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:

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

где $(\mathbf{A}_k^{i,j})^{-1}$ — диагональная матрица. Апирорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h})$ 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 $\boldsymbol{\Gamma}$ предлагается использовать произведение распределений Gumbel-Softmax [**?**]:

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

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. Реализацию $\hat{\gamma}_l$, т.е. l-й компоненты случайной величины $\boldsymbol{\gamma}$ можно породить 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 случайной величины концентрируется на вершинах симплекса:
$$p(\lim_{\lambda_{\text{temp}}\to 0} \gamma_l = 1 | \mathbf{s}, \lambda_{\text{temp}}) = \frac{s_l}{\sum_{l'} s_{l'}}.$$
4. При устремлении температуры к бесконечности плотность распределения концентрируется в центре симплекса:
$$\lim_{\lambda_{\text{temp}}\to\infty} p(\boldsymbol{\gamma}|\mathbf{s}, \lambda_{\text{temp}}) = \begin{cases} \infty, \boldsymbol{\gamma}_l = \frac{1}{K}, l \in \{1, \dots, K\}, \\ 0, \text{ иначе.} \end{cases} \tag{1.1}$$

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

Доказательство. Формула плотности записывается следующим образом с точностью до множителя:
$$\frac{\lambda_{\text{temp}}^{K-1}}{\left(\sum_{l=1}^{K} s_l \gamma_l^{-\frac{-K-1}{K}\lambda_{\text{temp}}} \sum_{l'=1}^{K} [l \neq l'] s_l \gamma_l^{-\frac{1}{K}\lambda_{\text{temp}}}\right)^K}$$

Заметим, что числитель $\lambda_{\text{temp}}^{K-1}$ имеет меньшую скорость сходимости, чем знаменатель. Знаменатель является суммой слагаемых вида:
$$\left(\frac{\prod_{l'\neq l} \gamma_{l'}^{\frac{1}{K}}}{\gamma_l^{\frac{K-1}{K}}}\right)^{\lambda_{\text{temp}}}. \tag{1.2}$$

Пусть хотя бы для одного l: $\gamma_l \neq \frac{1}{K}$. 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.

$\square$

Первое свойство Gumbel-Softmax распределения позволяет использовать репараметризацию при вычислении градиента в вариационном выводе (англ. reparametrization trick). Идея подхода заключается в следующем. Рассмотрим для примера математическое ожидание логарифма правдоподобия выборки модели по некоторому непрерывному распределению q:

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

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

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

Первое слагаемое в общем видел сложновычислимо. Пусть распределение q можно представить как функцию от непараметрического распределения:

$$q(\mathbf{w}) = q(g(\varepsilon)).$$

Тогда

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

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 $\mathbf{A}_k^{i,j}$ 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 λ_1, λ_2 можно получить более сильную или

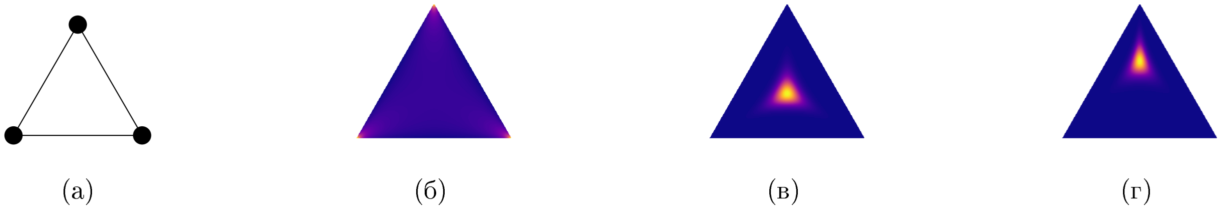

Рис. 1.1. Пример распределения Gumbel-Softmax при различных значениях параметров: а) $\lambda_{temp} \to 0$, б) $\lambda_{temp} = 1, \mathbf{s} = [1, 1, 1]$, в) $\lambda_{temp} = 5, \mathbf{s} = [1, 1, 1]$, г) $\lambda_{temp} = 5, \mathbf{s} = [10, 0.1, 0.1]$.

