My first scientific paper

Week 6

Write the theory: axioms → theorems, methods → algorithms

Vadim Strijov

Moscow Institute of Physics and Technology

2022

Как в моей дипломной работе найти теорему?

Борхес Х. Л. «Аналитический язык Джона Уилкинса»[1]

Животные делятся на:

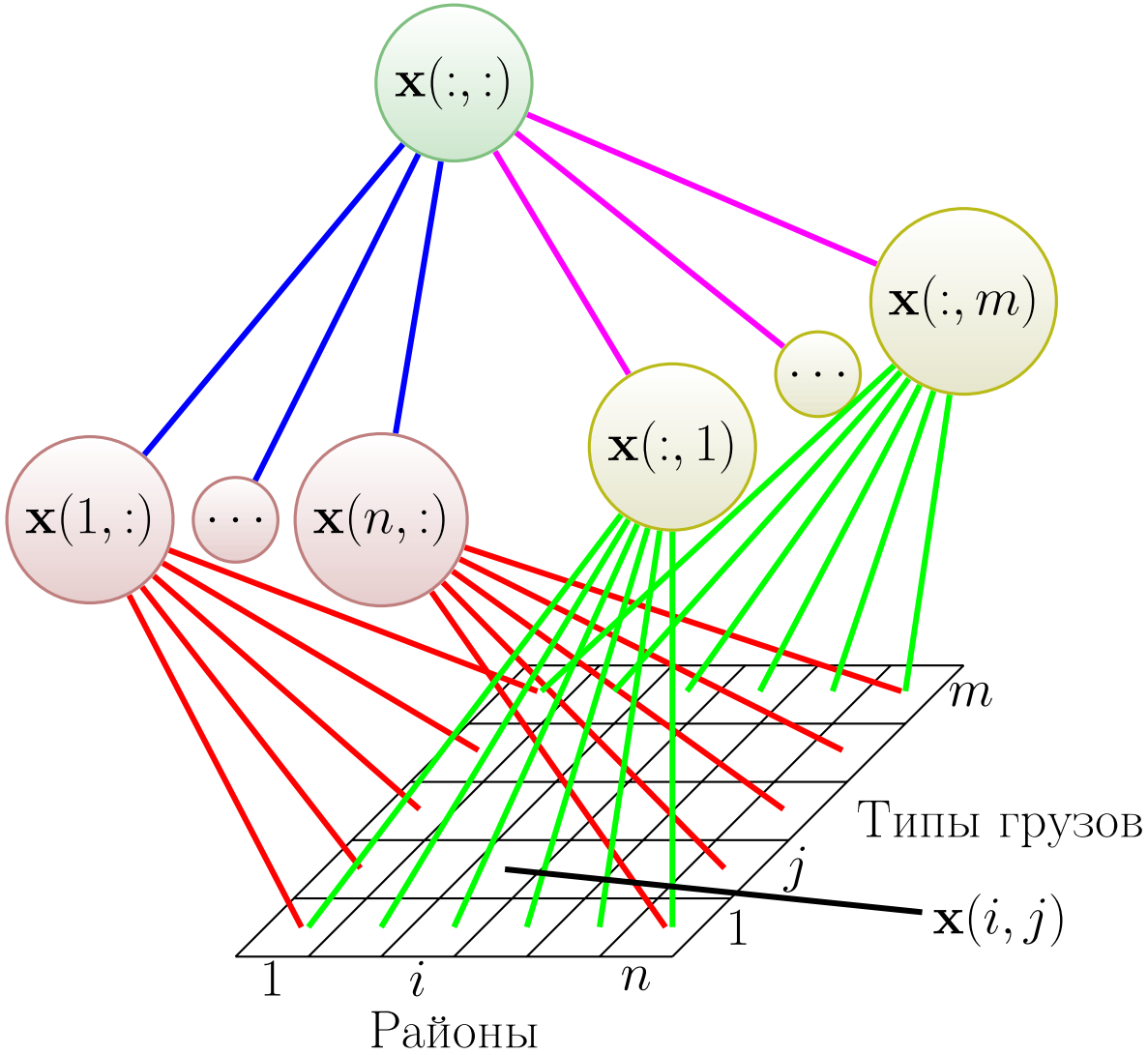
a) принадлежащих Императору,
b) набальзамированных,
c) прирученных,
d) молочных поросят,
e) сирен,
f) сказочных,
g) бродячих собак,
h) включённых в эту классификацию,
i) бегающих как сумасшедшие,
j) бесчисленных,
k) нарисованных тончайшей кистью из верблюжьей шерсти,
l) прочих,
m) разбивших цветочную вазу,
n) похожих издали на мух.

[1] Китайская энциклопедия «Божественное хранилище благотворных знаний»

Что доказываем?

a) существование,
b) единственность,
c) эквивалентность множеств или отображений, и даже решений,
d) сходимость,
e) корректность по Адамару,
f) сложность,
g) включая сложность $\mathcal{O}(n \ln n)$,
h) устойчивость не только по Ляпунову,
i) корректность статистического вывода,
j) и байесовского вывода,
k) состоятельность, несмещенность, эффективность,
l) оптимальность (в некотором смысле),
m) равновесность,
n) и даже сходимость.

