

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

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

где $\mathbf{s}^{j,k} \in (0, \infty)^K$ — гиперпараметр, отвечающий за смещенность плотности распределения относительно точек симплекса на $K^{j,k}$ 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 $\gamma^{j,k}$ порождается следующим образом:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Определение 1. Репараметризацией случайной величины ψ , распределенную по распределению q с параметрами θ_ψ назовем представление величины с помощью другой случайной величины, имеющей распределение, не зависящее от параметров θ :

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

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

Идею репараметризации поясним на следующем примере.

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

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

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

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

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

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

Тогда справедливо следующее выражение:

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



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) $\lambda_{\text{temp}} = 5, \mathbf{s} = [10, 0.1, 0.1]$.

Таким образом, распределение, позволяющее произвести репараметризацию, является более удобным для вычисления оценок интегральных оценок вида $\nabla_{\theta_{\mathbf{w}}} \mathbb{E}_{q(\mathbf{w}, \Gamma | \theta)} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma)$. Кроме того, данный подход позволяет значительно повысить точность вычисления градиента от функций, зависящих от случайных величин [?].

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 $\Gamma \in \mathbb{G}$ 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 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 $\lim_{\lambda_1, \lambda_2 \rightarrow 0} \text{inv-gamma}(\lambda_1, \lambda_2)$.

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

1. Параметры \mathbf{w} модели, распределенные нормально.
2. Структура модели Γ , содержащая все структурные параметры $\{\gamma^{j,k}, (j, k) \in E\}$ распределены по распределению Gumbel-Softmax.

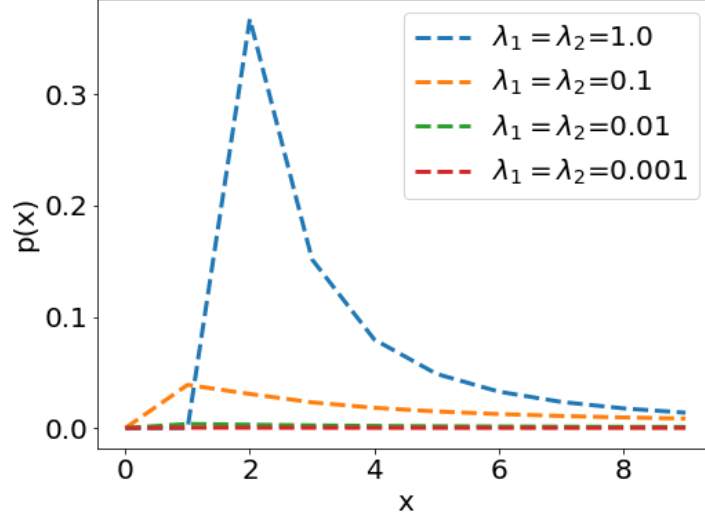


Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

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.

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

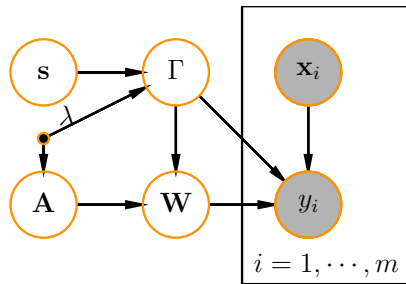


Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

1.2. Вариационная оценка для обоснованности вероятностной модели

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

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

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

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

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

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

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

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

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

Тогда справедлива следующая оценка:

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

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

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

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

Доказательство. Рассмотрим обоснованность:

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

Используя неравенство Йенсена получим

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

Декомпозируем распределение q по свойству условной дивергенции:

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

□

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

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

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

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

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

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

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

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

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

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

Теорема 2. Пусть

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

$$\text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{M} \quad \forall \mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \Gamma_1, \Gamma_2 \in U_{\Gamma};$$

3. вариационное распределение $q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ и априорное распределение $p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})$ являются абсолютно непрерывными и унимодальными на $U_{\mathbf{h}}, U_{\boldsymbol{\theta}}$;
4. вариационное распределение $q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ является липшецевым по \mathbf{w} ;
5. значение $q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ не равно нулю при $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$;
6. Решение задачи вида

