1 Аннотация

В работе рассматривается задача выбора структуры модели глубокого обучения. Модель — это вычислительный вероятностный граф, т.е. граф, в котором ребрами выступают нелинейные функции, а вершинами — результаты действия функцией на выборку. Каждому ребру поставлено в соответствие множество нелинейных функций, такое что линейная комбинация этих функций определяет дифференцируемую функцию заданной сигнатуры. Структурой модели назовем веса линейной комбинации этих функций.

Для нахождения оптимальной структуры предлагается ввести вероятностную интерпретацию модели, т.е. предположения о распределениях параметров и структуры модели. Проводится градиентная оптимизация параметров и гиперпараметров модели на основе байесовского вариационного вывода. Решается двухуровневая задача оптимизации: на первом уровне проводится оптимизация нижней оценки правдоподобия модели по вариационным параметрам модели. На втором уровне проводится оптимизация гиперпараметров модели. В качестве оптимизируемой функции для гиперпараметров модели предлагается обобщенная функция правдоподобия. Показано, что данная функция позволяет проводить оптимизацию несколькими алгоритмами: последовательным добавлением и удалением параметров, полным перебором, а также максимизацией нижней оценки правдоподобия модели.

Проводится сравнение с эвристическими алгоритмами выбора структуры модели. Вычислительный эксперимент проводится на синтетических данных и выборке рукописных цифр MNIST.

Цель работы: предложить метод выбора модели субоптимальной сложности, позволяющий проводить выбор модели в нескольких режимах (ELBO, AddDel, полный перебор, оптимизация без регуляризации и с регуляризацией).

2 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m, \tag{1}$$

состоящая из множества пар «объект-метка»

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{y} \subset \mathbb{Y}.$$

Метка y объекта $\mathbf{x}$ принадлежит либо множеству: $y \in \mathbb{Y} = \{1, \ldots, Z\}$ в случае задачи классификации, где Z — число классов, либо некоторому подмножеству вещественных чисел $y \in \mathbb{Y} \subseteq \mathbb{R}$ в случае задачи регрессии. Далее будем полагать, что пары объект $(\mathbf{x}, y)$ являются реализацией некоторой случайно величины и порождены независимо.

Определим семейство моделей глубокого обучения для дальнейшего выбора оптимальной модели. Будем рассматривать семейство моделей как граф V, E.

Каждому ребру $(i,j) \in E$ сопоставим вектор базовых функций $\mathbf{g}^{i,j}$ мощности $K^{i,j}$. Вершины V — промежуточные промежуточные представления выборки под действием данных функций.

Перейдем к формальному определению семейства моделей. Пусть задан граф V, E. Пусть для каждого ребра $(i,j) \in E$ определен вектор функций $\mathbf{g}^{i,j}$. Граф V, E называется семейством моделей, если функция, задаваемая рекурсивно как

$$f_j(\mathbf{x}) = \sum_{i \in \mathrm{Adj}(v_j)} \langle \boldsymbol{\gamma}^{i,j}, \mathbf{g}^{i,j} \rangle \left(\mathbf{f}_k(\mathbf{x}) \right), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x}$$

является дифференцируемой по параметрам функцией из $\mathbb{R}^n$ во множество $\mathbb{Y}$ при любых значениях векторов $\boldsymbol{\gamma}^{j,k}$.

Параметрами модели $\mathbf{W}$ будем называть конкатенацию всех параметров подмоделей $\mathbf{f}_i$. Структурой модели $\boldsymbol{\Gamma}$ будем называть конкатенацию всех структурных параметров $\boldsymbol{\gamma}^{i,j}$. Моделью будем называть совокупность параметров $\mathbf{W}$ и структуры $\boldsymbol{\Gamma}$.

