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 параметры модели 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 $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})$ 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 ($\mathcal{GS}$) [**?**]:

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

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:

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

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

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

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

Заметим, что числитель $\lambda_{\text{temp}}^{K-1}$ имеет меньшую скорость сходимости, чем знаменатель, поэтому для вычисления предела достаточно проанализировать только знаменатель. TODO: пояснение. Он представлен суммой слагаемых вида:

$$\left(\frac{\prod_{l' \neq l} \gamma_{l'}^{\frac{1}{K}}}{\gamma_l^{\frac{K-1}{K}}}\right)^{\lambda_{\text{temp}}}, \tag{1.2}$$

возведенных в степень $(-K) < 0$.

Рассмотрим два случая: когда вектор γ лежит центре симплекса и не лежит. 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 (1.2) при λ_{temp} стремится к бесконечности. Для $l \neq l'$ предел выражения (1.2) при λ_{temp} стремится к нулю. Возводя сумму пределов в степень $(-K)$ получаем предел плотности, равный нулю.

Пусть $\boldsymbol{\gamma} = \frac{1}{K}$. Тогда выражение 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.

□

Первое свойство Gumbel-Softmax распределения позволяет использовать репараметризацию при вычислении градиента в вариационном выводе (англ. reparametrization trick). 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 $\mathbf{w}$ можно представить как функцию от непараметрического распределения:

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

где $\boldsymbol{\varepsilon}$ — случайная величина с известным распределением, не зависящим от $\boldsymbol{\theta}$. Тогда

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

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

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 [**?**].

Пример распределения Gumbel-Softmax при различных параметрах представлен на Рис. 1.1. В качестве альтернативы для априорного распределения на структуре выступает распределение Дирихле и 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

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

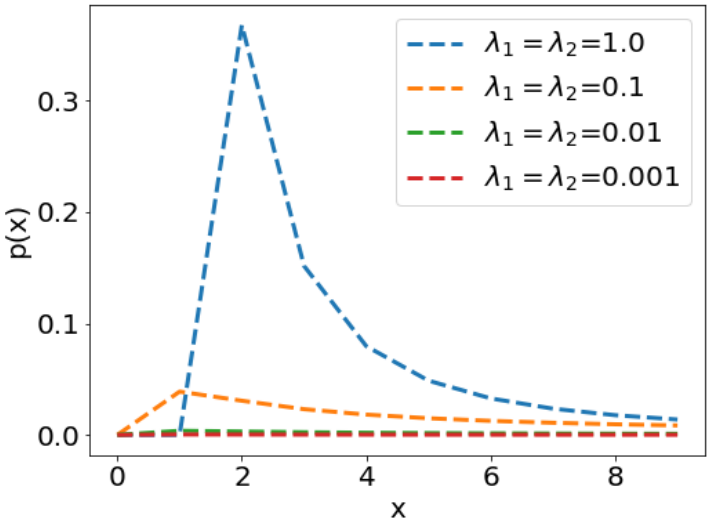

Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

$\mathbf{A}_l^{j,k}$ 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 inv-gamma(λ_1, λ_2) для разных значений метапараметров λ_1, λ_2 изображен на Рис. 1.2. Оптимизации без регуляризации соответствует случай предельного распределения $\lim_{\lambda_1, \lambda_2 \to 0}$ inv-gamma(λ_1, λ_2).

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

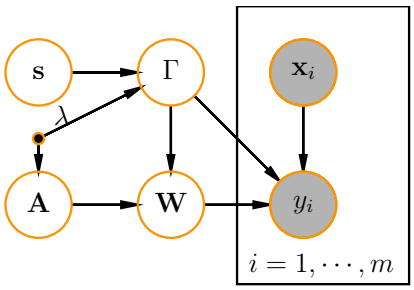

Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

1. Параметры $\mathbf{w}$ модели, распределенные нормально.
2. Структура модели $\mathbf{\Gamma}$, 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 параметрыне подлежат оптимизации и задаются экспертно.

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:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h} \in \mathbb{H}}, \tag{1.3}$$

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 (1.3).

