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:

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

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

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

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

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

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

где $\mathbf{s} \in (0, \infty)^K$ — гиперпараметр, отвечающий за смещенность плотности распределения относительно точек симплекса на K вершинах, λ_{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 $\boldsymbol{\gamma}$ 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 $\hat{\mathbf{g}} \sim -\log\bigl(-\log \mathcal{U}(0,1)^K\bigr)$.

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

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

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

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

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

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

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

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:

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

Рассмотрим два случая: когда вектор $\boldsymbol{\gamma}$ лежит не в центре симплекса, и когда $\boldsymbol{\gamma}$ 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 $\boldsymbol{\gamma}$. Для $l = l'$ предел выражения (1.3) при λ_{temp} стремится к бесконечности. Для $l \neq l'$ предел выражения (1.3) при λ_{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 λ^{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 ψ, распределенную по распределению q с параметрами $\boldsymbol{\theta}_\psi$ назовем представление величины с помощью другой случайной величины, имеющей распределение, не зависящее от параметров $\boldsymbol{\theta}$:

$$\psi \sim q \iff \hat{\psi} \sim g(\boldsymbol{\varepsilon}, \boldsymbol{\theta}_\psi),$$

где $\boldsymbol{\varepsilon}$ — случайная величина, чье распределение не зависит от параметров $\boldsymbol{\theta}_\psi$, g — некоторая детерминированная функция, $\hat{\psi}$ — реализация случайной величины ψ.

Идею репараметризации поясним на следующем примере.

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

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \int_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) q(\mathbf{w}) d\mathbf{w}.$$

Продифференцируем данное выражение по параметрам $\boldsymbol{\theta}$ вариационного распределения q:

$$\nabla_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \int_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \nabla_{\boldsymbol{\theta}} q(\mathbf{w}) d\mathbf{w}.$$

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

$$\mathbf{w} \sim q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \iff \hat{\mathbf{w}} \sim g(\boldsymbol{\varepsilon}, \boldsymbol{\theta}_{\mathbf{w}}).$$

Тогда справедливо следующее выражение:

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

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

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 [**?**].

(а) (б) (в) (г)

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

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 гиперпарамета $\mathbf{A}_l^{j,k}$ 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 λ_1, λ_2 получается более сильная или более слабая регуляризация [?]. Пример распределений $\text{inv-gamma}(\lambda_1, \lambda_2)$ для разных значений метапараметров λ_1, λ_2 изображен на Рис. 1.2. Оптимизации без регуляризации соответствует случай предельного распределения $\lim_{\lambda_1, \lambda_2 \to 0} \text{inv-gamma}(\lambda_1, \lambda_2)$.

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

1. Параметры $\mathbf{w}$ модели, распределенные нормально.
2. Структура модели $\boldsymbol{\Gamma}$, содержащая все структурные параметры $\{\boldsymbol{\gamma}^{j,k}, (j,k) \in E\}$ 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.

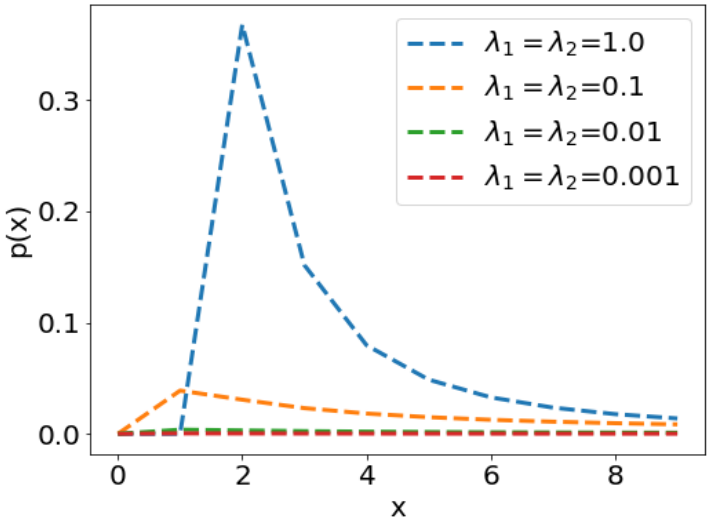


Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

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:

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

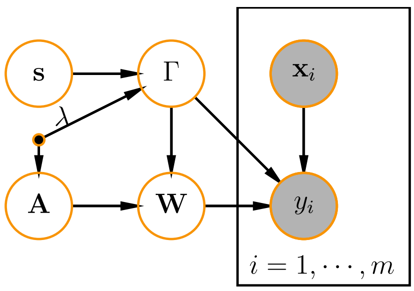


Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

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

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

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

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

Для вычисления обоснованности

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

из (1.4) предлагается использовать вариационную оценку обоснованности.

