# 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 объекты $\mathbf{x}$ 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 множество функций $\mathbf{g}^{i,j}$ мощности $K^{i,j}$. Вершины $V$ — промежуточные промежуточные представления выборки под действием данных функций.

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

$$f_j(\mathbf{x}) = \sum_{k \in \mathrm{Adj}(v_j)} < \gamma^{j,k}, \mathbf{g}^{j,k} > (\mathbf{f}_k(\mathbf{x})]), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x}$$

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

Параметрами модели $\mathbf{W}$ будем называть конкатенацию всех параметров подмоделей $\mathbf{f}_j$. Структурой модели $\mathbf{\Gamma}$ будем называть конкатенацию всех структурных параметров $\gamma^{j,k}$. Моделью будем называть совокупность параметров $\mathbf{W}$ и гиперпараметров $\mathbf{W}$.

Пусть все векторы $\gamma^{i,j}$ являются нормированными и положительными. Пусть для каждого структурного параметра $\gamma^{j,k} \in \mathbf{\Gamma}$ определено априорное Gumbel-softmax распределение $p(\gamma^{j,k}|\mathbf{m}^{j,k}, c_{\text{temp}})$ с параметром средних $\mathbf{m}$ и температурой $c_{\text{temp}}$.

Пусть для структуры модели определено априорное распределение $p(\mathbf{\Gamma}|\mathbf{m})$, где $\mathbf{m}$ — некоторое распределение. *(нужно ли определять его явно здесь?)* Пусть для каждого параметра $w \in \mathbf{W}$ определено множество $\mathcal{S}(w)$ структурных параметров $\gamma$, соответсвтующих базовым функциям $\mathbf{g}$, для которых определен этот параметр:

$$w \sim \mathcal{N}(\mathbf{0}, a^{-1} \cdot (\sum_{\gamma \in \mathcal{S}(w)} \gamma),$$

где $a$ — гиперпараметр, входящий в диагональную матрицу $\mathbf{A}^{-1}$. Пусть также определено правдоподобие выборки $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})$.

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

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

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 соотвествующие параметры и структур модели (см. вывод Байесва, первый уровень).

Докажем теорему о дискретности решиния задачи нахождения оптимальных параметров модели.

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

- $\mathbf{\Gamma}_1$ не содержит в себе точки внутри симплексов $\gamma$.

- $\mathbf{\Gamma}_2$ содержит в себе точки внутри симплексов $\gamma$.

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

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

**Доказательство**. По теореме из оригинальной статьи

$$p(\lim_{c_{\text{temp}} \to 0} \mathbf{m} \text{ лежит на вершинах произведения симплексов}) = 1.$$

Тогда апостериорная вероятность $\mathbf{\Gamma}$ :

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

будет стремиться к нулю при наличии точке вутри симплексов. 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 вычисление 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:

$$\log p(\mathbf{y} | \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) \geq \mathsf{E}_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c_{\text{temp}} - D_{KL}(q || p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})) = \tag{4}$$

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

Разница между верхней оценкой (4) и правдоподобием модели (2) определяется дивергенцией между вариацоинным распределение $q$ и апостериорным распределением $p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})$.

В дальнейшем будем использовать следующую форму вариационного распределения:

$$q = q_{\mathbf{W}} q_{\mathbf{\Gamma}} :$$

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

В дальнейшем будем обозначать за $\mathbf{m}_q$ конкатенацию всех векторов средних $\mathbf{m}^{j,k}$. TODO: расписать Монте-Карло.

Докажем теорему о дискретности задачи, аналогичную первой теореме. **Лемма** При устремлении температуры $c_{\text{temp}}$ к нулю $D_K L$ стремится к плюс бесконечности, для $c! = 0$

**Доказательство** Очевидно, расписать.

**Теорема** Для любых значений ковариационных матриц $\mathbf{A}, \mathbf{A}_q$, любого вектора $\boldsymbol{\mu}_q$ существуют такие точка $\mathbf{m}_q^1, \mathbf{m}^1$ на вершинах симплексов структуры $\boldsymbol{\Gamma}$, что для любой точки $\mathbf{m}_q^2$ и $\mathbf{m}^2$ внутри симплексов справедливо выражение:

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

**Доказательство** По лемме, требуется рассматривать только $c = 0$. Для $D_{\text{KL}}$ оптимум — совпадение распределений. Выберем точку, которая будет соответствовать главной структуре.

## 3.1 Общая постановка задачи

Определим основные величины, которые характеризуют сложность модели.

