

# Глава 1

## Выбор субоптимальной структуры модели

В данной главе рассматривается задача выбора структуры модели глубокого обучения. Предлагается ввести вероятностные предположения о распределениях параметров и 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 позволяет проводить оптимизацию, соответствующую нескольким критериям выбора структуры модели: методу максимального правдоподобия, последовательному увеличению и снижению сложности модели, полному перебору структуры модели, а также получению 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:

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

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 [?]:

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

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

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

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

Перечислим свойства, которыми обладает распределение Gumbel-Softmax:

1. Реализация  $\hat{\gamma}_l$ , т.е.  $l$ -й компоненты случайной величины  $\gamma$  порождается следующим образом:

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

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

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

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

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

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

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

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

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

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

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

возведенных в степень  $(-K) < 0$ .

Пусть хотя бы для одной компоненты  $l$  выполнено:  $\gamma_l \neq \frac{1}{K}$ . Пусть  $l'$  соответствует индексу максимальной компоненты вектора  $\gamma$ . Для  $l = l'$  предел выражения (1.2) при  $\lambda_{\text{temp}}$  стремится к бесконечности. Для  $l \neq l'$  предел выражения (1.2) при  $\lambda_{\text{temp}}$  стремится к нулю. Возводя сумму пределов в степень  $(-K)$  получаем предел плотности, равный нулю.

Пусть  $\gamma = \frac{1}{K}$ . Тогда выражение с точностью до множителя упрощается до  $\lambda^{K-1}$ . 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  $\Gamma$  определена для модели  $\mathbf{f}$  однозначно. Рассмотрим математическое ожидание логарифма правдоподобия выборки модели по некоторому непрерывному распределению  $q$ :

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

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

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

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

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

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

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 гиперпараметра



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

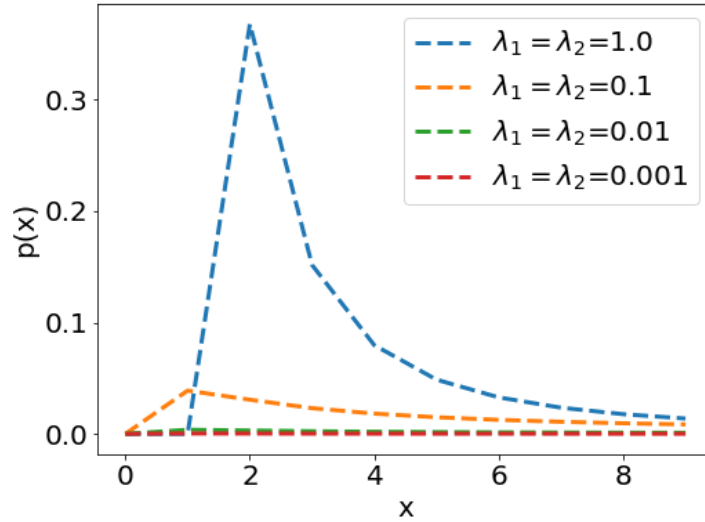


Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

$\mathbf{A}_l^{j,k}$  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. Калибруя метапарамы  $\lambda_1, \lambda_2$  можно получить более сильную или более слабую регуляризацию [?]. Пример распределений  $\text{inv-gamma}(\lambda_1, \lambda_2)$  для разных значений метапараметров  $\lambda_1, \lambda_2$  изображен на Рис. 1.2.

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

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

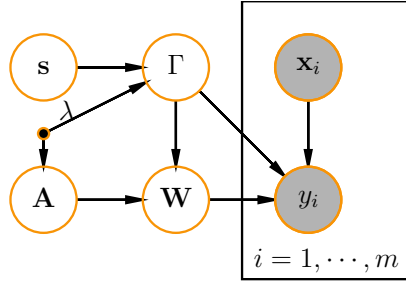


Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

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:  $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ .

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

## 1.2. Вариационная оценка для обоснованности вероятностной модели

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

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

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

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

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

где  $\mathbf{h}^*$  — решение задачи оптимизации (1.3).

Для вычисления обоснованности

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

из (1.3) предлагается использовать вариационную оценку обоснованности.