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

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

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

Тогда справедлива следующая оценка:

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

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

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

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

Доказательство. Рассмотрим обоснованность:

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

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

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

Используя неравенство Йенсена получим

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

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

Декомпозируем распределение q по свойству условной дивергенции:

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

$$= D_{\mathrm{KL}}\left(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})|p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\right) + D_{\mathrm{KL}}\left(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\Gamma})|p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\right).$$

□

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

$$q_{\mathbf{w}} = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q),$$

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

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

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

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

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

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

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

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 [?]:

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

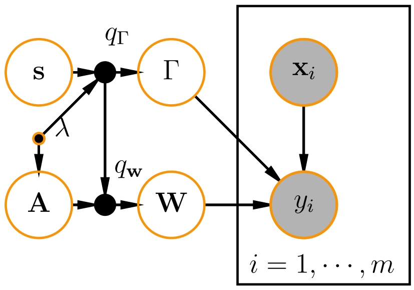


Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

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

Обобщим понятие относительной вариационной плотности на случай произвольных непрерывных распределений.

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

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

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

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

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

Теорема 2. Пусть

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

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

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
$$\mathbf{h} = \arg\min_{\mathbf{h}\in U_{\mathbf{h}}} \mathrm{D}_{\mathrm{KL}}\left(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h})\right) \qquad (1.6)$$
единственно для любого $\boldsymbol{\theta} \in U_{\boldsymbol{theta}}$.
5. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по маргинальным распределениям $q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda})$:
$$\mathsf{E}_q \mathbf{w}^2 = \mathsf{E}_{q_{\mathbf{\Gamma}}} \mathsf{E}_{q_{\mathbf{w}}} \mathbf{w}^2 < \infty;$$
$$\mathsf{E}_{p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})} \mathbf{w}^2 = \mathsf{E}_{p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})} \mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})} \mathbf{w}^2 < \infty,$$
где
$$p(\mathbf{w}|\mathbf{h}, \boldsymbol{\lambda}) = \int_{\mathbf{\Gamma}} p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}), \quad q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) = \int_{\mathbf{\Gamma}} q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}).$$
6. мода и матожидание вариационного распределения $q_{\mathbf{w}}$ и априорного распределения $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ совпадают:
$$\mathrm{mode}\ p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})} \mathbf{w};$$
$$\mathrm{mode}\ q(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) = \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w};$$
7. задана бесконечная последовательность векторов вариационных параметров $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_i \in U_{\boldsymbol{\theta}}$, такая что $\lim_{i\to\infty} C_p(\boldsymbol{\theta_i}|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$.

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

Доказательство. Воспользуемся неравенством Пинскера:
$$||F_q(\boldsymbol{\theta}) - F_p(\mathbf{h})||_{\mathrm{TV}}, \leq \sqrt{2\mathrm{D}_{\mathrm{KL}}\left(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h})\right)},$$
где $||\cdot||_{\mathrm{TV}}$ — расстояние по вариации, F_q, F_p — функции распределения $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ и $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$. Отсюда $\lim_{i\to\infty} ||F_q(\boldsymbol{\theta}) - F_p(\mathbf{h})||_{\mathrm{TV}} = 0$. По теореме Шеффе данное выражение можно переписать как:
$$\lim \frac{1}{2} \iint_{\mathbf{w},\mathbf{\Gamma}} |p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) - q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})| d\mathbf{\Gamma} d\mathbf{w} = 0.$$