Пусть значения каждого структурного параметра $\boldsymbol{\gamma}^{i,j}$ лежат на симплексе $\Delta^{K^{i,j}-1}$. Пусть для каждого структурного параметра $\boldsymbol{\gamma}^{i,j} \in \boldsymbol{\Gamma}$ определено априорное Gumbel-softmax распределение:

$$\boldsymbol{\gamma}^{i,j} \sim \mathcal{GS}(\mathbf{m}^{i,j}, c_{\text{temp}}), \quad \mathbf{m}^{i,j} \in \mathbb{R}^{K^{i,j}}, \quad c_{\text{temp}} > 0,$$

где $\mathbf{m}^{i,j}$ — параметр средних, c_{temp} — температура (или концентрация) распределения. (вообще здесь можно и Дирихле, нужно ли здесь об этом говорить или обобщить?) Обозначим за $\mathbf{m}$ объединение всех векторов средних $\mathbf{m}^{i,j}, (i,j) \in E$.

Обозначим за S биективное отношение между параметром модели $w \in \mathbf{W}$ и весами $\boldsymbol{\gamma}$ базовых функцией $\mathbf{g}$, которой принадлежит данный параметр. Априорное распределение параметров зададим следующим образом:

$$\mathbf{W} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{A}^{-1} S(\mathbf{W})\right).$$

где $\mathbf{A}$ — диагональная матрица с положительными элементами на диагонали.

Пусть также определено правдоподобие выборки $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$.

Определение. Правдоподобием модели $\mathbf{f}$ назовем следующее выражение:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) = \int_{\mathbf{w}, \boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}|\mathbf{A}) p(\boldsymbol{\Gamma}|\mathbf{m}, c_{\text{temp}}) d\mathbf{w} d\boldsymbol{\Gamma}. \tag{2}$$

Требуется найти гиперпараметры модели $\mathbf{A}, \mathbf{m}$ доставляющие максимум правдоподобия модели:

$$\arg\max_{\mathbf{A}, \mathbf{m}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}), \tag{3}$$

а также соответствующие параметры и структуру модели:

$$\arg\max_{\mathbf{W}, \boldsymbol{\Gamma}} p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}). \tag{4}$$

Докажем теорему об оптимальности решения задачи (4), лежащего на вершинах симплексов $\times_{(i,j)\in E}\Delta^{K^{i,j}-1}$. Обозначим за $\bar{\Delta}^K$ — множество вершин K-мерного симплекса.

Теорема.
Пусть $\mathbf{\Gamma}_1$ и $\mathbf{\Gamma}_2$ — реализации $\mathbf{\Gamma}$, такие что:

- $\mathbf{\Gamma}_1 \in \times_{(i,j)\in E}\bar{\Delta}^{K^{i,j}-1}$.
- $\mathbf{\Gamma}_2 \notin \times_{(i,j)\in E}\bar{\Delta}^{K^{i,j}-1}$.

Тогда для любых положительно определенных матриц $\mathbf{A}_1$ и $\mathbf{A}_2$ и векторов $\mathbf{m}_1, \mathbf{m}_2, \min(\mathbf{m}_1) > 0$ справедлива следующая формула:

$$\lim_{c_{\text{temp}}\to 0} \frac{p(\mathbf{\Gamma}_2|\mathbf{y}, \mathbf{W}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}_1, c_{\text{temp}})}{p(\mathbf{\Gamma}_1|\mathbf{y}, \mathbf{W}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}_1, c_{\text{temp}})} = \infty.$$

Доказательство.
По теореме из оригинальной статьи

$$p(\lim_{c_{\text{temp}}\to 0} \gamma_k^{i,j} = 1) = m_k^{i,j}.$$

Тогда:

$$p(\lim_{c_{\text{temp}}\to 0} \boldsymbol{\gamma}^{i,j} \in \bar{\Delta}^{K^{i,j}}) = 1.$$

Тогда апостериорная вероятность $\mathbf{\Gamma}$: в пределе равняется нулю, если структура не лежит на произведении вершин симплекса.

$$p(\mathbf{\Gamma}_2|\mathbf{y}, \mathbf{W}, \mathbf{X}, \mathbf{A}_2, \mathbf{m}_2, c_{\text{temp}}) \propto p(\mathbf{\Gamma})p(\mathbf{y}|\mathbf{\Gamma}, \mathbf{W}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}) \to 0,$$
$$p(\mathbf{\Gamma}_1|\mathbf{y}, \mathbf{W}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}_1, c_{\text{temp}}) \propto p(\mathbf{\Gamma})p(\mathbf{y}|\mathbf{\Gamma}, \mathbf{W}, \mathbf{X}, \mathbf{A}_1, \mathbf{m}) \to C,$$

где C — константа, большая нуля, т.к. $\min(\mathbf{m}_1) > 0$. что и требовалось доказать.