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

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

Тогда справедлива следующая оценка:

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

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

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

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

*Доказательство.* Используя неравенство Йенсена получим

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

Декомпозируем распределение  $q$  по свойству условной дивергенции:

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

□

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

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

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

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

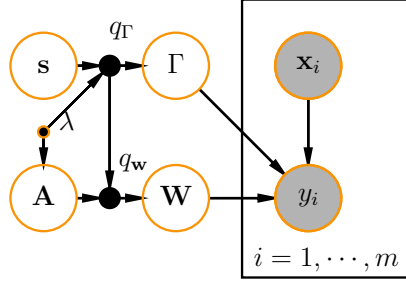


Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

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

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

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:

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

Параметрическая сложность модели соответствует ожидаемой длине описания параметров модели при условии заданного параметрического априорного распределения [?].

Одним из критериев удаления неинформативных параметров в вероятностных моделях является отношение вариационной плотности параметров в моде распределения к вариационной плотности параметра в нуле [?]:

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

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

Обобщим понятие относительной вариационной плотности на случай произвольных распределений.

**Определение 2.** Относительной вариационной плотностью параметра  $w \in \mathbf{w}$  при условии структуры  $\mathbf{\Gamma}$  и гиперпараметров  $\mathbf{h}$  назовем отношение вариационной плотности в моде вариационного распределения параметра к вариационной плотности в моде априорного распределению параметра:

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

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

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

**Теорема 2.** Пусть

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

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

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

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

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

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

Тогда следующее выражение стремится к единице:

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

*Доказательство.* Воспользуемся неравенством Пинскера:

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

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



Рассмотрим разность мод:

$$\begin{aligned} & \mathbb{E}_{q_{\Gamma} \text{ mode } q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)} - \mathbb{E}_{p(\Gamma|\mathbf{h}, \boldsymbol{\lambda}) \text{ mode } p(\mathbf{w}|\Gamma, \mathbf{h})} = \\ &= \mathbb{E}_{q_{\Gamma}} \mathbb{E}_{q_{\mathbf{w}}} \mathbf{w} - \mathbb{E}_{p(\Gamma|\mathbf{h}, \boldsymbol{\lambda})} \mathbb{E}_{p(\mathbf{w}|\Gamma, \mathbf{h}, \boldsymbol{\lambda})} \mathbf{w} = \\ &= \mathbb{E}_q \mathbf{w} - \mathbb{E}_{p(\mathbf{w}, \Gamma|\mathbf{h})} \mathbf{w}. \end{aligned}$$

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

$$\lim_{i \rightarrow \infty} (\mathbb{E}_q \mathbf{w} - \mathbb{E}_{p(\mathbf{w}, \Gamma|\mathbf{h})} \mathbf{w}) = 0.$$

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

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

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

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

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

$$\begin{aligned} & \lim_{i \rightarrow \infty} \mathbb{E}_q \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1} = \\ &= \mathbb{E}_q \lim_{i \rightarrow \infty} \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda})^{-1} = \\ & \mathbb{E}_q \lim_{i \rightarrow \infty} \prod_{w \in \mathbf{w}} \frac{q(\text{mode } p(w|\Gamma, \mathbf{h}, \boldsymbol{\lambda})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})}{q(\text{mode } q(w|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})|\Gamma, \boldsymbol{\theta}_{\mathbf{w}})} = \\ & \mathbb{E}_q \frac{q_{\mathbf{w}}(\mathbf{M})}{q_{\mathbf{w}}(\mathbf{M})} = 1. \end{aligned}$$

□

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

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

Априорным распределением  $p(\mathbf{w}, \Gamma | \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w} | \mathbf{h}, \boldsymbol{\lambda})$  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  $p(\mathbf{w}, \Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ .

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

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

Данное обобщение эквивалентно методу правдоподобия при выборе в качестве  $q$  эмпирического распределения параметров 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.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

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

В рамках данной задачи оптимизации параметры априорных распределений  $\mathbf{A}, \mathbf{s}$  выступают в качестве метапараметров и не подлежат оптимизации.

3. Перебор структуры:

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

где  $p'$  — некоторое распределение на структуре, выступающее в качестве метапараметра.

4. Критерий Акаике:

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

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:

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

где

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

,  $\lambda$  — метапараметр алгоритма,  $U_{\mathbf{h}} \subset \mathbb{H}$  — область определения задачи по гиперпараметрам.

5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

$$L = Q = \text{BIC}_{\lambda} = \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) - \log(m) |\{w : D_{\text{KL}}(\theta, \mathbf{h}) < \lambda\}|, \quad (1.10)$$

гиперпараметр  $\mathbf{h}$  определ аналогично (1.9).