Для вычисления обоснованности

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

из (1.3) предлагается использовать вариационную оценку обоснованности.

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

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

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

Тогда справедлива следующая оценка:

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) \geq \tag{1.4}$$

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

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

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

Доказательство. TODO: пояснения. Используя неравенство Йенсена получим

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

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

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 $\boldsymbol{\alpha}_q$.

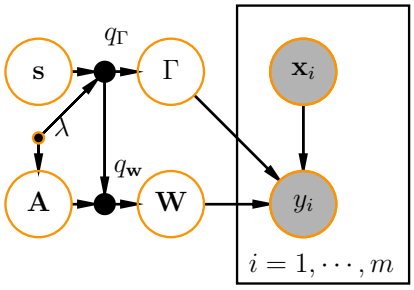

Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

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

Вариационными параметрами распределения q являются параметры распределений $q_\mathbf{w}, q_{\mathbf{\Gamma}}$:

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

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

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

Определение 1. 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:

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

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 [**?**]:

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

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 2. 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:

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

Относительнойвариационной 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 $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}} \subset \Theta$;
2. Усредненные по распределению мода априорного распределения $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ не зависят от гиперпараметров $\mathbf{h}$ и структуры $\boldsymbol{\Gamma}$ на $U_{\mathbf{h}}, \mathbb{\Gamma}$:

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

3. вариационное распределение $q_{\mathbf{w}}$ и априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h})$ являются абсолютно непрерывными и унимодальными на $U_{\mathbf{h}}, U_{\boldsymbol{\theta}}$.
4. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по распределениям $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}), p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.
5. мода и матожидание вариационного распределение $q_{\mathbf{w}}$ и априорного распределения $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ совпадают:

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

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

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

Тогда следующее выражение стремится к единице:

$$\mathsf{E}_q \boldsymbol{\rho}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1} \to 1.$$

Доказательство. Воспользуемся неравенством Пинскера:

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

где $||\cdot||_{\text{TV}}$ — расстояние по вариации, F_q, F_p — функции распределения $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ и $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$. Отсюда $\lim_{i\to\infty} ||F_q(\boldsymbol{\theta}) - F_p(\mathbf{h})||_{\text{TV}} = 0$. Из сходимости по вариации следует слабая сходимость распределений.

Рассмотрим разность мод:

$$\mathsf{E}_{q_{\boldsymbol{\Gamma}}}\text{mode } q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\Gamma}) - \mathsf{E}_{p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})}\text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}) =$$

$$= \mathsf{E}_{q_{\boldsymbol{\Gamma}}}\mathsf{E}_{q_{\mathbf{w}}}\mathbf{w} - \mathsf{E}_{p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})}\mathsf{E}_{p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w} =$$

$$= \mathsf{E}_q\mathbf{w} - \mathsf{E}_{p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h})}\mathbf{w}.$$

Т.к. вторые моменты величины $\mathbf{w}$ конечны для вариационного и априорного распределения, то функции $\mathsf{E}_{q(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}},\boldsymbol{\Gamma})}\mathbf{w}, \mathsf{E}_{p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h})}\mathbf{w}$ абсолютно интегрируемы, что в сочетании со слабой сходимостью позволяет записать:

$$\lim_{i\to\infty}\left(\mathsf{E}_q\mathbf{w} - \mathsf{E}_{p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h})}\mathbf{w}\right) = 0.$$

Т.к. вторые моменты случайных величин конечны, то конечны и первые моменты:

$$\lim_{i\to\infty}\mathsf{E}_q\mathbf{w} = \lim_{i\to\infty}\mathsf{E}_{p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h})}\mathbf{w} = \mathbf{M}.$$

Таким образом в пределе усредненные по структуре моды вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta})$ и априорного распределения $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h})$ совпадают. Т.к. наибольшее значение распределения $q_{\mathbf{w}}$ сосредоточено в моде распределения $q_{\mathbf{w}}$, то $\boldsymbol{\rho}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1}$ ограничена сверху единицей. Рассмотрим матожидание функции, обратной к отношению вариационных плотностей:

$$\mathsf{E}_q\boldsymbol{\rho}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1}$$