TODO: еще бы хотелось расписать, что гамма должна в дискретном случае концентрироваться на одной вершине, но пока непонятно как сформулировать.

3 Вариационный вывод

В общем виде получение значения интеграла (2) является вычислительно сложной процедурой. В качестве приближенного значения интеграла будем использовать вариационную верхнюю оценку правдоподобия модели.

Пусть задано непрерывное параметрические распределение q, аппроксимирующие апостериорные распределение $p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$. Тогда верно следующее выражение:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) \geq \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) - D_{KL}(q||p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}, \mathbf{m}, c_{\text{temp}})) = \tag{5}$$

$$= \log_q p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}).$$

Разница между верхней оценкой (5) и правдоподобием модели (2) определяется дивергенцией между вариацоинным распределение q и апостериорным распределением $p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$.

Определим вариацонное распределение q следующим образом. Декомпозируем q на два распределения:

$$q = q_{\mathbf{W}} q_{\mathbf{\Gamma}} : q_{\mathbf{W}} \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad q_{\mathbf{\Gamma}} = \prod_{(i,j)\in E} q_\gamma^{i,j}, \quad q_\gamma \sim \mathcal{GS}(\mathbf{m}_q^{i,j}, c_q).$$

В дальнейшем будем обозначать за $\mathbf{m}_q$ конкатенацию всех векторов средних $\mathbf{m}_q^{i,j}$.

Для получения значения $\log_q p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$ будем использовать следующие методы сэмплирования:

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) \approx \frac{1}{N} \sum_{u=1}^{N} \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{W}}_u, \hat{\mathbf{\Gamma}}_u, \mathbf{A}, \mathbf{m}, c_{\text{temp}}),$$

$$D_{KL}(q||p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}, \mathbf{m}, c_{\text{temp}}) = D_{KL}(q_{\mathbf{\Gamma}}||p(\mathbf{\Gamma}|\mathbf{m}, c_{\text{temp}}) + D_{KL}(q_{\mathbf{W}||p(\mathbf{W}|\mathbf{A}}) \approx$$

$$\frac{1}{N} \sum_{u=1}^{N} (\log\, q_{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}_u}) - \log p(\hat{\mathbf{\Gamma}_u}) + 0.5(\text{tr}(\hat{\mathbf{S}}_u(\mathbf{W})\mathbf{A}\mathbf{A}_q^{-1}) +$$

$$+ \boldsymbol{\mu}^{\mathrm{T}} \hat{\mathbf{S}}_u(\mathbf{W})\mathbf{A}\boldsymbol{\mu} - |\mathbf{W}| + \log\, \det \hat{\mathbf{S}}_u(\mathbf{W})\mathbf{A} - \log\, \det \mathbf{A}_q)),$$

где N — количество реализаций случайных величин, $\hat{\mathbf{\Gamma}}_u, \hat{\mathbf{W}}_u$ — реализации случайных величин, $\hat{S_u}(\mathbf{W})$ — соответствие между параметрами и реализацией весов базовых функций.

Сэмплирование происходит следующим образом:

$$\hat{\mathbf{W}} = \boldsymbol{\mu} + \varepsilon \mathbf{A}_q, \quad \varepsilon \in \mathcal{N}(\mathbf{0}, \mathbf{1}),$$

$$\hat{\boldsymbol{\gamma}}_k = \frac{\exp((\log m_k + a_k)/c)}{\sum_{i=1}^{K} (\log m_i + a_i)/c)}, \quad \mathbf{a} \in -\log(\log(\mathcal{U}(0,1)^K).$$

Численную оценку, полученную описанным выше способом будет обозначать как

$$\hat{\log}_q p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) = \hat{\mathsf{E}}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) - \hat{D_{KL}}(q||p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}, \mathbf{m}, c_{\text{temp}})).$$

(Возможно нужно доказать корректность оценки).