Рассмотрим разность усредненных мод:
$$|\mathsf{E}_{q_{\mathbf{\Gamma}}} \mathrm{mode}\ q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \mathbf{\Gamma}) - \mathsf{E}_{p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})} \mathrm{mode}\ p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})| =$$
$$= |\mathsf{E}_{q_{\mathbf{\Gamma}}} \mathsf{E}_{q_{\mathbf{w}}} \mathbf{w} - \mathsf{E}_{p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})} \mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})} \mathbf{w}| =$$

$$= |\mathsf{E}_{q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})}\mathbf{w} - \mathsf{E}_{p(\mathbf{w}|\mathbf{h})}\mathbf{w}|.$$

Т.к. вторые моменты $\mathsf{E}_{q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})}\mathbf{w}^2, \mathsf{E}_{p(\mathbf{w}|\mathbf{h})}\mathbf{w}^2$ 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 ε существует число t_0, такое что для любого $t \geq t_0$ справедливо выражение:

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

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

$$|\mathsf{E}_{q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})}\mathbf{w} - \mathsf{E}_{p(\mathbf{w}|\mathbf{h})}\mathbf{w}| \leq \int_\mathbf{w} |\mathbf{w} - \boldsymbol{\nu}(t)||p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w}+$$

$$+\int_\mathbf{w} |\boldsymbol{\nu}(t)||p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w} \leq$$

$$\leq \int_\mathbf{w} |\mathbf{w} - \boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{h})d\mathbf{w} + \int_\mathbf{w} |\mathbf{w} - \boldsymbol{\nu}(t)|q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w}+$$

$$+\int_\mathbf{w} |\boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w}.$$

Обозначим за $\mathbf{h}_i$ — решение задачи (1.6) для вектора $\boldsymbol{\theta}_i$. Т.к. $|\boldsymbol{\nu}(t)|$ — ограничена, то

$$\int_\mathbf{w} |\boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{h}_i) - q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w} \leq t\int_\mathbf{w} |p(\mathbf{w}|\mathbf{h}) - q(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})|d\mathbf{w} =$$

$$= t\iint_{\mathbf{w},\boldsymbol{\Gamma}} |p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_i, \boldsymbol{\lambda}) - q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|d\boldsymbol{\Gamma}d\mathbf{w} \to_{i\to\infty} 0.$$

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

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

Рассмотрим предел интегралов:

$$0 = \lim_{i\to\infty} \mathsf{E}_{q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma},i})}\mathsf{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w},i})}\mathbf{w} - \mathbf{M}| = \lim_{i\to\infty} |\int_{\boldsymbol{\Gamma}}\int_\mathbf{w} |(\mathbf{w} - \mathbf{M})|q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma},i})q(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w},i})d\mathbf{w}d\boldsymbol{\Gamma}.$$

Пусть $\boldsymbol{\theta}_j$ — некоторые вариационные параметры, принадлежащие компакту $U_{\boldsymbol{\theta}}$. Т.к. q — непрерывно-дифференцируемая функция, то она является липшецевой. Пусть C_L — его констанста Липшица. Тогда:

$$\lim_{i\to\infty} |\int_{\mathbf{\Gamma}} q(\int_{\mathbf{w}} \mathbf{w}d\mathbf{w}) - q(\mathbf{M})q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma},i})d\mathbf{\Gamma} \leq$$

$$\leq L\lim_{i\to\infty} |\int_{\mathbf{\Gamma}}\int_{\mathbf{w}} (\mathbf{w}-\mathbf{M})q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma},i})q(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w},i})d\mathbf{w}d\mathbf{\Gamma} = 0.$$

Тогда

$$\lim_{i\to\infty} |\int_{\mathbf{\Gamma}} \frac{q(\int_{\mathbf{w}} \mathbf{w}}{q(\mathbf{M}} d\mathbf{w}) - 1| = 0.$$

Устремим $j \to \infty$. Т.к. выражение — непрерывно-дифференцируемое, то возможно перестановка пределов:

$$\lim_{j\to\infty}\lim_{i\to\infty} |\int_{\mathbf{\Gamma}} \frac{q(\int_{\mathbf{w}} \mathbf{w}}{q(\mathbf{M}} d\mathbf{w}) - 1| = 0.$$

□

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

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

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, испольюзующих 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:

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

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

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

Данное обобщение (1.7) эквивалентно критерию правдоподобия при выборе в качестве q эмпирического распределения парамтетров и структуры. Метод не предполагает оптимизации гиперпараметров $\mathbf{h}$. Для формального соответствия данной задачи задаче выбора модели (**??**), т.е. двухуровневой задачи оптимизации, положим $L = Q$:

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

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

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

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

Т.к. в рамках данной задачи (1.8) не предполагается оптимизации гиперпараметров $\mathbf{h}$, положим параметры распределения $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ фиксированными:

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

3. Перебор структуры:

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

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

4. Критерий Акаике:

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

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:

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

где

$$\mathbf{h} = \mathop{\arg\min}_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\mathrm{KL}}(q(\mathbf{w}, \mathbf{\Gamma})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}', \boldsymbol{\lambda}), \tag{1.11}$$

λ_{prune} — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение (1.10) применимо только в случае, если выражение (1.11) определено однозначно, т.е. существует единственный вектор гиперпараметров на $U_{\mathbf{h}}$, доставляющий минимум дивергенции $D_{\mathrm{KL}}(q(\mathbf{w}, \mathbf{\Gamma})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.

5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

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

метапараметр λ_{prune} определен аналогично (1.11).

6. Метод вариационной оценки обоснованности:

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - D_{\mathrm{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{\Gamma}, \mathbf{w}|\mathbf{h}, \boldsymbol{\lambda})) + \log p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta}}, \tag{1.13}$$

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

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

7. Валидация на отложенной выборке:

$$L = \mathsf{E}_q \log p(\mathbf{y}_{\mathrm{train}}, \mathbf{w}, \mathbf{\Gamma}|\mathbf{X}_{\mathrm{train}}, \mathbf{h}, \boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta}}, \tag{1.14}$$

$$Q = \mathsf{E}_q \log p(\mathbf{y}_{\mathrm{test}}|\mathbf{X}_{\mathrm{test}}, \mathbf{w}, \mathbf{\Gamma}) \to \max_{\mathbf{h}},$$

где $(\mathbf{X}_{\mathrm{train}}, \mathbf{y}_{\mathrm{train}}), (\mathbf{X}_{\mathrm{test}}, \mathbf{y}_{\mathrm{test}})$ — 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 значинях метапараметров, будет называть *обобщающей.*

Определение 4. Двухуровневую задачу оптимизации будем называть *обобщающей* на компакте $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \mathbb{\Theta} \times \mathbb{H} \times \mathbb{\Lambda}$, если она удовлетворяет следующим критериям.

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

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

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

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

   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

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

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

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

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

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

   параметрическая сложность первой модели меньше, чем второй:

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

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

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

для всех $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}$.
6. Критерий перебора оптимальных структур: существует набор метапараметров $\boldsymbol{\lambda}$ и константа

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

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

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

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

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 нерперывности: функции L и Q непрерывны по метапараметрам $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.

Первый критерий является техническим и используется для исключения из рассмотрения вырожденных задач оптимизации. Второй критерий говорит о том, что решение первого и второго уровня должны быть согласованы и определены однозначно. Критерии 3-5 определяют возможные критерии оптимизации, которые должны приближаться обобщающей задачей. Критерий 6 говорит о возможности перехода между различными структурами модели. Данный критерий говорит о том, что мы можем перейти от одного набора гиперпараметров $\mathbf{h}_1$ к другим $\mathbf{h}_2$, если они соответствуют локальным оптимумам задачи оптимизации, и дивергенция соответствующих априорных распределений на структурах $p(\boldsymbol{\Gamma}|\mathbf{s}, \boldsymbol{\lambda})$ значимо высока. При этом соответствующие вариационные распределения $q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ могут оказаться достаточно близки. Возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения $q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ 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 (1.7),(1.8),(1.9),(1.10),(1.12),(1.14) не являются обобщающими.

Доказательство. Задачи (1.7),(1.8),(1.9),(1.10),(1.12) не имеют гиперпараметров $\mathbf{h}$, подлежащих оптимизации, поэтому не могут оптимизировать вариационную оценку.

При использовании валидации на отложенной выборки (1.14) в функцию валидации Q не входит ни один метапараметр, поэтому критерий перебора структур 6 для нее также не выполняется.

