

# 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:

$$\mathbf{w}_k^{i,j} \sim \mathcal{N}(\mathbf{0}, \gamma_k^{i,j} (\mathbf{A}_k^{i,j})^{-1}),$$

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

В качестве априорного распределения для структуры  $\mathbf{\Gamma}$  предлагается использовать произведение **распределение** 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:

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

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. Реализацию  $\hat{\gamma}_p$ , т.е.  $p$ -й компоненты случайной величины  $\gamma$  можно породить следующим образом:

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

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

2. Свойство округления:  $p(\gamma_{p_1} > \gamma_{p_2}, p_1 \neq p_2) = \frac{s_{p_1}}{\sum_{p'} s_{p'}}$ .
3. При устремлении температуры к нулю реализация случайной величины концентрируется на вершинах симплекса:

$$p(\lim_{\lambda_{\text{temp}} \rightarrow 0} \gamma_p = 1) = \frac{s_p}{\sum_{p'} s_{p'}}.$$

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

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

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

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

$$\frac{\lambda_{\text{temp}}^{K-1}}{\left( \sum_{p=1}^K s_p \gamma_p^{-\frac{K-1}{K} \lambda_{\text{temp}}} \sum_{p'=1}^K [p \neq p'] s_{p'} \gamma_{p'}^{-\frac{1}{K} \lambda_{\text{temp}}} \right)^K}$$

Заметим, что числитель  $\lambda_{\text{temp}}^{K-1}$  имеет меньшую скорость сходимости, чем знаменатель. Знаменатель является суммой слагаемых вида:

$$\left( \frac{\prod_{p' \neq p} \gamma_{p'}^{\frac{1}{K}}}{\gamma_p^{\frac{K-1}{K}}} \right)^{\lambda_{\text{temp}}}. \quad (1.2)$$

Пусть хотя бы для одного  $p$ :  $\gamma_p \neq \frac{1}{K}$ . Пусть  $p'$  соответствует индексу максимальной компоненты вектора  $\gamma$ . Для  $p = p'$  предел выражения (1.2) при  $\lambda_{\text{temp}}$  стремится к бесконечности. Для  $p \neq p'$  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.

□

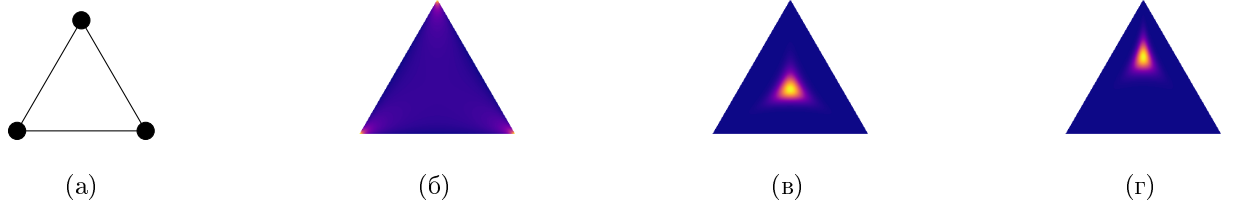


Рис. 1.1. Пример распределения Gumbel-Softmax при различных значениях параметров: а)  $\lambda_{temp} \rightarrow 0$ , б)  $\lambda_{temp} = 1, \mathbf{s} = [1, 1, 1]$ , в)  $\lambda_{temp} = 5, \mathbf{s} = [1, 1, 1]$ , г)  $\lambda_{temp} = 5, \mathbf{s} = [10, 0.1, 0.1]$ .

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

Заметим, что предлагаемое априорное распределение неоднозначно: одно и то же распределение можно получить с различными значениями гиперпараметра  $\mathbf{A}_k^{i,j}$  и структурного параметра  $\gamma_k^{i,j}$ . В качестве регуляризатора для матрицы  $(\mathbf{A}_k^{i,j})^{-1}$  предлагается использовать обратное гамма-распределение:

$$(\mathbf{A}_k^{i,j})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2),$$

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

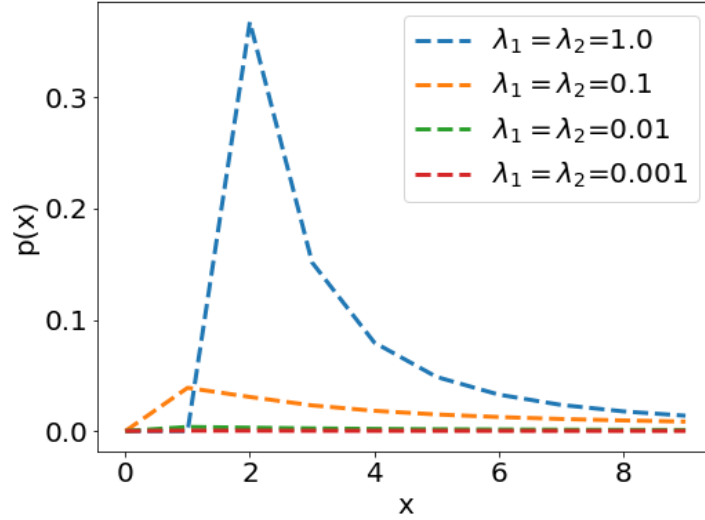


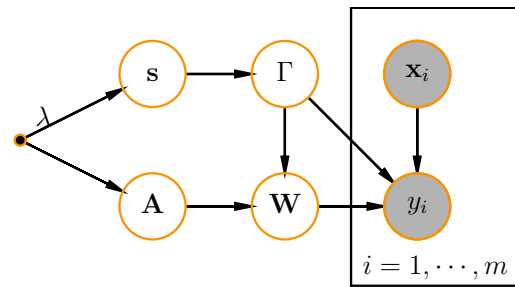
Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

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

## 1.2. Вариационная оценка для обоснованности вероятностной модели

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

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



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

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

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

из (1.3) предлагается использовать вариационную оценку обоснованности.

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

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

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

$$q_{\Gamma}(\Gamma | \theta_{\Gamma}) \approx p(\Gamma | \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda).$$

Тогда справедлива следующая оценка:

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

$$\mathbb{E}_{\Gamma \sim q_{\Gamma}} \mathbb{E}_{\mathbf{w} \sim q_{\mathbf{w}}} \log p(\mathbf{y} | \mathbf{w}, \Gamma, \mathbf{X}) - D_{\text{KL}}(q_{\Gamma}(\Gamma | \theta_{\Gamma}) | p(\Gamma | \mathbf{h}, \lambda)) - D_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}, \Gamma) | p(\mathbf{w} | \Gamma, \mathbf{h})),$$

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

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

*Доказательство.* Используя неравенство Йенсена получим

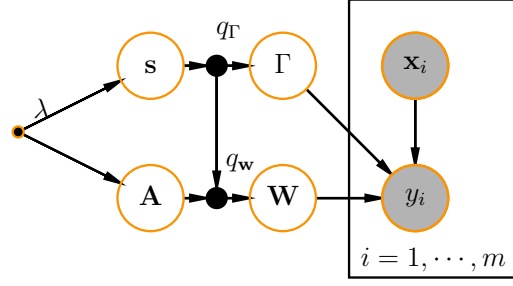
$$\log p(\mathbf{y} | \mathbf{X}, \mathbf{h}, \lambda) \geq$$

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

Декомпозируем распределение  $q$  по свойству условной дивергенции:

$$D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta) | p(\mathbf{w}, \Gamma | \mathbf{h})) = D_{\text{KL}}(q_{\Gamma}(\Gamma | \theta_{\Gamma}) | p(\Gamma | \mathbf{h}, \lambda)) + D_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}, \Gamma) | p(\mathbf{w} | \Gamma, \mathbf{h})).$$

□



**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  $q_{\mathbf{W}}$  предлагается использовать нормальное распределение, не зависящее от структуры модели  $\mathbf{\Gamma}$ :

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

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

В качестве вариационного распределения  $q_{\mathbf{\Gamma}}$  предлагается использовать произведение распределений Gumbel-Softmax. **конкатенацию** параметров концентрации распределений обозначим как  $\mathbf{s}_q$ . Температуру вариационного распределения на структуре  $\mathbf{\Gamma}$  обозначим как  $\theta_{\text{temp}}$ .