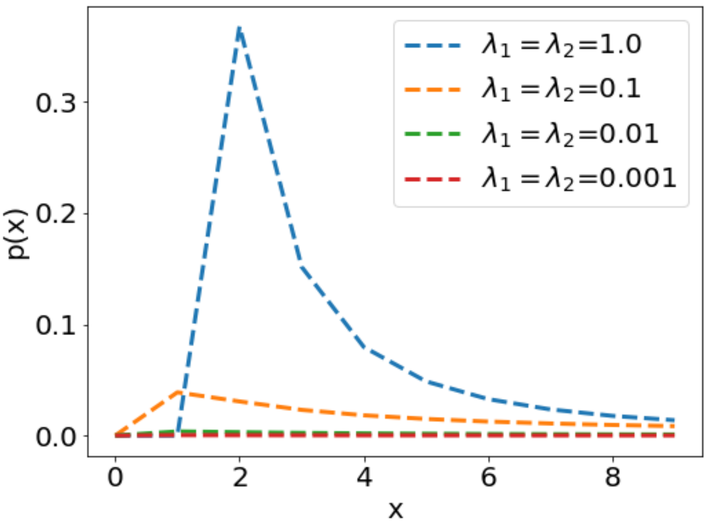

Рис. 1.2. Графики обратных гамма распределений для различных значений метапараметров.

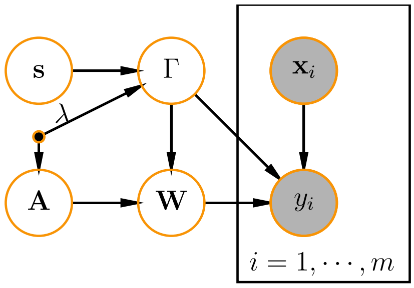

Рис. 1.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

более слабую регуляризацию [?]. Пример распределений inv-gamma(λ_1, λ_2) для разных значений метапараметров λ_1, λ_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: $\mathbf{h} = [\mathrm{diag}(\mathbf{A}), \mathbf{s}]$, 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:

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

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 $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) = q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_\mathbf{w}) q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ — 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 [?]:

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

Доказательство. Используя неравенство Йенсена получим

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

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

Декомпозируем распределение q по свойству условной дивергенции:

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

$\square$

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

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

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

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

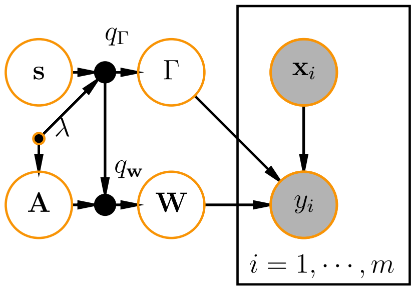

Рис. 1.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

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

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

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 $C_p(\boldsymbol{\theta})$ модели с вариационными параметрами $\boldsymbol{\theta}$ на компакте $U_\mathbf{h} \subset \mathbb{H}$ назовем минимальную дивергенцию между вариационным и априорным распределением:

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

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 [**?**]:

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

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 $\mathbf{\Gamma}$ и гиперпараметров $\mathbf{h}$ назовем отношение моды вариационного распределения параметра к моде априорного распределению параметра:

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

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

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

Теорема 2. Пусть

1. заданы компактные множества $U_\mathbf{h} \subset \mathbb{H}, U_{\boldsymbol{\theta}} \subset \mathbb{\Theta}$;
2. мода априорного распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}))$ не зависит от гиперпараметров $\mathbf{h}$ на $U_\mathbf{h}$:
$$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}_1) = p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}_2) = p(\mathbf{w}, \mathbf{\Gamma}) \forall \mathbf{h}_1, \mathbf{h}_2 \in U_\mathbf{h}.$$
3. вариационное распределение $q_\mathbf{w}$ и априорное распределение $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}))$ являются абсолютно непрерывными и унимодальными на $U_\mathbf{h} \subset \mathbb{H}, U_{\boldsymbol{\theta}}$.
4. мода и матожидание вариационного распределение q и априорного распределение $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}))$ распределения совпадают.
5. задана последовательность $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ — бесконечная последовательность векторов вариационных параметров, такая что $\lim_{i\to\infty} C_p(\boldsymbol{\theta_i}|U_\mathbf{h}) = 0, \boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.
6. $\mathbf{h}_i$.

$\mathbf{h}_i$. Тогда матожидание вариационной плотности данной последовательности стремится к единице:

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

Доказательство. Воспользуемся неравенством Пинскера:

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

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 $\lim_{i\to\infty} ||F_q(\boldsymbol{\theta}) - F_p(\mathbf{h})||_{\text{TV}} = 0$. Из сходимости по вариации следует слабая сходимость распределений.

Рассмотрим разность мод:

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

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

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

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

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

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

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

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

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

□

Теорема утверждает, что при устремлении параметрической сложности модели к нулю, параметры модели становятся неинформативными и подлежащими 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 расп распределений, чье матожидание стремится к нулю. Априорным распределением $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}) = p(\mathbf{w}|\mathbf{h})$ при этом может являться семейство нормальных распределений с нулевым средним.

1.3. Обобщающая задача

Рассмотрим основные критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

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

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 положим $L = Q$.

2. Метод максимальной апостериорной вероятности.

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

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L = Q = \mathsf{E}_q \mathrm{loglog} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \mathrm{log} p(\mathbf{w}|\boldsymbol{\lambda}) + \mathrm{log} p(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{w}). \qquad (1.6)$$

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

3. Перебор структуры:

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

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

4. Критерий Акаике:

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

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:

$$L = Q = \mathrm{log} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - |\{w : C_p(\theta|U_{\mathbf{h}}) < \lambda\}|, \qquad (1.8)$$

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

5. Информационный критерий Шварца:

$$\mathrm{log} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - 0.5 \mathrm{log}(m)|\{w : C_p(\theta|U_{\mathbf{h}}) < \lambda\}|.$$

Переформулируем данный критерий аналогично критерию AIC:

$$L = Q = BIC_\lambda = \mathrm{log} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \log(m)|\{w : C_p(w) < \lambda\}|. \qquad (1.9)$$

6. Метод вариационной оценки обоснованности.

$$L = Q = \mathsf{E}_q \mathrm{log} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - D_{\mathrm{KL}}(q|p). \qquad (1.10)$$

7. Hold-out кросс-валидация.

$$L = \mathsf{E}_q \mathrm{log} p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}), \qquad (1.11)$$

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

Каждый из рассмотренных критерии удовлетворяет хотя бы одному из перечисленных свойтсв:

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 $U \subset \mathbb{\Theta} \times \mathbb{H} \times \mathbb{\Lambda}$, 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 $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров, удовлетворяющих неравенству $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$, выполняется неравенство $\mathsf{E}_q\log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) > \log\mathsf{E}_q\; p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$.
3. Свойство минимизации параметрической сложности: существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h$, удовлетворяющих неравенству $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$ и при этом имеющие равенство ожидаемых правдоподобий выборок $\mathsf{E}_q\log\; p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) = \log\mathsf{E}_q\; p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$, параметрическая сложность первой модели меньше, чем второй: $C_p(\boldsymbol{\theta}^*(\mathbf{h}_1)|U_\mathbf{h}) < C_p(\boldsymbol{\theta}^*(\mathbf{h}_2|U_\mathbf{h})$.
4. Свойства приближения оценки обоснованности: существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели: $\arg\max_{\mathbf{h}\in U_h} Q(\arg\max_{\boldsymbol{\theta}} \in U_{\boldsymbol{\theta}} L) \approx \arg\max_{\mathbf{h}\in U_h} \mathsf{E}_q p(\mathbf{y}|\mathbf{w}, \mathbf{X}) - D_{KL}(q|p)$.
5. Свйоство перебора структур: существует константа K_3, такая что для любых двух векторов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих векторов $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*$: $D_{\text{KL}}(q_{\boldsymbol{\Gamma}_2}, q_{\boldsymbol{\Gamma}_1}) > K_3, D_{\text{KL}}(q_{\boldsymbol{\Gamma}_1}, q_{\boldsymbol{\Gamma}_2}) > K_3$ существуют значения гиперпараметров $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$, 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 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),(1.6),(1.7),(1.8),(1.9),(1.10),(1.11) не являются обобщающими.

