow \max_{\boldsymbol{\theta}}, \quad (1.13)$$

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

В рамках данной задачи функции  $L$  и  $Q$  совпадают, все гиперпараметры  $\mathbf{h}$  подлежат оптимизации.

7. Валидация на отложенной выборке:

$$L = \mathbb{E}_q \log p(\mathbf{y}_{\text{train}}, \mathbf{w}, \Gamma|\mathbf{X}_{\text{train}}, \mathbf{h}, \boldsymbol{\lambda}) \rightarrow \max_{\boldsymbol{\theta}}, \quad (1.14)$$

$$Q = \mathbb{E}_q \log p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{w}, \Gamma) \rightarrow \max_{\mathbf{h}},$$

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 критерии 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 модели.

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  $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \Theta \times \mathbb{H} \times \mathbb{A}$ , 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

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

определено однозначно при любых значениях метапараметров  $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ .

3. Критерий максимизации правдоподобия выборки: существует  $\boldsymbol{\lambda} \in U_\lambda$  и

$$K_1 > 0, \quad K_1 < \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_h} Q(\mathbf{h}_1) - Q(\mathbf{h}_2),$$

такие что для любых векторов гиперпараметров, удовлетворяющих неравенству

$$\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1,$$

выполняется неравенство

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

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

$$K_2 > 0, \quad K_2 < \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_h} Q(\mathbf{h}_1) - Q(\mathbf{h}_2),$$

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

$$Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2,$$

параметрическая сложность первой модели меньше, чем второй:

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

5. Критерий приближения оценки обоснованности: существует значение гиперпараметров  $\boldsymbol{\lambda}$ , такое что значение функций потерь  $L$  и валидации  $Q$  пропорционален вариационной оценке обоснованности модели:

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

для всех  $\boldsymbol{\theta} \in U_\theta, \mathbf{h} \in U_h$ .

6. Критерий перебора оптимальных структур: существует набор метапараметров  $\boldsymbol{\lambda}$  и константа

$$K_3 > 0, \quad K_3 < \max_{\mathbf{h}_1, \mathbf{h}_2} D_{KL}(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})),$$

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

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

$$Q(\mathbf{h}_1) > Q(\mathbf{h}_2),$$

существует значение метапараметров  $\boldsymbol{\lambda}'$ , такие что

- (a) Соответствие между вариационными параметрами  $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$  сохраняется при  $\boldsymbol{\lambda}'$ .
- (b)  $Q(\mathbf{h}_1) < Q(\mathbf{h}_2)$  при  $\boldsymbol{\lambda}'$ .

7. Критерий непрерывности: функции  $L$  и  $Q$  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  $p(\Gamma|\mathbf{s}, \lambda)$  значимо высока. При этом соответствующие вариационные распределения  $q(\Gamma|\theta_\Gamma)$  могут оказаться достаточно близки. Возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения  $q(\Gamma|\theta_\Gamma)$  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  $(??), (??), (??), (??), (??), (??)$  не являются обобщающими.

*Доказательство.* Задачи  $(??), (??), (??), (??), (??)$  не имеют гиперпараметров  $\mathbf{h}$ , подлежащих оптимизации, поэтому не могут оптимизировать вариационную оценку.

При использовании валидации на отложенной выборке  $(??)$  в функцию валидации  $Q$  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  $\nabla_{\theta_\Gamma} q(\Gamma|\theta_\Gamma)$  является нулевым не более чем счетное количество раз.
2. выражение  $\nabla_{\theta_\Gamma} q(\Gamma|\theta_\Gamma) \log p(\Gamma|\mathbf{h}, \lambda)$  ограничено на  $U_\theta$  некоторой случайной величиной с конечным первым моментом.

Тогда задача  $(??)$  не является обобщающей.