Докажем теорему о дискретности задачи оптимизации вариационной оценки в предельном случае.

Теорема. Пусть $c = c_{\text{temp}}$. Для любых значений ковариационных матриц $\mathbf{A}, \mathbf{A}_q$, любого вектора $\boldsymbol{\mu}_q$ существуют такие точки $\mathbf{m}_q^1 \in \times_{(i,j)\in E} \bar{\Delta}^{K^{i,j}-1}, \mathbf{m}^1 \in \times_{(i,j)\in E} \bar{\Delta}^{K^{i,j}-1}$

на вершинах симплексов структуры $\mathbf{\Gamma}$, что для любой точки $\mathbf{m}_q^2 \in \times_{(i,j)\in E}\Delta^{K^{i,j}-1}$ и $\mathbf{m}^2 \in \times_{(i,j)\in E}\Delta^{K^{i,j}-1}$ внутри симплексов справедливо выражение:

$$\lim_{c_{\text{temp}}\to 0}\frac{\log\hat{p}_{q_{\mathbf{W}}q_{\mathbf{\Gamma}}^2}(\mathbf{y}|\mathbf{X})}{\log\hat{p}_{q_{\mathbf{W}}q_{\mathbf{\Gamma}}}(\mathbf{y}|\mathbf{X})} \geq 1, \quad \text{где} q_{\mathbf{\Gamma}}^1 = \max_c q_{\mathbf{\Gamma}}(\mathbf{m}_q^1, c).$$

Доказательство. По свойству предельного распределения $\mathcal{GS}$ задача сводится к задаче с сингулярным распределением на структурах. Расписав $\log_q p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$ через двойную сумму находим максимальный элемент.

3.1 Общая постановка задачи

Определим основные величины, которые характеризуют сложность модели.

Определение Параметрической сложностью $C_{\mathbf{W}}$ модели назовем наименьшую дивергенцию вариационных параметров при условии априорного распределения параметров:

$$C_{\mathbf{W}} = \arg\min_{\mathbf{A}} D_{\text{KL}}\left(q|p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}, \mathbf{m}, \boldsymbol{c}_{\text{temp}})\right).$$

Определение Структурной сложностью $C_{\mathbf{\Gamma}}$ модели назовем энтропию распределения структуры:

$$C_{\mathbf{\Gamma}} = -\mathsf{E}_{q_{\mathbf{\Gamma}}}\log q_{\mathbf{\Gamma}}.$$

Сформулируем основные требования к оптимизационной задаче и оптимизируемым функционалам:

1. Оптимизируемые функции должны быть дифференцируемы.
2. Оптимизация должна позволять проводить максимизацию апостериорной вероятности правдоподобия.
3. Степень регуляризации структуры $\mathbf{\Gamma}$ и параметров $\mathbf{W}$ должна быть контролируемой.
4. Оптимизация должна приводить к максимуму вариационной оценки $\log_q p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$.
5. Оптимизация должна позволять калибровать параметрическую сложность модели $C_{\mathbf{W}}$.
6. Оптимизация должна позволять калибровать структурную сложность $\mathbf{\Gamma}$ модели.
7. Оптимизация должна позволять проводить полный перебор структуры $\mathbf{\Gamma}$.

Сформулируем задачу как двухуровневую задачу оптимизации. Обозначим за $\boldsymbol{\theta}$ оптимизируемые на первом уровне величины. Обозначим за $\mathbf{h}$ величины, оптимизируемые на втором уровне. Положим $\boldsymbol{\theta}$ равным параметрам распределений $q_{\mathbf{W}}, q_{\boldsymbol{\Gamma}} : \boldsymbol{\theta} = [\boldsymbol{\mu}_q, \mathbf{A}_q, \mathbf{m}_q, c]^{\mathrm{T}}$. Положим $\mathbf{h} = [\mathbf{A}, \mathbf{m}]$.