Доказательство. TODO $\square$

Теорема 4. Пусть задано непустое множество непрерывных по параметрам распределний на структуре $\mathbf{P}$. Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на компакте $U \subset \mathbb{\Theta} \times \mathbb{H} \times \mathbb{\Lambda}$, где параметры распределений $\mathbf{P} \in \mathbb{\Lambda}$. Тогда следующая задача является обобщающей на U.

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

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

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

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

где

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

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

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

Докажем критерий 2. Пусть $\lambda_{\text{Q}}^{\text{prior}} = 0, \boldsymbol{\lambda}_{\text{Q}}^{\text{struct}} = \mathbf{0}$. Зафиксируем некоторое значение метапараметров λ_1, λ_2. Т.к. $U_{\mathbf{h}}$ — компакт, возьмем в качестве константы K_1 разницу между максимальным и минимальным значением $p(\mathbf{h}|\mathbf{f})$:

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

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

Докажем критерий 3. Пусть $\lambda_{\text{Q}}^{\text{likelihood}} = 0, \boldsymbol{\lambda}_{\text{Q}}^{\text{struct}} = \mathbf{0}$. Зафиксируем некоторое значение метапараметров λ_1, λ_2. Т.к. $U_{\mathbf{h}}$ — компакт, возьмем в качестве константы K_1 разницу между максимальным и минимальным значением $p(\mathbf{h}|\mathbf{f})$: TODO

Докажем критерий 4. Пусть $\lambda_{\text{Q}}^{\text{likelihood}} = \lambda_{\text{Q}}^{\text{prior}} = \lambda_{\text{L}}^{\text{prior}} = 1, \boldsymbol{\lambda}_{\text{Q}}^{\text{struct}} = \mathbf{0}$. Тогда оптимизационную задачу можно записать как: TODO, что и требовалось доказать.

Докажем критерий 5. TODO

Докажем критерий 6. TODO $\square$

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

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

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 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-Каарло с порождением R реализаций величин $\mathbf{w}, \mathbf{\Gamma}$:

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

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

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

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

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

Для решения двухуровневой задачи предлагается использовать градиентные методы.

Теорема 5. Пусть 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 стремится к локальному минимуму $\mathbf{h}^* \in U$ исходной задачи оптимизации (Q^*) 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 задана выборка $\mathbf{X}, \mathbf{y}$ мощности m.
Пусть задана модель $\mathbf{f}(\mathbf{w}, \mathbf{X})$ и распределение q, апппроксимирующее апостериорное распределеине параметров $\mathbf{w}$ этой модели.
Рассмотрим выражение $\frac{1}{m}\text{ELBO}_\gamma$:

$$\frac{1}{m}\text{ELBO}_\gamma(\mathbf{X}, \mathbf{y}, q) = \frac{1}{m}\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \frac{\gamma}{m}\text{KL}(q|p(\mathbf{w})),$$

где $\gamma > 0$.

Пусть $\frac{1}{m}\text{ELBO}_\gamma$ сходится п.н. при $m \to \infty$ к функции $L(q)$
(вообще, она еще от гиперпараметров зависит, но здесь это будет лишним, прим. Олег).