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

$$\nabla_{\theta} E_{q(\mathbf{w}, \Gamma|\theta_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma) - \nabla_{\theta} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\theta_1)|p(\mathbf{w}, \Gamma|\mathbf{h}_1, \lambda)) =$$

$$= \nabla_{\theta} \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta_1)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \nabla_{\theta} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta_1) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \lambda')).$$

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\theta} D_{\text{KL}}(q(\Gamma | \theta_2) | p(\Gamma | \lambda)) = \nabla_{\theta} D_{\text{KL}}(q(\Gamma | \theta_2) | p(\Gamma | \lambda')),$$

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

$$\int_{\Gamma \in \Gamma} \nabla_{\theta} q(\Gamma | \theta_2) (\log p(\Gamma | \lambda) - \log p(\Gamma | \lambda')) d\Gamma = 0.$$

Т.к. выражение  $\nabla_{\theta} q(\Gamma | \theta_2)$  принимает нулевое значение в счетном количестве точек, то выражение  $\log p(\Gamma | \lambda) - \log p(\Gamma | \lambda')$  равно нулю почти всюду, что означает что метапараметр температуры  $\lambda_{\text{temp}}$  равен:

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

Таким образом, метапараметры  $\lambda, \lambda'$  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  $Q$  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  $Q$  не меняется при изменении метапараметров  $\lambda$ .  $\square$

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

$$\begin{aligned} \mathbf{h}^* &= \arg \max_{\mathbf{h}} Q = \\ &= \lambda_{\text{likelihood}}^Q \mathbb{E}_{q^*} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - \\ &\quad - \lambda_Q^{\text{prior}} D_{\text{KL}}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})) - \\ &\quad - \sum_{p' \in \mathbf{P}, \lambda \in \lambda_Q^{\text{struct}}} \lambda D_{\text{KL}}(\Gamma | p') + \log p(\mathbf{h} | \mathbf{f}), \\ q^* &= \arg \max_q L = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) \\ &\quad - \lambda_L^{\text{prior}} D_{\text{KL}}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})), \end{aligned} \tag{1.15} \tag{L^*}$$

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



**Теорема 5.** Пусть:

- 1) задано непустое множество непрерывных по параметрам распределений на структуре  $\mathbf{P}$ , где хотя бы одно распределение принадлежит Gumbel-Softmax-распределению.
- 2) вариационное распределение  $q = q_{\Gamma}(\Gamma|\theta_{\Gamma})q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\Gamma})$  является абсолютно непрерывным, плотность которого непрерывна по метапараметрам  $\lambda$ ;
- 3) задан компакт  $U = U_{\theta} \times U_{\mathbf{h}} \times U_{\lambda} \subset \Theta \times \mathbb{H} \times \Lambda$ , где параметры распределений  $\mathbf{P} \in \Lambda$ , область  $U_{\theta}$  декомпозируется на две области  $U_{\theta} = U_{\theta_{\mathbf{w}}} \times U_{\theta_{\Gamma}}$ ;
- 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  $\mathbf{h}$  оптимальное решение нижней задачи оптимизации  $\theta^*$  определено однозначно при любых значениях метапараметров  $\lambda \in U_{\lambda}$ ;
- 6) область значений метапараметров  $\lambda_{\text{likelihood}}^Q, \lambda_Q^{\text{prior}}, \lambda_Q^{\text{struct}}, \lambda_L^{\text{prior}}$  включает отрезок от нуля до единицы;
- 7) существует значение метапараметров  $\lambda_1, \lambda_2, \lambda_{\text{likelihood}}^Q$ , такое что

$$\max_{\mathbf{h}} \log p(\mathbf{h}|\lambda) - \min_{\mathbf{h}} \log p(\mathbf{h}|\lambda) < \max_{\mathbf{h}} Q(\mathbf{h}) - \min_{\mathbf{h}} Q(\mathbf{h}).$$

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

- 8) существует значение метапараметров  $\lambda_Q^{\text{prior}}, \lambda_1, \lambda_2, \lambda_{\text{temp}}$ , такое что

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

- 9) существуют значения метапараметров  $\lambda_Q^{\text{prior}}, \lambda_Q^{\text{likelihood}}, \lambda_1, \lambda_2, \lambda_{\text{temp}}$ , такие что

$$\max_{\mathbf{h}} D_{\text{KL}} - \min_{\mathbf{h}} D_{\text{KL}} < \frac{\max_{\mathbf{h}} Q - \min_{\mathbf{h}} Q}{\max_{\lambda} \lambda_{\text{struct}}}$$

при  $\lambda_Q^{\text{struct}} = 0$ .