Т.к. функция ограничена, то предел можно внести под знак интеграла:

$$\lim_{i\to\infty}\mathsf{E}_q\boldsymbol{\rho}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1} =$$

$$= \mathsf{E}_q\lim_{i\to\infty}\boldsymbol{\rho}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1} =$$

$$\mathsf{E}_q\lim_{i\to\infty}\prod_{w\in\mathbf{w}}\frac{q\big(\text{mode } p\left(w|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}\right)|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}\big)}{q\big(\text{mode } q\left(w|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}\right)|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}\big)} =$$

$$\mathsf{E}_q\frac{q_{\mathbf{w}}(\mathbf{M})}{q_{\mathbf{w}}(\mathbf{M})} = 1.$$

□

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 последовательность q_i не обязана иметь предел.

1.3. Обобщающая задача

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

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

   Метод заключается в максимизации правдоподобия обучающей выборки и подвержен переобучению. Для использования данного метода в качестве задачи выбора модели предлагается следующее обобщение:

$$L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}). \tag{1.5}$$

   Данное обобщение эквивалентно методу правдоподобия при выборе в качестве q эмпирического распределения парамтетров и структуры. Метод не предполагает оптимизации гиперпараметров. Для формального соответствия данной задачи задаче выбора модели, т.е. двухуровневой задачи оптимизации, положим $L = Q$.

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L = Q = \mathsf{E}_q\big(\log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) + \log p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}). \tag{1.6}$$

В рамках данной задачи оптимизации гиперпараметры $\mathbf{h}$ фиксированы и не подлежат оптимизации.

3. Перебор структуры:

$$L = Q = \mathsf{E}_q \log p(\mathbf{y}, \mathbf{w}|\mathbf{X})[q_{\mathbf{\Gamma}} = p'] \tag{1.7}$$

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

4. Критерий Акаике:

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

Заметим,что в условия выбора модели на параметрическом множестве моделей данный критерий не имеет смысла, т.к. количество параметров для каждой модели одинаково. Предлагается следующая переформулировка:

$$L = Q = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - -|\{w : D_{\text{KL}}(\theta, \mathbf{h}) < \lambda\}|, \tag{1.8}$$

где

$$\mathbf{h} = \arg\min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma})|p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}', \boldsymbol{\lambda}), \tag{1.9}$$

λ — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение применимо только в случае,если выражение (**??**) определено однозначно.

5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

$$L = Q = BIC_\lambda = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \tag{1.10}$$

$$\log(m)|\{w : D_{\text{KL}}(\theta, \mathbf{h}) < \lambda\}|,$$

гиперпараметр $\mathbf{h}$ определ аналогично (**??**).

6. Метод вариационной оценки обоснованности:

$$L = Q = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})|p(\mathbf{\Gamma}, \mathbf{w}|\mathbf{h}, \boldsymbol{\lambda})). \tag{1.11}$$

В рамках данной задачи, все гиперпараметры $\mathbf{h}$ подлежат оптимизации.

7. Валидация на отложенной выборке:

$$L = \mathsf{E}_q \log p(\mathbf{y}_{\text{train}}, \mathbf{w}, \mathbf{\Gamma} | \mathbf{X}_{\text{train}}, \mathbf{h}, \boldsymbol{\lambda}), \tag{1.12}$$

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

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 значинях метапараметров, будет называть *обобщающей.*

Определение 3. Двухуровневую задачу оптимизации будем называть *обобщающей* на компакте $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \Theta \times \mathbb{H} \times \mathbb{\Lambda}$, если она удовлетворяет следующим свойствам.