Условие согласованности прогнозов



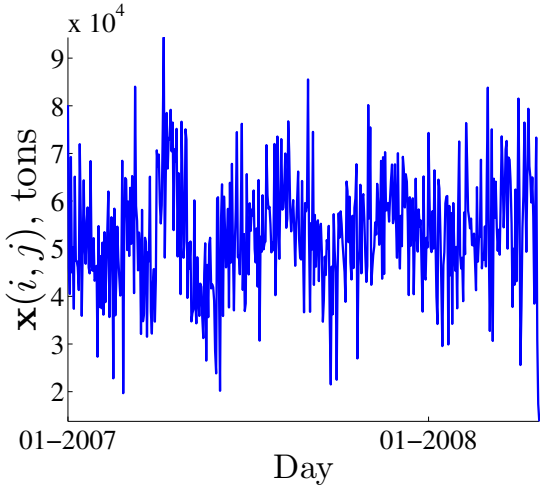
$$x_t(:,:) = \sum_{i=1}^{n} x_t(i,:);$$

$$x_t(:,:) = \sum_{j=1}^{m} x_t(:,j);$$

$$x_t(i,:) = \sum_{j=1}^{m} x_t(i,j), \quad i = 1,\dots n;$$

$$x_t(:,j) = \sum_{i=1}^{n} x_t(i,j), \quad j = 1,\dots m;$$

$$t = 1,\dots,T.$$



Постановка задачи согласования прогнозов

Дано

Матрица связей $\mathbf{S}$, множества $\mathcal{A}$, $\mathcal{B}$ и вектор независимых прогнозов $\hat{\chi}$

$$\hat{\chi} \notin \mathcal{A}, \quad \hat{\chi} \in \mathcal{B}.$$

Требуется построить

вектор согласованных прогнозов $\hat{\varphi}$, который удовлетворяет следующим требованиям:

- $\hat{\varphi} \in \mathcal{A}$, $\mathcal{A} = \{\chi \in \mathbb{R}^d \mid \mathbf{S}\chi = \mathbf{0}\}$ — согласованность;
- $\hat{\varphi} \in \mathcal{B}$ — физические ограничения;
- $l_h(\chi_{T+1}, \hat{\varphi}) \leq l_h(\chi_{T+1}, \hat{\chi})$ для любого среза действительных значений $\chi_{T+1} \in \mathcal{A} \cap \mathcal{B}$ — качество.

Согласование как антагонистическая игра

Игрок, выбирающий вектор согласованных прогнозов $\hat{\varphi}$, играет с природой, выбирающей срез иерархии в момент времени $(T+1)$. Цель игрока — минимизировать свои потери при любом ходе природы.

	Стратегия	Потери
Игрок	$\hat{\varphi} \in \mathcal{A} \cap \mathcal{B}$	$L(\hat{\varphi}, \chi_{T+1}) = I_h(\chi_{T+1}, \hat{\varphi}) - I_h(\chi_{T+1}, \hat{\chi})$
Природа	$\chi_{T+1} \in \mathcal{A} \cap \mathcal{B}$	$-L(\hat{\varphi}, \chi_{T+1})$

Равновесие Нэша в антагонистической игре — это

пара стратегий $(\hat{\varphi}, \chi_{T+1})$, таких что для любых стратегий $\hat{\varphi}'$, χ'_{T+1} выполнено неравенство

$$L(\hat{\varphi}, \chi'_{T+1}) \leq L(\hat{\varphi}, \chi_{T+1}) \leq L(\hat{\varphi}', \chi_{T+1}).$$

Цена игры (Дж. Нэш)

$$V = \min_{\hat{\varphi}} \max_{\chi_{T+1}} L(\hat{\varphi}, \chi_{T+1}) = \max_{\chi_{T+1}} \min_{\hat{\varphi}} L(\hat{\varphi}, \chi_{T+1})$$

определена тогда и только тогда, когда в игре существует равновесие Нэша.

Существование равновесия Нэша и выбор согласованных прогнозов

Теорема 1 (Стенина, 2014)

Пусть $\mathcal{A} \cap \mathcal{B} \neq \varnothing$ и для функции потерь l_h выполнено

1. $l_h(\chi_{T+1}, \hat{\chi}) \geq 0$ для произвольных векторов $\chi_{T+1},\ \hat{\chi}$, причем $l_h(\chi_{T+1}, \hat{\chi}) = 0 \quad \Leftrightarrow \quad \chi_{T+1} = \hat{\chi}$;
2. существует проекция $\chi_{proj} = \arg\min\limits_{\chi \in \mathcal{A} \cap \mathcal{B}} l_h(\chi, \hat{\chi})$;
3. для всех $\chi \in \mathcal{B}$ и для всех $\psi \in \mathcal{A} \cap \mathcal{B}$ выполняется неравенство $l_h(\psi, \chi) \geq l_h(\psi, \chi_{proj}) + l_h(\chi_{proj}, \chi)$.