Тогда задача (??) 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  $\lambda_Q^{\text{prior}} = 0, \lambda_Q^{\text{struct}} = 0$ . Пусть  $\lambda_1, \lambda_2, \lambda_{\text{likelihood}}^Q$  удовлетворяют седьмому условию теоремы. Возьмем в качестве  $K_1$  следующее выражение:

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

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

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

Отсюда следует выполнение критерия 3:

$$\lambda_Q^{\text{likelihood}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) - \lambda_Q^{\text{likelihood}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma) > 0.$$

Докажем критерий 4. Пусть  $\boldsymbol{\lambda}$  удовлетворяют восьмому условию и  $\lambda_Q^{\text{likelihood}} = 0, \boldsymbol{\lambda}_Q^{\text{struct}} = \mathbf{0}$ . Пусть

$$K_2 = \max_{\mathbf{h}} \log p(\mathbf{h} | \boldsymbol{\lambda}) - \min_{\mathbf{h}} \log p(\mathbf{h} | \boldsymbol{\lambda}) + \max_{\mathbf{h}} \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - \\ - \min_{\mathbf{h}, \boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) + \max_{\boldsymbol{\theta}} \frac{1}{\lambda_L^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) - \\ - \min_{\boldsymbol{\theta}} \frac{1}{\lambda_L^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}).$$

Пусть  $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$ . Рассмотрим разность параметрических сложностей двух векторов:

$$C_p(\boldsymbol{\theta}_2) - C_p(\boldsymbol{\theta}_1) = \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - \\ - \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) \geq \\ \geq \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \boldsymbol{\lambda})) + \\ + D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2) | p(\mathbf{w}, \Gamma | \mathbf{h}_2, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2) | p(\mathbf{w}, \Gamma | \mathbf{h}_2, \boldsymbol{\lambda})) = \\ = Q(\mathbf{h}_1) - Q(\mathbf{h}_2) - \log p(\mathbf{h}_1 | \boldsymbol{\lambda}) + \log p(\mathbf{h}_2 | \boldsymbol{\lambda}) + \\ + \min_{\mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_2) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \boldsymbol{\lambda})) > \\ > K_2 - \log p(\mathbf{h}_1 | \boldsymbol{\lambda}) + \log p(\mathbf{h}_2 | \boldsymbol{\lambda}) + \min_{\boldsymbol{\theta}, \mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - \\ - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \boldsymbol{\lambda})).$$

Рассмотрим разность:

$$\min_{\boldsymbol{\theta}, \mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \boldsymbol{\lambda})) = \\ = \min_{\boldsymbol{\theta}, \mathbf{h}} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda})) - \frac{1}{\lambda_L^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \Gamma | \boldsymbol{\theta}_1) \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) + \\ + \max_{\boldsymbol{\theta}} \left( \frac{1}{\lambda_L^{\text{prior}}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \boldsymbol{\theta})} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) - D_{\text{KL}}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) | p(\mathbf{w}, \Gamma | \mathbf{h}_1, \boldsymbol{\lambda})) \right) \geq$$

$$\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{L}}^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \\
&+ \max_{\boldsymbol{\theta}} \left( \min_{\boldsymbol{\theta}'} \frac{1}{\lambda_{\text{L}}^{\text{prior}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}')} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})) \right) \geq \\
&\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{L}}^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \\
&+ \min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{L}}^{\text{prior}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})) \geq \\
&\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{L}}^{\text{prior}}} \mathbb{E}_q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \\
&+ \min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{L}}^{\text{prior}}} \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \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  $K_2 - \log p(\mathbf{h}_2|\boldsymbol{\lambda}) + \log p(\mathbf{h}_1|\boldsymbol{\lambda})$  получаем разность параметрических сложностей больше нуля.

Докажем критерий 5. Пусть  $\lambda_{\text{Q}}^{\text{likelihood}} = \lambda_{\text{L}}^{\text{prior}} = \lambda_{\text{Q}}^{\text{prior}} > 0$ ,  $\boldsymbol{\lambda}_{\text{Q}}^{\text{struct}} = \mathbf{0}$ . Тогда функции  $L$  и  $Q$  можно записать как:

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

что и требовалось доказать.