1. Для каждого значения гиперпараметров $\mathbf{h}$ оптимальное решение нижней задачи оптимизации $\boldsymbol{\theta}^*$ определено однозначно.
2. Свойство максимизации правдоподобия выборки: существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_1 > 0, K_1 < \max_{\mathbf{h}_1, \mathbf{h}_2} Q(\mathbf{h}_1) - Q(\mathbf{h}_2)$,, такие что для любых векторов гиперпараметров, удовлетворяющих неравенству $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) \geq K_1$, выполняется неравенство $\mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) > \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \lambda_{\text{temp}}, \mathbf{f})$.
3. Свойство минимизации параметрической сложности: существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 > 0, K_2 \leq \max_{\mathbf{h}_1, \mathbf{h}_2} Q(\mathbf{h}_1) - Q(\mathbf{h}_2)$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h$, удовлетворяющих неравенству $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$ и при этом имеющие равенство ожидаемых правдоподобий выборок $\mathsf{E}_q \log\, p(\mathbf{y}|\boldsymbol{\theta}^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) = \mathsf{E}_q \log\, p(\mathbf{y}|\boldsymbol{\theta}^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f})$, параметрическая сложность первой модели меньше, чем второй: $C_p(\boldsymbol{\theta}^*(\mathbf{h}_1)|U_{\mathbf{h}}, \boldsymbol{\lambda}) < C_p(\boldsymbol{\theta}^*(\mathbf{h}_2)|U_{\mathbf{h}}, \boldsymbol{\lambda})$.

4. Свойства приближения оценки обоснованности: существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:

$$\max \mathbf{h} \in U_h Q(\arg\max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} L) \approx$$

$$\approx \max_{\mathbf{h} \in U_h, \boldsymbol{\theta} \in \Theta} \left(\mathsf{E}_q p(\mathbf{y}|\mathbf{w}, \mathbf{X}) - D_{KL}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))\right).$$

5. Свойство перебора структур: существует набор метапараметров $\boldsymbol{\lambda}$ и константа K_3, такие что для любых двух векторов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих векторов $\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}_2^* = \boldsymbol{\theta}^*(\mathbf{h}_2)$, полученных при метапараметрах λ и удовлетворяющих неравенству $D_{\mathrm{KL}}(q_{\boldsymbol{\Gamma}_2}, q_{\boldsymbol{\Gamma}_1}) > K_3, D_{\mathrm{KL}}(q_{\boldsymbol{\Gamma}_1}, q_{\boldsymbol{\Gamma}_2}) > K_3$ существуют значения гиперпараметров $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$, такие что:
   (a) соответствие между $\boldsymbol{\theta}_1$ и $\mathbf{h}_1$ и соответсвие между между $\boldsymbol{\theta}_2$ и $\mathbf{h}_2$ сохраняются при метапараметрах $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$.
   (b) $Q(\mathbf{h}_1, \lambda_1) > Q(\mathbf{h}_2, \lambda_1), Q(\mathbf{h}_1, \lambda_1) < Q(\mathbf{h}_2, \lambda_2)$.
6. Свойство нерперывности: $\mathbf{h}^*, \boldsymbol{\theta}^*$ непрерывны по метапараметрам.

Первое свойство говорит о том, что решение первого и второго уровня должны быть согласованы и определены однозначно. Свойства 2-4 определяют возможные критерии оптимизации, которые должны приближаться обобщающей задачей. Свойство 5 говорит о возможности перехода между различными структурами модели. Отметим, что данное условие крайне важно в условиях оптимизации моделей глубокого обучения, которые отличаются многоэкстремальностью. Последнее свойство говорит о том, что обобщающая задача должна позволять производить переход между различными критериями выбора параметров и структуры модели непрерывно.

Теорема 3. Рассмотренные задачи (1.5),(??),(??),(??),(??),(??),(??) не являются обобщающими.

Доказательство. Задачи (1.5),(??),(??),(??),(??) не имеют гиперпараметров, подлежащих оптимизации, поэтому не могут оптимизировать вариационную оценку.

Докажем, что задача (??) не является обобщающей. Пусть выполнены условия свойства 5. Возьмем в качестве векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2$ гиперпараметры, отличающиеся только параметрами распределения структуры. Т.к. для набора метапараметров должны быть выполнены условия соответствия, то градиенты $\nabla_{\boldsymbol{\theta}}$ должны быть нулевыми. Тогда

$$\nabla \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \nabla D_{\mathrm{KL}}(q|p_1) = \nabla \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - D_{\mathrm{KL}}(q|p_2).$$

Отсюда следует, что параметр температуры постоянен для всех наборов метапараметров. Метапараметры λ_1, λ_2 влияют на значение функции Q только при разных значениях гиперпараметров $\mathbf{A}$. Таким образом, приходим к противоречию.