□

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

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

Тогда задача (1.13) 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 $\boldsymbol{\lambda}$. По условию критерия соответствтие $\boldsymbol{\theta}^*(\mathbf{h}_1)$ и $\boldsymbol{\theta}^*(\mathbf{h}_2)$ должны сохраняться, т.е. для некоторого $\boldsymbol{\lambda}'$ решение нижней задачи оптимизации $\boldsymbol{\theta}^*(\mathbf{h}_1)$ должно совпадать с решением $\boldsymbol{\theta}^*(\mathbf{h}_1)$ при метапараметрах $\boldsymbol{\lambda}$. Тогда

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

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

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\boldsymbol{\theta}} \text{D}_{\text{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\theta}} \text{D}_{\text{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}')),$$

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

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

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

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

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

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

Метапараметры λ_1, λ_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 λ_1, λ_2. Приходим к противоречию: значение Q не меняется при изменении метапараметров $\boldsymbol{\lambda}$.

$\square$

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q = \tag{1.15}$$

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

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

$$- \sum_{p' \in \mathbf{P}, \lambda \in \boldsymbol{\lambda}^{\mathrm{struct}}_{\mathrm{Q}}} \lambda \mathrm{D}_{KL}(\boldsymbol{\Gamma}|p') + \log p(\mathbf{h}|\mathbf{f}),$$

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

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

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

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 $q = q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ является абсолютно непрерывным, плотность которого непрерывна по метапараметрам $\boldsymbol{\lambda}$;
3) задан компакт $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \mathbb{\Theta} \times \mathbb{H} \times \mathbb{\Lambda}$, где параметры распределений $\mathbf{P} \in \mathbb{\Lambda}$, область $U_{\boldsymbol{\theta}}$ декомпозируется на две области $U_{\boldsymbol{\theta}} = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}}$;
4) область определения каждого параметра $w \in \mathbf{w}$, гиперпараметра $h \in \mathbf{h}$ и метапараметра $\lambda \in \boldsymbol{\lambda}$ не является является пустым и не является точкой;
5) для каждого значения гиперпараметров $\mathbf{h}$ оптимальное решение нижней задачи оптимизации $\boldsymbol{\theta}^*$ определено однозначно при любых значениях метапараметров $\boldsymbol{\lambda} \in U_{\lambda}$;
6) область значений метапараметров $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{prior}}_{\mathrm{Q}}, \boldsymbol{\lambda}^{\mathrm{struct}}_{\mathrm{Q}}, \lambda^{\mathrm{prior}}_{\mathrm{L}}$ включает отрезок от нуля до единицы;

7) существует значение метапараметров $\lambda_1, \lambda_2, \lambda_{\text{likelihood}}^{\text{Q}}$, такое что

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

при $\boldsymbol{\lambda}_Q^{\text{struct}} = \mathbf{0}, \lambda_Q^{\text{prior}} = 0.$

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

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

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

$$- \min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\text{L}}^{\text{prior}}} \mathsf{E}_q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) < \max_{\boldsymbol{\theta},\mathbf{h}} D_{\text{KL}} - \min_{\boldsymbol{\theta},\mathbf{h}} D_{\text{KL}}.$$

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

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

при $\boldsymbol{\lambda}_{\text{Q}}^{\text{struct}} = 0.$

Тогда задача (1.15) является обобщающей на U.

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

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

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

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

$$Q(\mathbf{h}_1) - Q(\mathbf{h}_2) = \lambda_{\text{Q}}^{\text{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})-$$

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

Отсюда следует выполнение критерия 3:

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

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

$$K_2 = \max_{\mathbf{h}} \log\, p(\mathbf{h}|\boldsymbol{\lambda}) - \min_{\mathbf{h}} \log\, p(\mathbf{h}|\boldsymbol{\lambda}) + \max_{\mathbf{h}} \min_{\boldsymbol{\theta}} D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))-$$

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

$$-\min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\mathsf{E}_q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})\mathrm{log}p(\mathbf{y}|\mathbf{X},\mathbf{w}).$$

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:

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

$$-\min_{\mathbf{h}} D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})) \geq$$

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

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

$$= Q(\mathbf{h}_1) - Q(\mathbf{h}_2) - \log\, p(\mathbf{h}_1|\boldsymbol{\lambda}) + \log\, p(\mathbf{h}_2|\boldsymbol{\lambda}) +$$