Докажем критерий 6. Пусть задан вектор метапараметров  $\boldsymbol{\lambda}$ , удовлетворяющий девятому условию теоремы и  $\boldsymbol{\lambda} = \mathbf{0}$ . Возьмем в качестве  $K_4$  следующее выражение:

$$K_4 = \frac{\max_{\mathbf{h}} Q - \min_{\mathbf{h}} Q}{\max_{\lambda_{\text{struct}}} \lambda_{\text{struct}}}.$$

Пусть вектор метапараметров  $\boldsymbol{\lambda}'$  отличается от  $\boldsymbol{\lambda}$  лишь метапараметром  $\lambda_{\text{struct}}$ . Для обоих векторов метапараметров нижняя задача оптимизации  $L$  совпадает, поэтому выполняется первое условие критерия.

Без ограничения общности предположим, что  $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > 0$  при  $\boldsymbol{\lambda}$ . Также без ограничения общности будем полагать, что множество  $\mathbf{P}$  состоит только из одного распределения на структуре  $\boldsymbol{\Gamma}$ , равного распределению на структуре  $p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})$ .

Положим для  $\boldsymbol{\lambda}'$  параметр  $\lambda_{\text{struct}}$  равным максимальному значению:  $\lambda_{\text{struct}} = \max \lambda'_{\text{struct}}$ . Тогда при  $\boldsymbol{\lambda}'$  неравенство

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

что и требовалось доказать.

Докажем критерий 7. Достаточным условием непрерывности функций  $L, Q$  является непрерывность входящих в нее слагаемых. Т.к. априорные распределения задаются непрерывными функциями плотности  $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}), p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ , и функция плотности  $p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$  распределения структуры  $\mathbf{\Gamma}$  ограничена на компакте, то дивергенция  $D_{KL}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))$  непрерывна по метапараметрам. Т.к. остальные слагаемые функций оптимизаций  $L, Q$  также непрерывны по метапараметрам, то непрерывна и сами функции оптимизации.  $\square$

Метапараметрами данной задачи (??) являются коэффициенты  $\lambda_Q^{\text{prior}}, \lambda_L^{\text{prior}}$ , 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  $\boldsymbol{\lambda}_Q^{\text{struct}}$ .

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  $\boldsymbol{\lambda}_Q^{\text{struct}}$  порождает последовательность задач оптимизаций, схожую с перебором структур. Рассмотрим следующий пример.

**Пример 2.** Рассмотрим вырожденный случай поведения функции  $Q$ , когда  $\lambda_{\text{likelihood}}^Q = \lambda_Q^{\text{prior}} = 0$ . Пусть модель использует один структурный параметр, в качестве априорного распределения на структуре задано распределение Gumbel-Softmax с  $\lambda_{\text{temp}} = 1.0$ . 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. ?? изображены значения функции  $Q$  в зависимости от метапараметров  $\boldsymbol{\lambda}_Q^{\text{struct}}$  и значений гиперпараметра  $\mathbf{s}$  распределения на структуре. Видно, что варьируя коэффициенты метапараметров получается последовательность оптимизаций, схожая с полным перебором структуры.

#### 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

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

эквивалентна оптимизации вариационной оценки обоснованности  $\mathbb{E}_q \log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - D_{KL}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}))$  для произвольной случайной подвыборки  $\hat{\mathbf{y}}, \hat{\mathbf{X}}$  мощности  $\frac{m}{\lambda_{\text{prior}}^L}$  из генеральной совокупности.