Обозначим за L функцию потерь:

$$L = c_{\text{reg}}\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) - D_{KL}(q_{\boldsymbol{\Gamma}}||p(\boldsymbol{\Gamma})) - D_{KL}(q_{\mathbf{W}}||p(\mathbf{w})), \qquad (6)$$

где c_{reg} — коэффициент регуляризации регуляризации структуры $\boldsymbol{\Gamma}$ и параметров $\mathbf{W}$ априорным распределением.

Лемма. Пусть $\mathbf{A}_q$ фиксирована и близка к нулю, $c_{\text{reg}} = 1$. Тогда максимизация L эквивалентна оптимизации апостериорной вероятности параметров при $c \to 0$.

Доказательство. $L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) - D_{KL}(q||p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{A}, \mathbf{m}, c_{\text{temp}}))$. Полагая ковариационную матрицу близкой к нулю $\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) \approx \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\mu}_q, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$.

$$D_{KL}(q||p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{A}, \mathbf{m}, c_{\text{temp}}) =$$

$$\frac{1}{N}\sum_{u=1}^{N}(\log\, q_{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\Gamma}}_u) - \log p(\hat{\boldsymbol{\Gamma}}_u) + 0.5(\boldsymbol{\mu}^{\mathrm{T}}\hat{\mathbf{S}}_u(\mathbf{W})\mathbf{A}\boldsymbol{\mu} - |\mathbf{W}| + \log\, \det\hat{\mathbf{S}}_u(\mathbf{W})\mathbf{A}).$$

Следующая теорема говорит о том, что варьируя c_{reg} мы проводим оптимизацию, ассимптотически аналогичную оптимизации выборки из того же распределения, но другой мощности.

Теорема. Пусть $c_{\text{reg}} > 0$, $c_{\text{reg}} m \in \mathbb{N}$. Тогда функция L сходится почти наверно к вариационной нижней оценке правдоподобия для произвольной подвыборки $\mathfrak{D}$ мощностью $m_0 = \frac{m}{c_{\text{reg}}}$, поделенной на данную константу.

Доказательство. Рассмотрим произвольную подвыборку $\hat{\mathfrak{D}}$ мощностью m_0. Нижняя оценка правдоподобия модели для подвыборки имеет вид:

$$\mathsf{E}_{q_w, q_\gamma} \log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c) - D_{KL}(q_\gamma||p(\boldsymbol{\Gamma})) - D_{KL}(q_w||p(\mathbf{w})).$$

$$\log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c) = \sum_i \log p(\hat{\mathbf{y}}_i|\hat{\mathbf{x}}_i, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c) \to^{\text{п.н.}}_{m\to\infty} m_0 \mathsf{E} \log p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c).$$

Таким образом, ассимптотическая формула вариационной нижней оценки правдоподобия для подвыборки мощностью m_0 выглядит следующим образом:

$$m_0 \mathsf{E} \log p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{A}, \mathbf{m}, c) - D_{KL}(q_\gamma||p(\boldsymbol{\Gamma})) - D_{KL}(q_w||p(\mathbf{w})).$$

Домножив на выражение на $\frac{m}{m_0}$ получаем ассимптотику для L, что и требовалось доказать.

Пусть Q — валидационная функция:

$$Q = c_{\text{train}}\mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}}\mathrm{D}_{KL}(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}, \mathbf{\Gamma}))-$$

$$c_{\text{comb}} \sum_{p' \in \mathbf{P}} \mathrm{D}_{KL}(\mathbf{\Gamma}|p') \to \max,$$

где $\mathbf{P}$ — множество (возможно пустое) распределений на структуре модели, c_{prior} — коэффициент регуляризации параметрической сложности модели, c_{comb} — коэффициент перебора.