6. Метод вариационной оценки обоснованности.

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

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

7. Hold-out кросс-валидация.

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

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

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

1. Для каждого значения гиперпараметров  $\mathbf{h}$  оптимальное решение нижней задачи оптимизации  $\theta^*$  определено однозначно.
2. Свойство максимизации правдоподобия выборки: существует  $\lambda \in U_{\lambda}$  и  $K_1 > 0$ , такие что для любых векторов гиперпараметров, удовлетворяющих неравенству  $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$ , выполняется неравенство  $E_q \log p(\mathbf{y}|\mathbf{X}, \theta^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) > E_q \log p(\mathbf{y}|\mathbf{X}, \theta^*(\mathbf{h}_2), \lambda_{\text{temp}}, \mathbf{f})$ .
3. Свойство минимизации параметрической сложности: существует  $\lambda \in U_{\lambda}$  и  $K_2 > 0$ , 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  $E_q \log p(\mathbf{y}|\theta^*(\mathbf{h}_1), \lambda_{\text{temp}}, \mathbf{f}) = E_q \log p(\mathbf{y}|\theta^*(\mathbf{h}_2), \lambda_{\text{temp}}, \mathbf{f})$ , параметрическая сложность первой модели меньше, чем второй:  $C_p(\theta^*(\mathbf{h}_1)|U_{\mathbf{h}}, \lambda) < C_p(\theta^*(\mathbf{h}_2)|U_{\mathbf{h}}, \lambda)$ .
4. Свойства приближения оценки обоснованности: существует значение гиперпараметров  $\lambda$ , такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:

$$\begin{aligned} & \arg \max_{\mathbf{h} \in U_h} Q(\arg \max_{\theta \in U_{\theta}} L) \approx \\ & \approx \arg \max_{\mathbf{h} \in U_h} \arg \max_{\theta} (E_q p(\mathbf{y}|\mathbf{w}, \mathbf{X}) - D_{KL}(q(\mathbf{w}, \Gamma|\theta)|p(\mathbf{w}, \Gamma|\mathbf{h}, \lambda))) . \end{aligned}$$

5. Свойство перебора структур: существует набор метапараметров  $\lambda$  и константа  $K_3$ , такие что для любых двух векторов  $\mathbf{h}_1, \mathbf{h}_2$  и соответствующих векторов  $\theta_1 = \theta^*(\mathbf{h}_1), \theta_2 = \theta^*(\mathbf{h}_2)$ , полученных при метапараметрах  $\lambda$  и удовлетворяющих неравенству  $D_{KL}(q_{\Gamma_2}, q_{\Gamma_1}) > K_3, D_{KL}(q_{\Gamma_1}, q_{\Gamma_2}) > K_3$  существуют значения гиперпараметров  $\lambda_1, \lambda_2$ , такие что:

(a) Соответствие между  $\theta_1$  и  $\mathbf{h}_1$  и соответствие между  $\theta_2$  и  $\mathbf{h}_2$  сохраняются при метапараметрах  $\lambda_1, \lambda_2$ .

(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 непрерывности:  $\mathbf{h}^*, \theta^*$  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), (1.6), (1.7), (1.8), (1.10), (1.11), (1.12) не являются обобщающими.

*Доказательство.* TODO □

**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  $U = U_\theta \times U_{\mathbf{h}} \times U_\lambda \subset \Theta \times \mathbb{H} \times \mathbb{A}$ , где параметры распределений  $\mathbf{P} \in \mathbb{A}$ . Пусть также определены для метапараметров, определенных ниже,  $\lambda_{\text{likelihood}}^Q, \lambda_Q^{\text{prior}}, \lambda_Q^{\text{struct}}, \lambda_L^{\text{prior}}$  область определения включает значения  $\{0, 1\}$ . Тогда следующая задача является обобщающей на  $U$ .