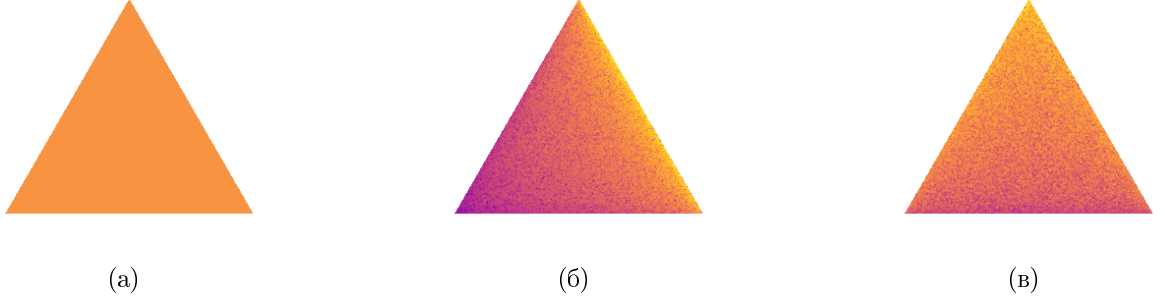


Рис. 1.5. Пример зависимости функции  $Q$  от гиперпараметра  $\mathbf{s}$  при различных значениях метапараметров  $\lambda_Q^{\text{struct}}$ . Темные точки на графике соответствуют наименее предпочтительным значениям гиперпараметра. а)  $\lambda_Q^{\text{struct}} = [0, 0]$ , б)  $\lambda_Q^{\text{struct}}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

*Доказательство.* Рассмотрим величину  $\frac{1}{m}L$ :

$$\frac{1}{m}L = \frac{1}{m}\mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - \frac{\lambda_{\text{prior}}^L}{m} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)). \quad (1.16)$$

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

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

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

$$\begin{aligned} \frac{1}{m_0} \mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - \frac{1}{m_0} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)) &\approx \\ &\approx \mathbb{E}_{y, \mathbf{x}} \mathbb{E}_q \log p(y|\mathbf{x}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - \frac{1}{m_0} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)) = \\ &= \mathbb{E}_{y, \mathbf{x}} \mathbb{E}_q \log p(y|\mathbf{x}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda) - \frac{\lambda_{\text{prior}}^L}{m} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda)). \end{aligned} \quad (1.17)$$

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

Таким образом, для достаточно большого  $m$  и  $\lambda_L^{\text{prior}} > 0$ ,  $\lambda_L^{\text{prior}} \neq 1$  оптимизация параметров и гиперпараметров эквивалентна нахождению оценки обоснованности для выборки другой мощности: чем выше значение  $\lambda_L^{\text{prior}}$ , 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  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^L = \lambda_{\text{prior}}^Q = 1, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Тогда:

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

$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) + \log p(\mathbf{h}|\mathbf{f}) \rightarrow \max_{\mathbf{h}}.$$

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

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

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

*Доказательство.* При  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^L = 1$  как верхняя, так и нижняя задачи оптимизации (??) эквивалентны оптимизации вариационной оценки обоснованности, поэтому первое утверждение выполняется.

Докажем второе утверждение. Рассмотрим логарифм обоснованности модели:

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

Из данного равенства следует:

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

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

Докажем третье утверждение. Т.к. вариационное распределение  $q$  декомпозируется на  $q(\Gamma|\boldsymbol{\theta}_{\Gamma}), q(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ , апостериорное распределение  $p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$  декомпозируется на  $p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda), p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \Gamma, \mathbf{h}, \lambda)$ , поэтому достижимо значение нулевого значения дивергенции:  $D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)) = 0$ . Она представима в следующем виде:

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

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. Калибруя верхнюю оптимизацию;
2. Калибруя нижнюю оптимизацию;
3. Калибруя обе оптимизации.

Последний вариант соответствует теореме о калибровке. Рассмотрим различие варианта 1 и 2 на примере.

**Пример 3.** Пусть задана модель и выборка и мы хотим уменьшить вес априорного распределения. В случае, если мы калибруем нижнюю оптимизацию ( $\rightarrow 0$ ), на первом уровне задача совпадает с задачей поиска наиболее правдоподобных параметров, при этом на верхнем уровне мы ищем те параметры, которые отвечают наилучшим с точки зрения обоснованности.

Если мы калибруем верхнюю оптимизацию или обе оптимизации, то это приведет к поиску наиболее правдоподобных параметров и гиперпараметров.