При использовании кросс-валидации в функцию валидацию не входит ни один метапараметр, поэтому свойство 5 длянее также не выполняется. □

Теорема 4. Пусть задано непустое множество непрерывных по параметрам распределений на структуре $\mathbf{P}$. Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на компакте $U = U_{\boldsymbol{\theta}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \Theta \times \mathbb{H} \times \mathbb{\Lambda}$, где параметры распределений $\mathbf{P} \in \mathbb{\Lambda}$. Пусть также определены для метапараметров, определенных ниже, $\lambda^{\text{Q}}_{\text{likelihood}}, \lambda^{\text{prior}}_{\text{Q}}, \boldsymbol{\lambda}^{\text{struct}}_{\text{Q}}, \lambda^{\text{prior}}_{\text{L}}$ область определения включает значения $\{0, 1\}$. Тогда следующая задача является обобщающей на U.

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

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

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

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

где

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

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

Доказательство. Для доказательсta теоремы требуется доказать критерии 1-6 из определения обобщающей задачи. Критерий 1 следует из условий задачи.

Докажем критерий 2. Пусть $\lambda_{\text{temp}}, \lambda_1, \lambda_2, \lambda^{\text{prior}}_{\text{Q}}, \lambda^{\text{likelihood}}_{\text{Q}}$ удовлетворяют условиям, $\boldsymbol{\lambda}^{\text{struct}}_{\text{Q}} = \mathbf{0}$. Возьмем в качестве K_1 следующее выражение:

$$A = B + C$$

$$K_1 = \big(\max_{\mathbf{h}} \log p(\mathbf{h}|\mathbf{f}) - \min_{\mathbf{h}} \log\, p(\mathbf{h}|\mathbf{f}) + \max_{\mathbf{h}, \boldsymbol{\theta}} D_{\text{KL}}(q||p) - \min_{\mathbf{h}, \boldsymbol{\theta}} D_{\text{KL}}(q||p)\big).$$

Тогда

$$Q(\mathbf{h}_1) - Q(\mathbf{h}_2) = \lambda^{\text{likelihood}}_{\text{Q}} \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1) \lambda_{\text{temp}}, \mathbf{f}) - \lambda^{\text{likelihood}}_{\text{Q}} \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \lambda_{\text{ter}}$$

$$\lambda^{\text{prior}}_{\text{Q}} D_{\text{KL}}(q|p) - \lambda^{\text{prior}}_{\text{Q}} D_{\text{KL}}(q|p) + \log\, p(\mathbf{h}_2|\mathbf{f}) - \log\, p(\mathbf{h}_1|\mathbf{f}) > K_1.$$

Отсюда следует

$$\mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1) \lambda_{\text{temp}}, \mathbf{f}) - \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2) \lambda_{\text{temp}}, \mathbf{f}) > K_1 -$$

$$-\lambda^{\text{prior}}_{\text{Q}} D_{\text{KL}}(q|p) + \lambda^{\text{prior}}_{\text{Q}} D_{\text{KL}}(q|p) - \log\, p(\mathbf{h}_2|\mathbf{f}) + \log\, p(\mathbf{h}_1|\mathbf{f}) > 0.$$

Докажем критерий 3. Пусть гиперпараметры удовлетворяю условиям. Пусть

$$K_2 = -\big(\max_{\mathbf{h}} \log p(\mathbf{h}|\mathbf{f}) - \min_{\mathbf{h}} \log p(\mathbf{h}|\mathbf{f}) - \min_{\mathbf{h}, \boldsymbol{theta}} - \min \mathbf{h} \max \boldsymbol{\theta} (-\lambda^{\text{prior}}_{\text{L}} D_{\text{KL}}(q|p) + \log p(\mathbf{y}$$