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h} \in \mathbb{R}^h} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})), \tag{7}$$

где T — оператор оптимизации, решающий задачу оптимизации:

$$L(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})) \to \max.$$

Теорема. Пусть $D_{KL}(q_w|p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0, D_{KL}(q_\gamma|p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0$, пусть $c_{\text{prior}} = 1, c_{\text{reg}} = 1, c_{\text{comb}} = 0$. Тогда оптимизация (7) эквивалентна оптимизации (2).

Доказательство. При соблюдении условий теоремы оптимизация вариационной оценки эквивалента оптимизации правдоподобия модели. При $c_{\text{prior}} = 1, c_{\text{reg}} = 1, c_{\text{comb}} = 0$, функция Q становится равной вариационной нижней оценке. Таким образом, двухуровневая оптимизация становится эквивалентной оптимизации правдоподобия модели по $\mathbf{A}, \mathbf{m}$, что и требовалось доказать.

3.2 О параметрической сложности

Обозначим за $F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, c_{\text{comb}}, \mathbf{P}, c_{\text{temp}})$ множество экстремумов функции L при решении задачи двухуровневой оптимизации.

Теорема. Пусть $\mathbf{f}_1 \in F(1, 1, c^1_{\text{prior}}, 0, \varnothing, c_{\text{temp}}), \mathbf{f}_2 \in F(1, 1, c^2_{\text{prior}}, 0, \varnothing, c_{\text{temp}}),\ c^1_{\text{prior}} < c^2_{\text{prior}}$.
Пусть вариационные параметры моделей $\mathbf{f}_1$ и $\mathbf{f}_2$ лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми.
Тогда модель $\mathbf{f}_1$ имеет параметрическую сложность, не меньшую чем у $\mathbf{f}_2$.

$$C_{\text{param}}(\mathbf{f}_1) \geq C_{\text{param}}(\mathbf{f}_2).$$

Доказательство. Обозначим за q^1, q^2 — вариационные распределения моделей $\mathbf{f}_1, \mathbf{f}_2$, p^1, p^2 — априорные распределения моделей.

Отсюда справедливы следующие неравенства (по единственности точек экстремума L, Q):

$$\mathsf{E}_{q^1} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) - c^1_{\text{prior}} D_{\text{KL}}(q^1||p^1) - \mathsf{E}_{q^2} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) + c^1_{\text{prior}} D_{\text{KL}}(q^2||p^2) \geq 0,$$

$\mathsf{E}_{q^2}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) - c^2_{\text{prior}} D_{\text{KL}}(q^2||p^2) - \mathsf{E}_{q^1}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) + c^2_{\text{prior}} D_{\text{KL}}(q^1||p^1) \geq 0.$

Складывая неравенства получим:

$$D_{\text{KL}}(q^1||p^1) \geq D_{\text{KL}}(q^2||p^2),$$

$$\mathsf{E}_{q^2}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) \leq \mathsf{E}_{q^1}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c).$$

С учетом полученных неравенств распишем доказываемое утверждение:

$$\max_p \left(-D_{\text{KL}}(q^1||p)\right) - \max_p \left(-D_{\text{KL}}(q^2||p^2)\right) =$$

$$\max_p \left(-c^2_{\text{prior}} D_{\text{KL}}(q^1||p) + \mathsf{E}_{q^1}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) - \mathsf{E}_{q^1}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c)\right) -$$

$$- \max_p \left(-c^2_{\text{prior}} D_{\text{KL}}(q^2||p) + \mathsf{E}_{q^2}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) + \mathsf{E}_{q^2}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c)\right) \leq 0,$$

что и т.д.

Теорема. Пусть $\mathbf{f} \in F(1, 1, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. При устремлении c_{prior} к бесконечности параметрическая сложность модели $\mathbf{f}$ устремляется к нулю (или существтует?):

$$\lim_{c_{\text{prior}} \to \infty} C_{\text{param}}(\mathbf{f}) = 0.$$

Доказательство
В пределе: $Q = D_{KL}$.
Минимум достигается при совпадении параметров распределений: $mu = 0$.
Докажем существование решения L, которое удовлетворяет этому.
Рассмотрим значение L при $A \to 0$. Два случая: либо конечное значение, либо бесконечное.
Таким образом, калибруя A получаем значения, близкие к нулю.
Рассмотрим последовательность. Тогда lim inf ->0.
Доказано.