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

$$> K_2 - \log\, p(\mathbf{h}_1|\boldsymbol{\lambda}) + \log\, p(\mathbf{h}_2|\boldsymbol{\lambda}) + \min_{\boldsymbol{\theta},\mathbf{h}} D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}))$$

$$- D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda})).$$

Рассмотрим разность:

$$\min_{\boldsymbol{\theta},\mathbf{h}} D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})) - D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda})) =$$

$$= \min_{\boldsymbol{\theta},\mathbf{h}} D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})) - \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\mathsf{E}_q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)\mathrm{log}p(\mathbf{y}|\mathbf{X},\mathbf{w}) +$$

$$+\max_{\boldsymbol{\theta}}(\frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})}\mathrm{log}p(\mathbf{y}|\mathbf{X},\mathbf{w}) - D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda}))) \geq$$

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

$$+\max_{\boldsymbol{\theta}}(\min_{\boldsymbol{\theta}'} \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}')}\mathrm{log}p(\mathbf{y}|\mathbf{X},\mathbf{w}) - D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda}))) \geq$$

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

$$+\min_{\boldsymbol{\theta}} \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})}\mathrm{log}p(\mathbf{y}|\mathbf{X},\mathbf{w}) - \min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda}))) \geq$$

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

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

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

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

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:

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

Пусть вектор метапараметров $\boldsymbol{\lambda}'$ отличается от $\boldsymbol{\lambda}$ лишь метапараметром $\boldsymbol{\lambda}_{\mathrm{struct}}$. 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 полгаать, что множестве $\mathbf{P}$ 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 $\lambda_{\mathrm{struct}}$ равным максимальному значению: $\lambda_{\mathrm{struct}} = \max \lambda'_{\mathrm{struct}}$. Тогда при $\boldsymbol{\lambda}'$ неравенство

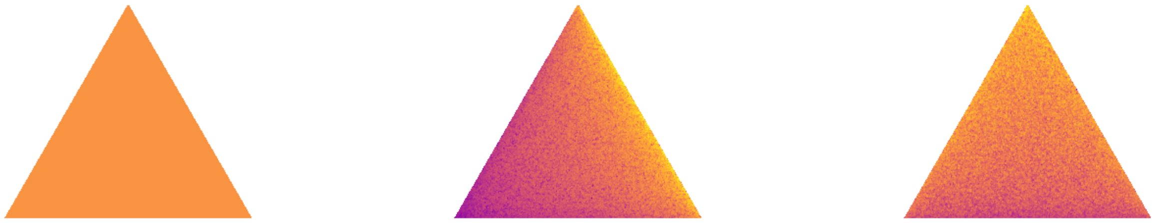
$$Q(\mathbf{h}_1|\boldsymbol{\lambda}') - Q(\mathbf{h}_2|\boldsymbol{\lambda}') = Q(\mathbf{h}_1|\boldsymbol{\lambda}) - Q(\mathbf{h}_2|\boldsymbol{\lambda}) + \lambda'_{\mathrm{struct}} D_{\mathrm{KL}}(p(\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}')|p(\boldsymbol{\Gamma}|\mathbf{h}_1,\boldsymbol{\lambda}')) >$$

$$> Q(\mathbf{h}_1|\boldsymbol{\lambda}) - Q(\mathbf{h}_2|\boldsymbol{\lambda}) + \lambda'_{\mathrm{struct}} K_4 = Q(\mathbf{h}_1|\boldsymbol{\lambda}) - Q(\mathbf{h}_2 + |\boldsymbol{\lambda}) + \max_{\mathbf{h}} Q - \min_{\mathbf{h}} Q = 0,$$

что и требовалось доказать.

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

Метапараметрами данной задачи (1.15) являются коэффициенты $\lambda_{\mathrm{Q}}^{\mathrm{prior}}, \lambda_{\mathrm{L}}^{\mathrm{prior}}$, отвечающие за регуляризацию верхней и нижней задачи оптимизации, коэффициент $\lambda_{\mathrm{likelihood}}^{\mathrm{Q}}$ отвечает за максимизацию правдоподобия, а также параметры распрделений $\mathbf{P}$ и вектор коэффициентов перед ними $\boldsymbol{\lambda}_{\mathrm{Q}}^{\mathrm{struct}}$.