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

где

$$\begin{aligned} q^* &= \arg \max_q L = \mathbb{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) & (L^*) \\ &- \lambda_L^{\text{prior}} D_{KL}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})). \end{aligned}$$

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

Докажем критерий 2. Пусть  $\lambda_Q^{\text{prior}} = 0, \lambda_Q^{\text{likelihood}} = 1, \lambda_Q^{\text{struct}} = \mathbf{0}$ . Зафиксируем некоторое значение метапараметров  $\lambda_1, \lambda_2$ . Т.к.  $U_{\mathbf{h}}$  — компакт, возьмем в качестве константы  $K_1$  разницу между максимальным и минимальным значением  $p(\mathbf{h} | \mathbf{f})$ :

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

Тогда

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

Отсюда следует

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

Докажем критерий 3. TODO

Докажем критерий 4. Пусть  $\lambda_Q^{\text{likelihood}} = \lambda_Q^{\text{prior}} = \lambda_L^{\text{prior}} = 1, \boldsymbol{\lambda}_Q^{\text{struct}} = \mathbf{0}$ . Тогда оптимизационную задачу можно записать как:

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

что и требовалось доказать.

Докажем критерий 5. TODO

Докажем критерий 6. TODO □

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

В предельном случае, когда температура  $\lambda_{\text{temp}}$  близка к нулю, а множество  $\mathbf{P}$  состоит из распределений, близких к дискретным соответствующим всем возможным структурам, калибровка  $\boldsymbol{\lambda}_Q^{\text{struct}}$  порождает последовательность задач оптимизаций, схожую с перебором структур. Для примера рассмотрим вырожденный случай поведения функции  $Q$ , когда  $\lambda_Q^{\text{likelihood}} = \lambda_Q^{\text{prior}} = 0$ . Пусть в 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:

$$\mathbf{P} = [\text{Gumbel-Softmax}([0.8, 0.1, 0.1]^T, 0.1), \text{Gumbel-Softmax}([0.1, 0.8, 0.1]^T, 0.1)].$$

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

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

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

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  $U_W \times U_{\boldsymbol{\Gamma}} \times U_H \times U_{\boldsymbol{\lambda}} \subset \mathbb{W} \times \mathbb{F} \times \mathbb{H} \times \mathbb{A}$ , при этом  $U_H \times U_{\boldsymbol{\lambda}}$  — компакт. Тогда решение задачи градиентной оптимизации

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

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

*Доказательство.* TODO □

## 1.4. Анализ обобщающей задачи

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

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

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

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

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

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

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

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

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

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

Таким образом, задачи оптимизации совпадают, что и требовалось доказать.  $\square$

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

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

**Теорема 7.** Пусть задано параметрическое множество вариационных распределений:  $q(\boldsymbol{\theta})$ . Пусть  $\lambda_{\text{likelihood}}^L = \lambda_{\text{prior}}^L = \lambda_{\text{prior}}^Q > 0, \boldsymbol{\lambda}_{\text{struct}}^Q = \mathbf{0}$ . Тогда:

1. Задача оптимизации ( $Q^*$ ) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:

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

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

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

*Доказательство.* TODO  $\square$



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

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

*Доказательство.* TODO □

Следующие теоремы посвящены асимптотическим свойствам представленной обобщающей задачи.

**Теорема 9.** Пусть  $\lambda_{\text{likelihood}}^Q = \lambda_{\text{prior}}^L > 0, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Тогда предел оптимизации

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

доставляет минимум параметрической сложности.

*Доказательство.* TODO □

**Теорема 10.** Пусть  $\lambda_{\text{likelihood}}^L = 1, \lambda_{\text{struct}}^Q = \mathbf{0}$ . Пусть  $\mathbf{f}_1, \mathbf{f}_2$  — результаты градиентной оптимизации при разных значениях гиперпараметров  $\lambda_{\text{prior}}^{Q,1}, \lambda_{\text{prior}}^{Q,2}, \lambda_{\text{prior}}^{Q,1} < \lambda_{\text{prior}}^{Q,2}$ , полученных при начальном значении вариационных параметров  $\boldsymbol{\theta}_0$  и гиперпараметров  $\mathbf{h}_0$ . Пусть  $\boldsymbol{\theta}_0, \mathbf{h}_0$  принадлежат области  $U$ , в которой соответствующие функции  $L$  и  $Q$  являются локально-выпуклыми. Тогда:

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

*Доказательство.* TODO □

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

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

$$C_s = -\mathbb{E}_q \mathbb{E}_{\mathbf{\Gamma}} \log p_{\mathbf{\Gamma}}.$$

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

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

*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

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

*Доказательство.* TODO

□

TODO: вывод Эксперимент: пример 1

Эксперимент: пример 2