Таким образом, основная разница между калибровкой верхней и нижней оптимизации заключается в следующем: при калибровке нижнего уровня мы получаем модель, соответствующую критерию максимального правдоподобия. В случае калибровки верхнего уровня мы получаем модель с параметрами, полученными в соответствии с методом максимальной обоснованности, но при минимально возможной регуляризации априорным распределением.

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

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

*Доказательство.* Параметры вариационного распределения  $q$  не зависят от слагаемых вида  $\log p(\mathbf{h} | \mathbf{f})$  и  $D_{KL}(\Gamma | p'), p' \in \mathbf{P}$ , поэтому нижняя задача оптимизации:

$$\begin{aligned}
&\log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - \\
&\quad - \lambda_L^{\text{prior}} D_{KL}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})) \rightarrow \max_{\boldsymbol{\theta}}
\end{aligned}$$

эквивалентна следующей задаче:

$$\log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$

$$\begin{aligned}
& -\lambda_L^{\text{prior}} D_{KL}(q(\mathbf{w}, \Gamma | \boldsymbol{\theta}) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})) \rightarrow \max_{\boldsymbol{\theta}} \\
& - \sum_{p' \in \mathbf{P}, \lambda \in \boldsymbol{\lambda}_Q^{\text{struct}}} \lambda D_{KL}(\Gamma | p') + \log p(\mathbf{h} | \mathbf{f}) \rightarrow \max_{\boldsymbol{\theta}}
\end{aligned}$$

для любого вектора  $\boldsymbol{\lambda}_Q^{\text{struct}}$ . Т.к. выполнено равенство  $\frac{\lambda_Q^{\text{prior}}}{\lambda_Q^{\text{likelihood}}} = \lambda_{\text{prior}}^L$ , то нижняя задача оптимизации эквивалентна следующей задаче:

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

а значит верхняя и нижняя задачи совпадают:

$$\mathbf{h} = \arg \max_{\mathbf{h}'} Q(\mathbf{h}, \boldsymbol{\theta}^*(\mathbf{h}')),$$

где

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

Из свойства

$$\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h}) = \max_{\boldsymbol{\theta}, \mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$$

следует доказательство теоремы. □

Для вычисления приближенного значения функций  $Q$  и  $L$  предлагается использовать приближение методом Монте-Карло с порождением  $R$  реализаций величин  $\mathbf{w}, \Gamma$ :

$$\begin{aligned}
\mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_1 \lambda_{\text{temp}}, \mathbf{f}) & \approx \sum_{r=1}^R \log p(\mathbf{y} | \boldsymbol{\mu} + \boldsymbol{\alpha}_q \circ \hat{\epsilon}_r, \hat{\Gamma}_r, \mathbf{X}), \\
D_{KL}(q_{\Gamma}(\Gamma | \boldsymbol{\theta}_{\Gamma}) || p(\Gamma | \mathbf{h}, \boldsymbol{\lambda})) & \approx \sum_{r=1}^R \left( \log q_{\Gamma}(\hat{\Gamma}_r | \boldsymbol{\theta}_{\Gamma}) - p(\hat{\Gamma}_r | \mathbf{h}, \boldsymbol{\lambda}) \right), \\
D_{KL}(q_{\mathbf{w}}(\mathbf{w} | \boldsymbol{\theta}_{\mathbf{w}}, \Gamma) || p(\mathbf{w} | \Gamma, \mathbf{h})) & = \sum_{(j,k) \in E} \sum_{l=1}^{K^{j,k}} D_{KL} \left( q_{\mathbf{w}}(\mathbf{w}_l^{j,k} | \boldsymbol{\theta}_{\mathbf{w}}, \gamma_l^{j,k}) || p(\mathbf{w}_l^{j,k} | \gamma_l^{j,k}, \mathbf{h}) \right) \approx \\
& \approx - \sum_{(j,k) \in E} \sum_{l=1}^{K_{j,k}} \sum_{r=1}^R \frac{1}{2} \left( (\hat{\gamma}_r^{j,k}[l])^{-1} \text{tr}((\mathbf{A}_l^{j,k})_q (\mathbf{A}_l^{j,k})^{-1}) + (\boldsymbol{\mu}_l^{j,k})^{\top} \hat{\gamma}_r^{j,k}[l]^{-1} (\mathbf{A}_l^{j,k})^{-1} \boldsymbol{\mu}_l^{j,k} - \right.
\end{aligned}$$



$$-|\mathbf{w}_l^{j,k}| + \log \frac{|\hat{\gamma}^{j,k}[l]_r \mathbf{A}_l^{j,k}|}{|(\mathbf{A}_l^{j,k})_q|},$$