3.3 О структурной сложности

Теорема Пусть для каждого ребра (i, j) семейства моделей $\mathfrak{F}$ априорное распределение

$$p(\boldsymbol{\gamma}_{i,j}) = \lim_{c_{\text{temp}} \to 0} \mathcal{GS}(c_{\text{temp}}).$$

Пусть $c_{\text{reg}} > 0, c_{\text{train}} > 0, c_{\text{prior}} > 0$. Пусть $\mathbf{f} \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. Тогда структурная сложность модели $\mathbf{f}$ равняется нулю.

$$C_{\text{struct}}(\mathbf{f}) = 0.$$

Доказательство 1. Доказываем, что гипер-концентрация будет лежать на вершине
2. У нас получается, что D_{KL} будет конечным только в случае совпадения.(???) 3. Итого, получили.

Теорема Пусть $\mathbf{f}_1 \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^1_{\text{temp}}), \mathbf{f}_2 \in \lim_{c^2_{\text{temp}} \to \infty} F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^2_{\text{temp}})$. Пусть вариационные параметры моделей f_1 и f_2 лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда разница структурных сложностей моделей ограничена выражением:

$$C_{\text{struct}}(\mathbf{f}_1) - C_{\text{struct}}(\mathbf{f}_2) \leq \mathsf{E}^1_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}.\mathbf{A}^{-1}, c^1_{\text{temp}}) - \mathsf{E}^2_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{A}^{-1}).$$

Доказательство 0. Доказываем равномерную сходимость.
1. расписываем неравенства вида: $L_1 - DKL(q_1|p1) < L_2 - DKL(q_2|p1)$
2. Замечаем, что при стремлении к бесконечности гумбель превращается в равномерное
3. выражаем все в равномерном
4. замечаем, что $D_K L = Entropy + const$ для равномерного

3.4 О переборе вариантов

Утверждение (очень предварительно). Изменение c позволяет избежать ухода в локальный минимум.

Утверждение (очень предварительно). Изменение c_2 позволяет избежать ухода в локальный минимум.

Утверждение (очень предварительно). Взаимосвязь структуры и параметров в prior позволяет получить «хорошие» модели.

Утверждение (предварительно). Пусть $c_1 = c_2 = c_3 = 0$. Пусть $q_w \sim \mathcal{N}(\mathbf{0}, \sigma), \sigma \sim 0$. Тогда оптимизация эквивалентна обычной оптимизации параметров с l_2 - регуляризацией.

3.5 Общая теорема

4 Вариационная постановка задачи

5 Вычислительный эксперимент

В качестве модельного эксперимента рассматривалась задача выбора модели линейной регрессии. Множество объектов $\mathbf{X}$ было сгенерировано из трехмерного стандартного распределения:

$$\mathbf{X} \sim \mathcal{N}(0, \mathbf{I}), n = 3.$$

Множество меток было определено следующим правилом:

$$\mathbf{y} = \arg\max_{0,1}(\mathbf{X}_1 + \mathbf{X}_2),$$

третья компонента не участвовала в генерации ответа.

Рассматривались четыре возможные структуры:

1. $f_1 = \mathbf{w}_1\mathbf{X}_1$ (модель — регрессия только по первому признаку),
2. $f_2 = \mathbf{w}_2\mathbf{X}_2$ (модель — регрессия только по первому признаку),
3. $f_3 = \mathbf{w}_3\mathbf{X}_3$ (модель — регрессия только по шумовому признаку),
4. $f_4 = \mathbf{w}_4\mathbf{X}$ (модель — регрессия по всем признакам).

Ожидаемое поведение оптимизации:

1. При $c_1 = c_2 = 1c \sim 0$ (Evidence с низкой температурой) будет произведен выбор структуры f_4.
2. При $c_1 = c_2 = 1, c >> 0$ (Evidence с высокой температурой) будет произведен выбор двух структур с одинаковым весом: f_1, f_2.