Рассмотрим разность параметрических сложностей двух векторов:

$$C_p(\boldsymbol{\theta}_1)-C_p(\boldsymbol{\theta}_2) = \min D_{\mathrm{KL}}(q_1|p)-\min D_{\mathrm{KL}}(q_2|p) \geq \min D_{\mathrm{KL}}(q_1|p)-D_{\mathrm{KL}}(q_2|p) = \min D$$

$$= Q_1-Q_2-\log p(\mathbf{h}_1|\mathbf{f})+\log p(\mathbf{h}_2|\mathbf{f})-\min D_{\mathrm{KL}}(q_1|p) \geq \max \min D_{\mathrm{KL}}(q_1|p)+K_2-\max \log$$

Рассмотрим слагаемое $D_{\mathrm{KL}}(q_1|p)$:

$$-D_{\mathrm{KL}}(q_1|p) = \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\big(-D_{\mathrm{KL}}(q_1|p) + \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta}) - \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta}) \geq$$

$$\geq \frac{1}{\lambda_{\mathrm{L}}^{\mathrm{prior}}}\big(-D_{\mathrm{KL}}(q_1|p) + \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta}) - \max_{\boldsymbol{theta}} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta}),$$

Т.к. $\boldsymbol{theta}$— решение задачи оптимизации, то справедлива оценка:

$$\frac{1}{\min_{\mathbf{h}} \max_{\boldsymbol{theta}} \lambda_{\mathrm{L}}^{\mathrm{prior}}}\big(-D_{\mathrm{KL}}(q_1|p) + \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta}) - \max_{\boldsymbol{theta}} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h},\boldsymbol{\theta})$$

Отсюда следует, что разнсоть больше нуля.

Докажем критерий 4. Пусть $\lambda_{\mathrm{Q}}^{\mathrm{likelihood}}$ и другие удовлетворяют условиями, $\boldsymbol{\lambda}_{\mathrm{Q}}^{\mathrm{struct}} = \mathbf{0}$. Тогда оптимизационную задачу можно записать как:

$$\arg\max_{\mathbf{h}\in U_h} \arg\max_{\boldsymbol{\theta}} \left(\mathsf{E}_q p(\mathbf{y}|\mathbf{w},\mathbf{X}) - D_{KL}(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}))\right),$$

что и требовалось доказать.

Докажем критерий 5. Пусть выполнены условия. Возьмем в качестве K_4 следующее выражение:

$$K_4 = \min_{\boldsymbol{theta}_1,\boldsymbol{theta}_2 : D_{\mathrm{KL}}(p_1|p_2) \geq \frac{1}{\max \lambda_{\mathrm{comb}}}} \max Q_1 - Q_2 D_{\mathrm{KL}}(q_1,q_2).$$

Пусть векторы метапараметров $\lambda_1, \lambda_2, \lambda_3$ отличаются лишь метапараметром λ_{comb}. Для всех трех наборов нижняя задача оптимизации L одинакова, поэтому выполняется первое условие свойства. Пусть $Q_1 - Q_2 > 0$ при λ_{comb}. Положим распределение из $\mathbf{P}$ равным распределению q_2. Тогда:

$$Q_1 - Q_2 - \max \lambda_{\mathrm{comb}} D_K L(p_1|p_2) < 0.$$

Докажем критерий 6. Т.к. априорные распределения задаются нерперывными функциями плотности, и функция плотностираспределения структуры ограничена на компакте, то дивергенция непрерывна по метапараметрам. Тк. слагаемые функций оптимизации непрерывны, то непрерывна и сами функции оптимизации. □

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

В предельном случае, когда температура λ_{temp} близка к нулю, а множество $\mathbf{P}$ состоит из распределений, близких к дискретным 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 $\lambda_{\text{temp}} = 0.1$. Пусть в качестве множества распределений $\mathbf{P}$ используется два распределения Gumbel-Softmax, сконцентрированных близко к вершинам симплекса:

$$\mathbf{P} = [\mathcal{GS}([0.8, 0.1, 0.1]^{\text{T}}, 0.1), \mathcal{GS}([0.1, 0.8, 0.1]^{\text{T}}, 0.1)].$$