(а) (б) (в)

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

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

Пример 2. Рассмотрим вырожденный случай поведения функции Q, когда $\lambda_{\text{likelihood}}^{\text{Q}} = \lambda_{\text{Q}}^{\text{prior}} = 0$. 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:

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

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 $\boldsymbol{\lambda}_{\text{Q}}^{\text{struct}}$ 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 ассимптотического 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

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

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

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

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

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

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

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

$$\frac{1}{m_0}\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda}) - \frac{1}{m_0}D_{\text{KL}}(q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})) \approx \quad (1.17)$$

$$\approx \mathsf{E}_{y,\mathbf{x}}\mathsf{E}_q\text{log}p(y|\mathbf{x},\mathbf{w},\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda}) - \frac{1}{m_0}D_{\text{KL}}(q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})) =$$

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

Таким образом, задачи оптимизации функций (1.16),(1.17) совпадают, что и требовалось доказать. □

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 λ_L^{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 $\lambda^Q_{\text{likelihood}} = \lambda^L_{\text{prior}} = \lambda^Q_{\text{prior}} = 1, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Тогда:

1. Задача оптимизации (1.15) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:
$$\text{log}\hat{p}(\mathbf{y}|\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f}) + \text{log}p(\mathbf{h}|\mathbf{f}) \to \max_{\mathbf{h}}.$$

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

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

Доказательство. При $\lambda^{Q}_{\text{likelihood}} = \lambda^{L}_{\text{prior}} = 1$ как верхняя, так и нижняя задачи оптимизации (1.15) эквивалентны оптимизации вариационной оценки обоснованности, поэтому первое утверждение выполняется.

Докажем второе утвреждение. Рассмотрим логарифм обоснованности модели:

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

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

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

Из данного равенства следует:

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

$$\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})),$$

где правая часть равенства соответствует вариационной оценки обоснованности. Выражение $\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ 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 декомпозируется на $q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}), q(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$, апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}))$ декомпозируется на $p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$, $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$, 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:

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

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

Отсюда следует что соотстветсвующие вариационные и апостериорные распределения совпадают. □

Теорема 8. Пусть $\frac{\lambda^{Q}_{\text{prior}}}{\lambda^{Q}_{\text{likelihood}}} = \lambda^{L}_{\text{prior}}$. Тогда задачи оптимизации (1.15) представима в виде одноуровневой задачи оптимизации:

$$= \lambda^{\text{Q}}_{\text{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$

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

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

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

$$\log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$
$$-\lambda_{\text{L}}^{\text{prior}} \mathrm{D}_{KL}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big) \to \max_{\boldsymbol{\theta}}$$

эквивалентна следующей задаче:

$$\log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$
$$-\lambda_{\text{L}}^{\text{prior}} \mathrm{D}_{KL}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big) \to \max_{\boldsymbol{\theta}}$$
$$-\sum_{p' \in \mathbf{P}, \lambda \in \boldsymbol{\lambda}_{\text{Q}}^{\text{struct}}} \lambda \mathrm{D}_{KL}(\mathbf{\Gamma}|p') + \log p(\mathbf{h}|\mathbf{f}) \to \max_{\boldsymbol{\theta}}$$

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

$$= \lambda_{\text{likelihood}}^{\text{Q}} \mathsf{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) -$$
$$-\lambda_{\text{Q}}^{\text{prior}} \mathrm{D}_{KL}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})\big) -$$
$$-\sum_{p' \in \mathbf{P}, \lambda \in \boldsymbol{\lambda}_{\text{Q}}^{\text{struct}}} \lambda \mathrm{D}_{KL}(\mathbf{\Gamma}|p') + \log p(\mathbf{h}|\mathbf{f}) \to \max_{\boldsymbol{\theta}},$$