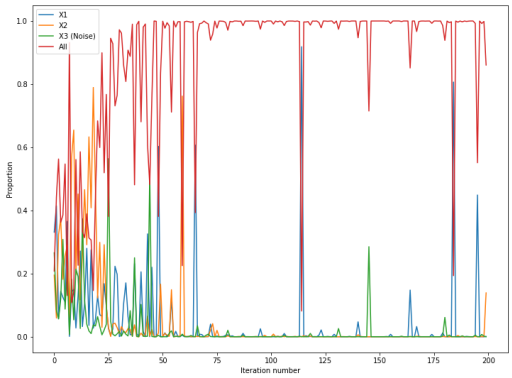

Figure 1: Evidence с низкой температурой

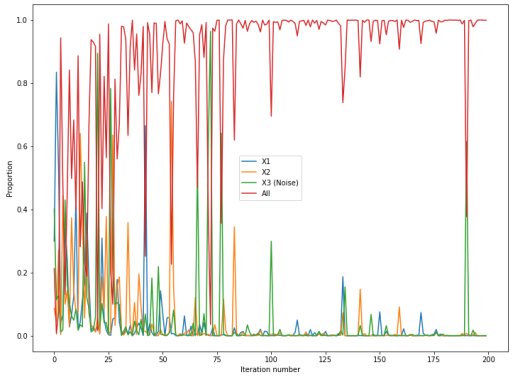

Figure 2: Evidence с высокой температурой

3. При $c_1 = c_2 = 0, c_3 = 1, \mathbf{p} = [[0.0, 0.0, 1.0, 0.0]], c \sim 0$ (Поощряется выбор структуры с шумовой компонентой) будет произведен выбор структуры f_4, при снижении параметра c_{reg} выбор будет меняться в сторону f_3.

Результаты

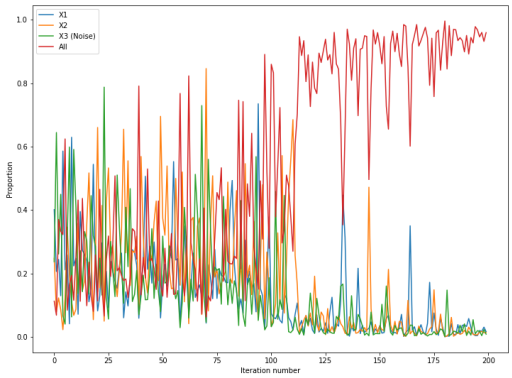

Figure 3: Evidence с высокой температурой, $\beta = 0.01$

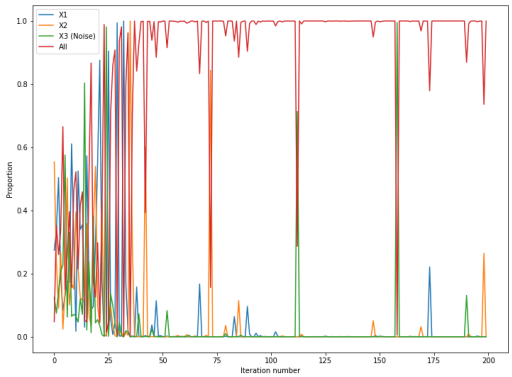

Figure 4: Поощрение выбора шумовой компоненты

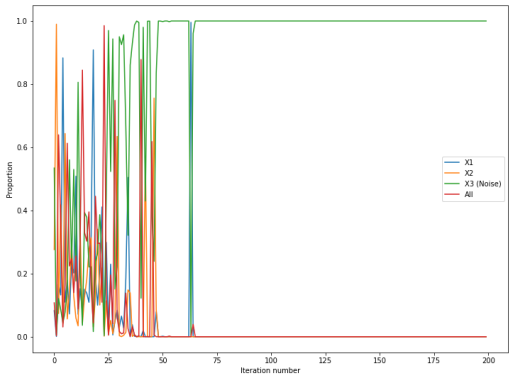

Figure 5: Поощрение выбора шумовой компоненты, $\beta = 0.01$