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

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

~~где  $\mathbf{s}_q, \theta_{\text{temp}}$  — параметры Gumbel Softmax распределений.~~

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

Для вычисления приближенного значения вариационной оценки обоснованности (1.4) предлагается использовать следующую формулу:

$$\begin{aligned} & \sum_{r=1}^R \log p(\mathbf{y} | \boldsymbol{\mu} + \boldsymbol{\alpha}_q \circ \hat{\epsilon}_r, \hat{\mathbf{\Gamma}}_r, \mathbf{X}) - \sum_{r=1}^R \left( \log q_{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}_r | \boldsymbol{\theta}_{\mathbf{\Gamma}}) - p(\hat{\mathbf{\Gamma}} | \mathbf{h}, \boldsymbol{\lambda}) \right) - \\ & - \sum_{r=1}^R \frac{1}{2} \left( \hat{\mathbf{\Gamma}}_r^{-1} \text{tr}(\mathbf{A}_q \mathbf{A}^{-1}) + \boldsymbol{\mu}^T \hat{\mathbf{\Gamma}}_r^{-1} \mathbf{A}^{-1} \boldsymbol{\mu} - |\mathbf{W}| + \log \frac{|\mathbf{\Gamma}_r \mathbf{A}|}{|\mathbf{A})_q|} \right), \end{aligned}$$

где  $r$  — множество реализаций случайных величин, по котором вычисляется значения вариационной оценки обоснованности,  $\hat{\epsilon}_r \sim \mathcal{N}(0, 1)$ ,  $\hat{\Gamma}_r$  — реализация случайной величины, соответствующей структуре  $\Gamma$ .

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

**Определение 1.** Параметрической сложностью  $C_p(\boldsymbol{\theta})$  модели с вариационными параметрами  $\boldsymbol{\theta}$  назовем минимальную дивергенцию между вариационным и априорным распределением:

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

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 **вариационная плотность** [?]: отношение вариационной плотности параметров в моде распределения к вариационной плотности параметра в нуле:

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

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

Обобщим понятие относительной вариационной плотности на случай произвольных распределений.

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

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

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

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

**Теорема 2.** Пусть вариационное распределение  $q_{\mathbf{w}}$  и априорное распределение  $p(w | \Gamma, \mathbf{h})$  являются унимодальными с ограниченным вторым моментом и свойством:

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

Пусть мода априорного распределения  $p(w | \Gamma, \mathbf{h})$  не зависит от гиперпараметров  $\mathbf{h}$ . Пусть также  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots$  — бесконечная последовательность векторов вариационных параметров, такая что  $\lim_{i \rightarrow \infty} C_p(\boldsymbol{\theta}_i) = 0$ . Тогда вариационная

плотность данной последовательности стремится к единице почти наверно по вероятностной мере  $q_{\Gamma}$ :

$$\rho(\mathbf{w}|\Gamma, \mathbf{h}_i) \xrightarrow{\text{п.н. по } q_{\Gamma}} 1.$$

где  $\mathbf{h}_i = \arg \min_{\mathbf{h} \in \mathbb{H}} D_{\text{KL}}(q(\mathbf{w}, \Gamma|\boldsymbol{\theta})|p(\mathbf{w}, \Gamma|\mathbf{h}))$ .

*Доказательство.* Предел параметрической сложности перепишем как

$$\lim_{i \rightarrow \infty} \min_{\mathbf{h}} D_{\text{KL}}(q_{\Gamma}(\Gamma|\boldsymbol{\theta}_{\Gamma})|p(\Gamma|\mathbf{h}, \boldsymbol{\lambda})) + D_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)|p(\mathbf{w}|\Gamma, \mathbf{h}))$$

Т.к. параметрическая сложность состоит из двух неотрицательных слагаемых, то в пределе оба слагаемых достигают нуля. Рассмотрим второе слагаемое:

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

Т.к. предел равен нулю, то для множества событий меры 1 по  $q_{\Gamma}$  выполняется:

$$\hat{D}_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)|p(\mathbf{w}|\Gamma, \mathbf{h})) = 0,$$

где  $\hat{D}_{\text{KL}}(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)|p(\mathbf{w}|\Gamma, \mathbf{h}))$  — дивергенция при фиксированном значении переменной  $\Gamma$ . Для каждого значения  $\Gamma$  за исключением счетного множества значений по неравенству Пинскера следует:

$$\|F_q - F_p\| \rightarrow 0,$$

где  $F_q, F_p$  — функции распределения для  $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma), p(\mathbf{w}|\Gamma, \mathbf{h})$ . Из теоремы Шеффе следует, что  $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma) - p(\mathbf{w}|\Gamma, \mathbf{h})$  сходится слабо к нулю. Т.к. второй момент параметров конечен, то последовательность равномерно интегрируема:

$$\begin{aligned} \lim_{i \rightarrow \infty} (\text{mode}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)} - \text{mode}_{p(\mathbf{w}|\Gamma, \mathbf{h})}) &= \\ &= \lim_{i \rightarrow \infty} \mathbb{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)} \mathbf{w} - \mathbb{E}_{p(\mathbf{w}|\Gamma, \mathbf{h})} \mathbf{w} = 0. \end{aligned}$$