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 $\mathbf{w}, \mathbf{\Gamma}$:

$$\mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1 \lambda_{\text{temp}}, \mathbf{f}) \approx \sum_{r=1}^{R} \log p(\mathbf{y}|\boldsymbol{\mu} + \boldsymbol{\alpha}_q \circ \hat{\epsilon}_r, \hat{\mathbf{\Gamma}}_r, \mathbf{X}),$$

$$D_{\mathrm{KL}}\left(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})|p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\right) \approx \sum_{r=1}^{R}\Big(\log q_{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}_r|\boldsymbol{\theta}_{\mathbf{\Gamma}})) - p(\hat{\mathbf{\Gamma}}|\mathbf{h},\boldsymbol{\lambda})\Big),$$

$$D_{\mathrm{KL}}\left(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}},\mathbf{\Gamma})|p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h})\right) = \sum_{(j,k)\in E}\sum_{l=1}^{K^{j,k}} D_{\mathrm{KL}}\Big(q_{\mathbf{w}}(\mathbf{w}_l^{j,k}|\boldsymbol{\theta}_{\mathbf{w}},\gamma_l^{j,k})|p(\mathbf{w}_l^{j,k}|\gamma_l^{j,k},\mathbf{h})\Big) \approx$$

$$\approx -\sum_{(j,k)\in E}\sum_{l=1}^{K_{j,k}}\sum_{r=1}^{R}\frac{1}{2}\left((\hat{\gamma}_r^{j,k}[l]\right)^{-1}\mathrm{tr}((\mathbf{A}_l^{j,k})_q(\mathbf{A}_l^{j,k})^{-1}) + (\boldsymbol{\mu}_l^{j,k})^{\mathsf{T}}\hat{\gamma}_r^{j,k}[l]^{-1}(\mathbf{A}_l^{j,k})^{-1}\boldsymbol{\mu}_{\boldsymbol{l}}^{\boldsymbol{j,k}} -$$

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

где R — количество реализаций случайных величин, по котором вычисляется значения вариационной оценки обоснованности, $\hat{\epsilon}_r \sim \mathcal{N}(0,1)$, $\hat{\mathbf{\Gamma}}_r = [\hat{\boldsymbol{\gamma}}_r^{j,k}, (j,k) \in E]$ — 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 $U_W \times U_\Gamma \times U_H \times U_\lambda \subset \mathbb{W}\times\mathbb{\Gamma}\times\mathbb{H}\times\mathbb{\Lambda}$, при этом $U_H \times U_\lambda$ — компакт. Тогда решение задачи градиентной оптимизации

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

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

Доказательство. TODO □

Следующие теоремы посвящены ассимптотическим свойствам представленной обобщающей задачи.

Теорема 10. Пусть $\lambda^Q_{\mathrm{likelihood}} = \lambda^L_{\mathrm{prior}} > 0, \boldsymbol{\lambda}^Q_{\mathrm{struct}} = \mathbf{0}$. Тогда предел оптимизации

$$\lim_{\lambda^Q_{\mathrm{prior}}\to\infty}\lim_{\eta\to\infty} T^\eta\big(Q,\mathbf{h},T^\eta(L,\boldsymbol{\theta}_0,\mathbf{h})\big)$$

доставляет минимум параметрической сложности.

Доказательство. TODO □

Теорема 11. Пусть $\lambda^L_{\mathrm{likelihood}} = 1, \boldsymbol{\lambda}^Q_{\mathrm{struct}} = \mathbf{0}$. Пусть $\mathbf{f}_1, \mathbf{f}_2$ — результаты градиентной оптимизации при разных значениях гиперпараметров $\lambda^{Q,1}_{\mathrm{prior}}, \lambda^{Q,2}_{\mathrm{prior}}, \lambda^{Q,1}_{\mathrm{prior}} < \lambda^{Q,2}_{\mathrm{prior}}$, 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:

$$C_p(\mathbf{f}_1) - C_p(\mathbf{f}_2) \geq \lambda^L_{\mathrm{prior}}(\lambda^L_{\mathrm{prior}} - \lambda^{Q,1}_{\mathrm{prior}})\mathrm{sup}_{\boldsymbol{\theta},\mathbf{h}\in U}|\nabla^2_{\boldsymbol{\theta},\mathbf{h}} D_{KL}(q|p)(\nabla^2_{\boldsymbol{\theta}}L)^{-1}\nabla_{\boldsymbol{\theta}} D_{KL}(q|p))|.$$

Доказательство. TODO □

TODO: выводы **Эксперимент: пример 1**
Эксперимент: пример 2