**Определение** Параметрической сложностью $C_w$ модели назовем наименьшую дивергенцию вариационныъ параметров при условии априорного распределения параметров:

$$C_w = \arg\min_{\mathbf{A}} D_{\text{KL}}(q|p).$$

(Примечание: кажется, здесь должны учитываться параметры и структура, т.к. в априорном распределении параметров зашита зависимость от структуры).

**Определение** Структурной сложностью $C_\gamma$ модели назовем энтропию распределения структуры:

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

Сформулируем основные требования к оптимизационной задаче и оптимизируемым функционалам:

1. Оптимизируемые функции должны быть дифференцируемы.
2. Оптимизация должна позволять проводить простое обучение модели.
3. Степень регуляризации структуры и параметров должна быть контролируемой.
4. Оптимизация должна приводить к максимуму вариационной оценки.
5. Оптимизация должна позволять калибровать параметрическую сложность модели
6. Оптимизация должна позволять калибровать структурную сложность модели.
7. Оптимизация должна позволять проводить полный перебор структуры.

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 $q_w, q_\gamma$. Положим $\mathbf{h} = [\mathbf{A}, \mathbf{m}]$.

Пусть $L$ — приближенное значение вариационной оценки правдоподобия:

$$L = c_{\text{reg}} \log p(\mathbf{y}|\hat{\mathbf{w}}, \hat{\boldsymbol{\Gamma}}) - D_{KL}(q_\gamma || p(\boldsymbol{\Gamma})) - D_{KL}(q_w || p(\mathbf{w})), \tag{5}$$

**Лемма.** Пусть $\mathbf{A}_q$ фиксированна и близка к нулю. Тогда оптимизация $L$ эквивалентна простой оптимизации с $l$-2 регуляризацией.

**Доказательство** расписать.

где $\hat{\mathbf{w}} \sim q_w$, $\hat{\boldsymbol{\Gamma}} \sim q_\gamma$, $c_{\text{reg}}$ — коэффициент, контролирующий степень регуляризации структуры и параметров.

Следующая теорема говорит о том, что калибруя $c_{\text{reg}}$ 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 подвыборки $\mathfrak{D}$ мощностью $c_{\text{reg}} m$, разделенной на данную константу.

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

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

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

Формула становится эквивалентна форумле $L$ с точностью до множителя, что и т.д.

Пусть $Q$ — валидационная функция:

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

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

где $\mathbf{P}$ — множество (возможно пустое) распределений на структуре модели.

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h} \in \mathbb{R}^h} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})), \tag{6}$$

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 $c_1 = 1, c_2 = 1, c_3 = 0$. Тогда оптимизация (6) эквивалентна оптимизации (2).

**Доказательство.** При соблюдении условий теоремы неравенство вариационной оценки превращается в равенство.

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

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

**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.

**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 $\mathsf{U}$, 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.** Заметим, что $q$ и $q'$ можно выразить следующим образом:

$$q = \arg \max_{\hat{q}: L(\hat{q}, \mathbf{A}, \mathbf{\Gamma}) = \max} Q(\mathbf{A}, \mathbf{\Gamma}, c_2),$$

$$q' = \arg\max_{\hat{q}: L(\hat{q}, \mathbf{A}, \mathbf{\Gamma}) = \max} Q(\mathbf{A}, \mathbf{\Gamma}, c_2').$$

Функция $Q$ является выпуклой как сумма выпуклых функций. Отсюда справедливы следующие неравенства (по единственности точек экстремума):

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) - c_2 D_{\text{KL}}(q||p) - \mathsf{E}_{q'} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) + c_2 D_{\text{KL}}(q'||p') \geq 0,$$

$$\mathsf{E}_{q'}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) - c_2' D_{\mathrm{KL}}(q'||p') - \mathsf{E}_q\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) + c_2' D_{\mathrm{KL}}(q||p) \geq 0.$$

Вычитая неравенства получим:

$$D_{\mathrm{KL}}(q||p) \geq D_{\mathrm{KL}}(q'||p'),$$

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

С учетом полученных неравенств распишем доказываемое утверждение:

$$\max_p -D_{\mathrm{KL}}(q||p) - \max_{p'} -D_{\mathrm{KL}}(q'||p') \propto$$

$$\propto \max_p -c_2' D_{\mathrm{KL}}(q||p) + \mathsf{E}_q\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) - \mathsf{E}_q\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) -$$

$$- \max_{p'} -c_2' D_{\mathrm{KL}}(q'||p') + \mathsf{E}_{q'}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) + \mathsf{E}_{q'}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{\Gamma},\mathbf{A},\mathbf{m},c) \leq 0,$$

что и т.д.