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

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

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

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

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

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

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

9. 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:

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

Доказательство. Обозначим за \mathbf{h}_i — решение задачи (1.7) для вектора вариационных параметров $\boldsymbol{\theta}_i$. Воспользуемся неравенством Пинскера:

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

где $\|\cdot\|_{\text{TV}}$ — расстояние по вариации, F_q, F_p — функции распределения $q(\mathbf{w}, \Gamma|\boldsymbol{\theta}_i), p(\mathbf{w}, \Gamma|\mathbf{h}_i, \boldsymbol{\lambda})$. Т.к. дивергенция (1.6) состоит из двух неотрицательных величин, то обе они стремятся к нулю. Рассмотрим вторую величину:

$$\begin{aligned} 0 &= \lim_{i \rightarrow \infty} \mathbb{E}_{\Gamma \sim q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma)} \mathbb{E}_{\mathbf{w} \sim q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \log \left(\frac{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \right) = \\ \lim_{i \rightarrow \infty} \left| \int_{\Gamma} \int_{\mathbf{w}} \log \left(\frac{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \right) q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w} d\Gamma \right| &\geq \lim_{i \rightarrow \infty} \geq \\ \lim_{i \rightarrow \infty} \int_{\Gamma} \|F_q(\boldsymbol{\theta}_i) - F_p(\mathbf{h}_i)\|_{\text{TV}} q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\Gamma. \end{aligned}$$

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

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

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

$$\begin{aligned} &\left| \int_{\Gamma} \frac{q_{\mathbf{w}}(\text{mode } q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{[q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})]} q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\Gamma \right| = \\ &\left| \int_{\Gamma} \frac{q_{\mathbf{w}}(\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{[q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})]} q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\Gamma \right| \leq \\ &\int_{\Gamma} \left| \frac{q_{\mathbf{w}}(\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} - \frac{q_{\mathbf{w}}(\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{[q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})]} q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\Gamma \right| \leq \\ &\frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}} \in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \int_{\Gamma} |\mathbb{E}_{q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \mathbf{w} - \mathbb{E}_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w}| q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\Gamma \leq \\ &\leq \frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}} \in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{M}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} \iint_{\Gamma, \mathbf{w}} |\mathbf{w}| |q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})| q_\Gamma(\Gamma|\boldsymbol{\theta}_\Gamma) d\mathbf{w} d\Gamma, \end{aligned}$$

где C_l — максимальная константа Липшица для $q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})$ на $U_{\boldsymbol{\theta}}$. Определим случайную величину $\boldsymbol{\nu}(t), t \geq 0$ следующим образом:

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

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

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

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

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

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

где математическое ожидание берется по распределениям $\int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}) d\Gamma$, $\int_{\Gamma} q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma}) p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) d\Gamma$. Тогда

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

для любого t . Устремляя t к бесконечности, получим

$$\lim_{i \rightarrow \infty} \iint_{\Gamma, \mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| |p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{h}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} d\Gamma = 0.$$

Рассмотрим второе слагаемое. Т.к. $|\boldsymbol{\nu}(t)|$ — ограничена, то

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

Данное выражение стремится к нулю при $i \rightarrow \infty$.

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

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

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

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

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

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

1.3. Обобщающая задача

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

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

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

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

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

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

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

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

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

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

3. Перебор структуры:

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

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

4. Критерий Акаике:

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

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

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

где

$$\mathbf{h} = \arg \min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})), \quad (1.13)$$

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

5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = BIC_\lambda = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \log m|\{w : D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}), p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))| + \} p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \quad (1.14)$$

метапараметр λ_{prune} определен аналогично (1.13).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}), p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})) + p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \quad (1.15)$$

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

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

7. Валидация на отложенной выборке:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) + p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \rightarrow \boldsymbol{\theta}, \quad (1.16)$$

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

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 $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \Theta \times \mathbb{H} \times \boldsymbol{\lambda}$, если она удовлетворяет следующим критериям.