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

Обобщающая задача: переформулировка через градиент

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

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

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

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

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

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

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 5. 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 □

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 $m >> 0$, $\lambda^L_{\text{prior}} > 0, m \gg 0, \frac{m}{\lambda^L_{\text{prior}}} \in \mathbb{N}$. Тогда оптимизация функции

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

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

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

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

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

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

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

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

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

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

Таким образом, задачи оптимизации 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 задано параметрическое множество вариационных распределений: $q(\boldsymbol{\theta})$. Пусть $\lambda^L_{\text{likelihood}} = \lambda^L_{\text{prior}} = \lambda^Q_{\text{prior}} > 0, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Тогда:

1. Задача оптимизации (**??**) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:
$$\log\hat{p}(\mathbf{y}|\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f}) + \log p(\mathbf{h}|\mathbf{f}) \to \max_{\mathbf{h}}.$$
2. Вариационное распределение q приближает апостериорное распределение $p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\lambda_{\text{temp}},,\mathbf{f})$ наилучшим образом:
$$D_{\mathrm{KL}}(q||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f})) \to \min_{\boldsymbol{\theta}}.$$

Доказательство. TODO □

Теорема 8. Пусть также распределение q декомпозируется на два независимых распределения для параметров $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$ модели $\mathbf{f}$. Тогда вариационные распределения $q_{\mathbf{w}}, q_{\boldsymbol{\Gamma}}$приближают апостериорные распределения $p(\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f}), p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{y},\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f})$ наилучшим образом:

$$D_{\mathrm{KL}}(q_{\boldsymbol{\Gamma}}||p(\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\lambda_{\text{temp}},\mathbf{f})) \to \min, \quad D_{\mathrm{KL}}(q_{\mathbf{w}}||p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h},\mathbf{f})) \to \min.$$

Доказательство. TODO □

Следующие теоремы посвящены ассимптотическим свойствам представленной обобщающей задачи.

Теорема 9. Пусть $\lambda^Q_{\text{likelihood}} = \lambda^L_{\text{prior}} > 0, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Тогда предел оптимизации

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

доставляет минимум параметрической сложности.

Доказательство. TODO □

Теорема 10. Пусть $\lambda^L_{\text{likelihood}} = 1, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Пусть $\mathbf{f}_1, \mathbf{f}_2$ — результаты градиентной оптимизации при разных значениях гиперпараметров $\lambda^{Q,1}_{\text{prior}}, \lambda^{Q,2}_{\text{prior}}, \lambda^{Q,1}_{\text{prior}} < \lambda^{Q,2}_{\text{prior}}$, 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:

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

Доказательство. TODO □

Для анализа свойств структуры модели $\boldsymbol{\Gamma}$ введем понятие структурной сложности.

Определение 4. Структурной сложностью C_s модели назовем энтропию структур $\boldsymbol{\Gamma}$, полученных из вариационного распределения q:

$$C_s = -\mathsf{E}_q \mathsf{E}_\Gamma \log p_{\boldsymbol{\Gamma}}.$$

TODO: пояснение

Теорема 11. Пусть $\lambda_{\text{train}} > 0$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ — вариационные параметры, такие что $\boldsymbol{\theta}_1$ лежит внутри произведения симплексов структуры, $\boldsymbol{\theta}_2$ — на вершинах симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \to 0} \frac{L(\boldsymbol{\theta}_2)}{L(\boldsymbol{\theta}_1)} \to 0.$$

Доказательство. TODO □

Теорема 12. Пусть $\lambda_{\text{train}} > 0$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ — вариационные параметры, такие что $\boldsymbol{\theta}_1$ лежит внутри произведения симплексов структуры, $\boldsymbol{\theta}_2$ — в центре симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \to \infty} \frac{L(\boldsymbol{\theta}_2)}{L(\boldsymbol{\theta}_1)} \to 0.$$

Доказательство. TODO □

TODO: вывод **Эксперимент: пример 1**
Эксперимент: пример 2