## 3.3 О структурной сложности

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

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

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

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

**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 $\mathbf{f}_1 \in F(c_{\mathrm{reg}}, c_{\mathrm{train}}, c_{\mathrm{prior}}, 0, \varnothing, c^1_{\mathrm{temp}}), \mathbf{f}_2 \in \lim_{c^2_{\mathrm{temp}}\to\infty} F(c_{\mathrm{reg}}, c_{\mathrm{train}}, c_{\mathrm{prior}}, 0, \varnothing, c^2_{\mathrm{temp}})$. 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:

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

**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: $f_1, f_2$.

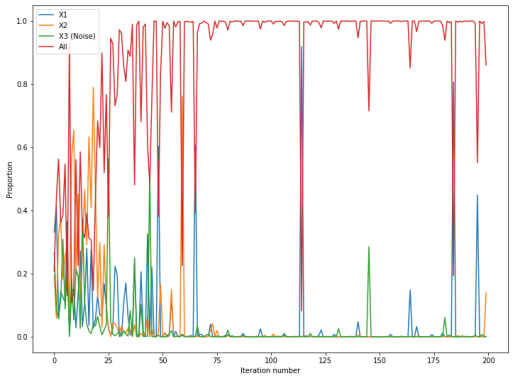


Figure 1: Evidence с низкой температурой

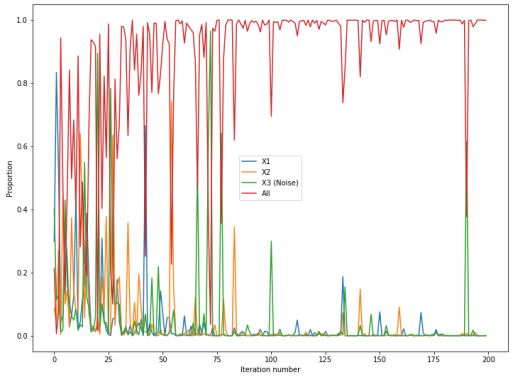


Figure 2: Evidence с высокой температурой

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 $c_{\text{reg}}$ выбор будет меняться в сторону $f_3$.

**Результаты**

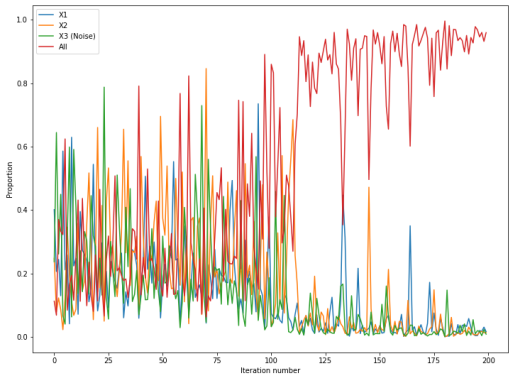


Figure 3: Evidence с высокой температурой, $\beta = 0.01$

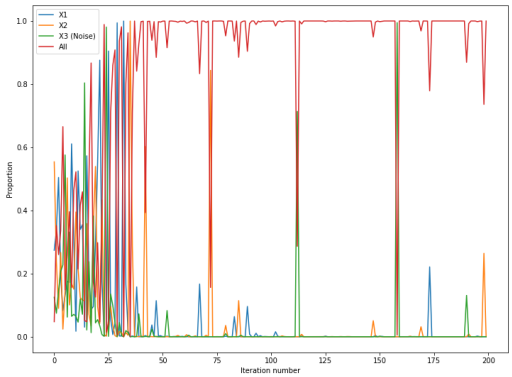


Figure 4: Поощрение выбора шумовой компоненты

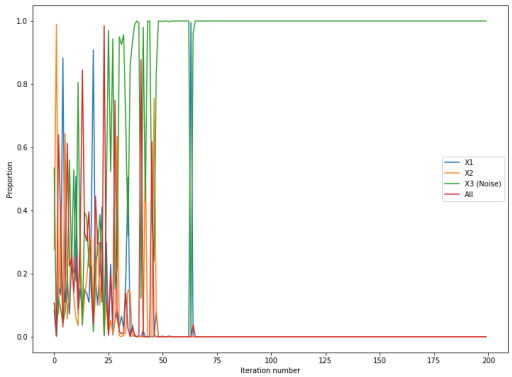


Figure 5: Поощрение выбора шумовой компоненты, $\beta = 0.01$