В пределе мода распределения  $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}, \Gamma)$  совпадает с модой априорного распределения, отсюда относительная плотность стремится к единице почти всюду.  $\square$

Теорема утверждает, что при устремлении параметрической сложности моделей к нулю, параметры модели становятся неинформативными и подлежащими удалению.



### 1.3. Обобщающая задача

Рассмотрим основные 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:

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

Данное обобщение эквивалентно методу правдоподобия при выборе в качестве  $q$  эмпирического распределения параметров и структуры. Метод не предполагает оптимизации гиперпараметров. Для формального соответствия данной задаче выбора 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:

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

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

3. Перебор структуры:

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

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

4. Критерий Акаике:

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

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:

$$L = Q = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - |\{w : C_p(w) < \lambda\}|,$$

где  $\lambda$  — метапараметр алгоритма.

5. Информационный критерий Шварца:

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

Переформулируем данный критерий аналогично критерию AIC:

$$L = Q = BIC_\lambda = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \log(m)|\{w : C_p(w) < \lambda\}|.$$

6. Метод вариационной оценки обоснованности.

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

7. Hold-out кросс-валидация.

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

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

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, будет называть *обобщающей*.

**Определение 3.** Двухуровневую задачу оптимизации будем называть *обобщающей* на области  $U \subset \Theta \times \mathbb{H} \times \mathbb{A}$ , 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  $K_1 \in \mathbb{R}_+$ , такие что для любых векторов гиперпараметров  $\mathbf{h}_1, \mathbf{h}_2 \in U_h$ ,  $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$  : матожидания правдоподобия выборок:  $\mathbb{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) > \log \mathbb{E}_q p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$ .
3. Свойство минимизации параметрической сложности: существует  $\boldsymbol{\lambda} \in U_\lambda$  и  $K_2 \in \mathbb{R}_+$ , такие что для любых векторов гиперпараметров  $\mathbf{h}_1, \mathbf{h}_2 \in U_h$ ,  $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$ ,  $\mathbb{E}_q \log p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) = \log \mathbb{E}_q p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$ , количество ненулевых параметров у первой модели меньше, чем у второй.

4. Свойства приближения оценки обоснованности: существует значение гиперпараметров  $\lambda$ , такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели.
5. Свойство перебора структур: существует константа  $K_3$ , такая что для любых двух векторов  $\mathbf{h}_1, \mathbf{h}_2$  и соответствующих векторов  $\theta_1^*, \theta_2^*$  :  $D_{KL}(q_{\Gamma_2}, q_{\Gamma_1}) > K_3, D_{KL}(q_{\Gamma_1}, q_{\Gamma_2}) > K_3$ : существуют значения гиперпараметров  $\lambda_1, \lambda_2$ , такие что  $Q(\mathbf{h}_1, \lambda_1) > Q(\mathbf{h}_2, \lambda_1), Q(\mathbf{h}_1, \lambda_1) < Q(\mathbf{h}_2, \lambda_2)$ .
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 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 некоторой области  $U \subset \Theta \times \mathbb{H} \times \mathbb{A}$ , где параметры распределений  $\mathbf{P} \in \mathbb{A}$ . Тогда следующая задача является обобщающей на  $U$ .

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

где

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

*Доказательство.* 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  $\lambda_Q^{\text{struct}}$ .

В предельном случае, когда множество температур  $\lambda_{\text{temp}}$  близка к нулю, а множество  $\mathbf{P}$  состоит из распределений, близких к дискретным, и соответствующих всем возможным структурам, калибровка  $\lambda_Q^{\text{struct}}$  порождает последовательность задач оптимизаций, схожую с перебором структур.

TODO

**Обобщающая задача: переформулировка через градиент**

**Обобщающая задача: адекватность задачи**

**Обобщающая задача: свойства коэффициентов**

**Решение задачи**

**Эксперимент: пример 1**

**Эксперимент: пример 2**