где  $R$  — количество реализаций случайных величин, по котором вычисляется значения вариационной оценки обоснованности,  $\hat{\epsilon}_r \sim \mathcal{N}(0, 1)$ ,  $\hat{\mathbf{\Gamma}}_r = [\hat{\gamma}_r^{j,k}, (j, k) \in E]$  — реализация случайной величины, соответствующей структуре  $\mathbf{\Gamma}$ .

Для решения двухуровневой задачи предлагается использовать градиентные методы.

**Теорема 9.** Пусть  $T$  — оператор градиентного спуска. Пусть  $Q, L$  — локально выпуклы и непрерывны в некоторой области  $U_W \times U_\Gamma \times U_H \times U_\lambda \subset \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \times \mathbb{A}$ , при этом  $U_H \times U_\lambda$  — компакт. Тогда решение задачи градиентной оптимизации

$$\mathbf{h}^* = T^\eta(Q, \mathbf{h}, T^\eta(L, \boldsymbol{\theta}_0, \mathbf{h}))$$

стремится к локальному минимуму  $\mathbf{h}^* \in U$  исходной задачи оптимизации при  $\eta \rightarrow \infty$ ,  $\mathbf{h}^*$  является непрерывной функцией по метопараметрам модели.

*Доказательство.* TODO □

Следующие теоремы посвящены асимптотическим свойствам представленной обобщающей задачи.

**Теорема 10.** Пусть  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^L > 0, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Тогда предел оптимизации

$$\lim_{\lambda_{\text{prior}}^Q \rightarrow \infty} \lim_{\eta \rightarrow \infty} T^\eta(Q, \mathbf{h}, T^\eta(L, \boldsymbol{\theta}_0, \mathbf{h}))$$

доставляет минимум параметрической сложности.

*Доказательство.* TODO □

**Теорема 11.** Пусть  $\lambda_{\text{likelihood}}^L = 1, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Пусть  $\mathbf{f}_1, \mathbf{f}_2$  — результаты градиентной оптимизации при разных значениях гиперпараметров  $\lambda_{\text{prior}}^{Q,1}, \lambda_{\text{prior}}^{Q,2}, \lambda_{\text{prior}}^{Q,1} < \lambda_{\text{prior}}^{Q,2}$ , полученных при начальном значении вариационных параметров  $\boldsymbol{\theta}_0$  и гиперпараметров  $\mathbf{h}_0$ . Пусть  $\boldsymbol{\theta}_0, \mathbf{h}_0$  принадлежат области  $U$ , в которой соответствующие функции  $L$  и  $Q$  являются локально-выпуклыми. Тогда:

$$C_p(\mathbf{f}_1) - C_p(\mathbf{f}_2) \geq \lambda_{\text{prior}}^L (\lambda_{\text{prior}}^L - \lambda_{\text{prior}}^{Q,1}) \sup_{\boldsymbol{\theta}, \mathbf{h} \in U} |\nabla_{\boldsymbol{\theta}, \mathbf{h}}^2 D_{KL}(q|p) (\nabla_{\boldsymbol{\theta}}^2 L)^{-1} \nabla_{\boldsymbol{\theta}} D_{KL}(q|p)|.$$

*Доказательство.* TODO □

TODO: выводы **Эксперимент: пример 1**

**Эксперимент: пример 2**

## Список основных обозначений

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

$\lambda$  — вектор метапараметров

$\lambda$  — пространство метапараметров

$U_\lambda$  — область определения метапараметров

$p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda)$  — 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

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

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

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

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

$q_{\Gamma}(\Gamma | \theta_{\Gamma})$  — вариационное распределение параметров модели

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

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

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

$\mathfrak{Q}$  — семейство вариационные распределений

$S$  — энтропия распределения

$M$  — множество моделей без общей параметризации