Тогда функция $\frac{1}{m_0}\text{ELBO}_1$ для выборки мощности $m_0 = \frac{m}{\gamma}$ из той же генеральной совокупности сходится почти наверно к этой же функции $L(q)$:

$$\frac{1}{m_0}\text{ELBO}_1(\hat{\mathbf{X}}, \hat{\mathbf{y}}, q) \to^{\text{п.н.}} L(q),$$

где $|\hat{\mathbf{X}}| = m_0$.

Доказательство. Рассмотрим величину $\frac{1}{m}\text{ELBO}_\gamma$:

$$\frac{1}{m}\text{ELBO}_\gamma(\mathbf{X}, \mathbf{y}) = \frac{1}{m}\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \frac{\gamma}{m}\text{KL}(q|p(\mathbf{w})).$$

По УЗБЧ:

$$\frac{1}{m}\text{ELBO}_\gamma(\mathbf{X}, \mathbf{y}) \to^{\text{п.н.}}_{m\to\infty} \mathsf{E}_\mathbf{X}\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \frac{\gamma}{m}\text{KL}(q|p(\mathbf{w})) = L(q).$$

Аналогично рассмотрим $\frac{1}{m_0}\text{ELBO}_1$ для выборки мощностью $m_0 = \frac{m}{\gamma}$:

$$\frac{1}{m_0}\text{ELBO}_1(\hat{\mathbf{X}}, \hat{\mathbf{y}}) \to^{\text{п.н.}}_{m\to\infty} \mathsf{E}_\mathbf{X}\mathsf{E}_q\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \frac{1}{m_0}\text{KL}(q|p(\mathbf{w})) =$$

$$= \mathsf{E}_{\mathbf{X}}\mathsf{E}_q \text{log} p(\mathbf{y}|\mathbf{X},\mathbf{w}) - \frac{\gamma}{m}\text{KL}(q|p(\mathbf{w})) = L(q),$$

предельные функции совпадают, что и требовалось доказать. □

Таким образом, для достаточно большого m и $\gamma > 0, \gamma \neq 1$ оптимизация параметров и гиперпараметров эквивалентна оптимизации ELBO для выборки другой мощности:

$$\max_q \text{ELBO}_\gamma(\mathbf{X},\mathbf{y},q) \propto \max_q \frac{1}{m}\text{ELBO}_\gamma(\mathbf{X},\mathbf{y},q) \sim \max_q \frac{1}{m_0}\text{ELBO}_1(\hat{\mathbf{X}},\hat{\mathbf{y}},q) \sim$$

$$\sim \max_q \text{ELBO}_1(\hat{\mathbf{X}},\hat{\mathbf{y}},q)$$

К примеру, оптимизация ELBO_γ при $\gamma > 1$ эквивалентна оптимизации ELBO для выборки меньшей мощности (и бОльшего вклада априорного распределения в оптимизацию).

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

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

Доказательство. TODO □

Теорема 8. Пусть также распределение q декомпозируется на два независимых распределения для параметров $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$ модели $\mathbf{f}$:

$$q = q_{\mathbf{w}}q_{\boldsymbol{\Gamma}}, q_{\boldsymbol{\Gamma}} \approx p(\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\mathbf{f}), q_{\mathbf{w}} \approx p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{y},\mathbf{X},\mathbf{h},\mathbf{f}).$$

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:

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

Доказательство. TODO □

Следующие теоремы посвящены ассимптотическим свойствам представленной обобщающей задачи.

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

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

доставляет минимум параметрической сложности. Существует компактная область U, такая что для любой точки $\boldsymbol{\theta}_0 \in U$ предел данной оптимизации доставляет нулевую параметрическую сложность: $C_p = 0$.

Доказательство. TODO □

Теорема 10. Пусть $\lambda_{\text{likelihood}}^{L} = 1, \boldsymbol{\lambda}_{\text{struct}}^{Q} = \mathbf{0}$. 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:

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

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:

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

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