1. Область определения каждого параметра $w \in \mathbf{w}$, гиперпараметра $h \in \mathbf{h}$ и метапараметра $\lambda \in \boldsymbol{\lambda}$ не является пустым множеством и не является точкой.

2. Для каждого значения гиперпараметров \mathbf{h} оптимальное решение нижней (??) задачи оптимизации

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

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

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

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

такие что для любых векторов гиперпараметров, удовлетворяющих неравенству

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

выполняется неравенство

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

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

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

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

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

параметрическая сложность первой модели меньше, чем второй:

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

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

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

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

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

6. Критерий перебора оптимальных структур: существует набор метапараметров λ и константа

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

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

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

существует значение метапараметров λ' , такие что

- (a) Соответствие между вариационными параметрами $\theta^*(\mathbf{h}_1), \theta^*(\mathbf{h}_2)$ сохраняется при λ' .
 - (b) $Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \theta, \lambda) < Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \theta, \lambda)$ при λ' .
7. Критерий непрерывности: функции $L(\theta|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda)$ и $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \theta, \lambda)$ непрерывны по метапараметрам $\lambda \in U_\lambda$.

Первый критерий является техническим и используется для исключения из рассмотрения вырожденных задач оптимизации. Второй критерий говорит о том, что решение первого и второго уровня должны быть согласованы и определены однозначно. Критерии 3-5 определяют возможные критерии оптимизации, которые должны приближаться обобщающей задачей. Критерий 6 говорит о возможности перехода между различными структурами модели. Данный критерий говорит о том, что мы можем перейти от одного набора гиперпараметров \mathbf{h}_1 к другим \mathbf{h}_2 , если они соответствуют локальным оптимумам задачи оптимизации, и дивергенция соответствующих априорных распределений на структурах $p(\Gamma|\mathbf{h}, \lambda)$ значимо высока. При этом соответствующие вариационные распределения $q_\Gamma(\Gamma|\theta_\Gamma)$ могут оказаться достаточно близки. Возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения $q_\Gamma(\Gamma|\theta_\Gamma)$ 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 (1.9), (1.10), (1.11), (1.12), (1.14), (1.16) не являются обобщающими.

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

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

Теорема 4. Пусть q_Γ — 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_{θ} некоторой случайной величиной с конечным первым моментом.

Тогда задача (1.15) не является обобщающей.

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

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

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\theta} D_{\text{KL}}(q(\Gamma|\theta_1)|p(\Gamma|\lambda)) = \nabla_{\theta} D_{\text{KL}}(q(\Gamma|\theta_1)|p(\Gamma|\lambda')),$$

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

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

Т.к. выражение $\nabla_{\theta_{\Gamma}} q(\Gamma|\theta_2)$ 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 λ_{temp} равен при разных значениях метапараметров:

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

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

□

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)
 Δ^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
 Γ — структура модели
 \mathbb{I} — множество значений структуры модели
 U_{Γ} — область определения параметров модели
 \mathbf{h} — гиперпараметры модели
 \mathbb{H} — пространство гиперпараметров модели
 $U_{\mathbf{h}}$ — область определения гиперпараметров
 $\boldsymbol{\theta}$ — вариационные параметры модели
 Θ — 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
 $\boldsymbol{\theta}_{\Gamma}$ — вариационные параметры модели, аппроксимирующие структуру модели
 Θ_{Γ} — пространство вариационных параметров модели, аппроксимирующих структуру модели
 $U_{\boldsymbol{\theta}_{\Gamma}}$ — область определения вариационных параметров модели, аппроксимирующих структуру модели

λ — вектор метапараметров

λ — пространство метапараметров

U_λ — 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

$q_{\mathbf{w}}(\mathbf{h} | \Gamma, \theta_{\mathbf{w}})$ — 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 вариационных распределений

S — энтропия распределения

M — множество моделей без общей параметризации