Тогда

- пара стратегий $(\chi_{proj}, \chi_{proj})$ является равновесием Нэша в антагонистической игре, описывающей задачу согласования прогнозов;
- пара $(\chi_{proj}, \chi_{proj})$ является седловой точкой функции $L(\hat{\varphi}, \chi_{T+1}) = l_h(\chi_{T+1}, \hat{\varphi}) - l_h(\chi_{T+1}, \hat{\chi})$.

Оптимизационная задача

Теорема 2: цена игры (Стенина, 2014)

При выполнении требований теоремы 1 цена игры определена и равна

$$V = \min_{\hat{\varphi}} \max_{\boldsymbol{\chi_{T+1}}} L(\hat{\varphi}, \boldsymbol{\chi}_{T+1}) = \max_{\boldsymbol{\chi_{T+1}}} \min_{\hat{\varphi}} L(\hat{\varphi}, \boldsymbol{\chi}_{T+1}) = -I_h(\boldsymbol{\chi}_{proj}, \hat{\boldsymbol{\chi}}) \leq 0.$$

Теорема 3: согласованные прогнозы (Стенина, 2014)

При выполнении требований теоремы 1 использование в качестве вектора согласованных прогнозов $\hat{\varphi}$ вектора

$$\hat{\varphi} = \boldsymbol{\chi}_{proj} = \arg\min_{\boldsymbol{\chi} \in \mathcal{A} \cap \mathcal{B}} I_h(\boldsymbol{\chi}, \hat{\boldsymbol{\chi}})$$

гарантирует, что вектор согласованных прогнозов будет удовлетворять требованиям согласованности и качества и физическим ограничениям.

Задача согласования прогнозов сводится к решению оптимизационной задачи.

Примеры функций потерь — дивергенции Брегмана

Удовлетворяют условию теоремы 1

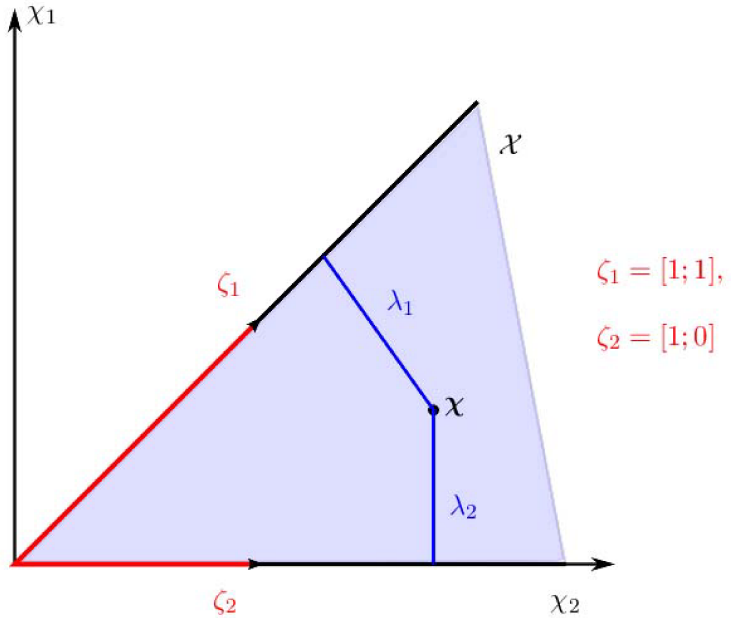
- $l_h(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|^2$ — квадрат евклидового расстояния,
- $l_h(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathbf{u} - \mathbf{v})^T Q(\mathbf{u} - \mathbf{v})$ — квадрат расстояния Махаланобиуса,
- $l_h(\mathbf{u}, \mathbf{v}) = \sum\limits_{i=1}^{d} u_i \log \frac{u_i}{v_i} - \sum\limits_{i=1}^{d} u_i + \sum\limits_{i=1}^{d} v_i$ — обобщенная дивергенция Кульбака-Лейблера,
- $l_h(\mathbf{u}, \mathbf{v}) = \sum\limits_{i=1}^{d} \left(\frac{u_i}{v_i} - \log \frac{u_i}{v_i} - 1 \right)$ — расстояние Itakura-Saito,
- $l_h(\mathbf{u}, \mathbf{v}) = \sum\limits_{i=1}^{d} |u_i|^a - \sum\limits_{i=1}^{d} |v_i|^a - a \sum\limits_{i=1}^{d} sign(v_i)|v_i|^{a-1}(u_i - v_i)$, $a > 1$,
- $l_h(\mathbf{u}, \mathbf{v}) = \frac{2}{a^2} \sum\limits_{i=1}^{d} (e^{au_i} - e^{av_i}) - \frac{2}{a} \sum\limits_{i=1}^{d} e^{av_i}(u_i - v_i)$, $a \neq 0$.

Порождающее представление конуса

Порождающее представление конуса

Полиэдральный конус $\mathcal{X}$ допускает представление через конечный набор порождающих элементов $\zeta_1, ..., \zeta_k$:

$$\mathcal{X} = \left\{ \sum_{k=1}^{r} \lambda_k \zeta_k \mid \lambda_k \geq 0 \right\}.$$



Теорема (о порождающем представлении конуса), [Кузнецов: 2013]

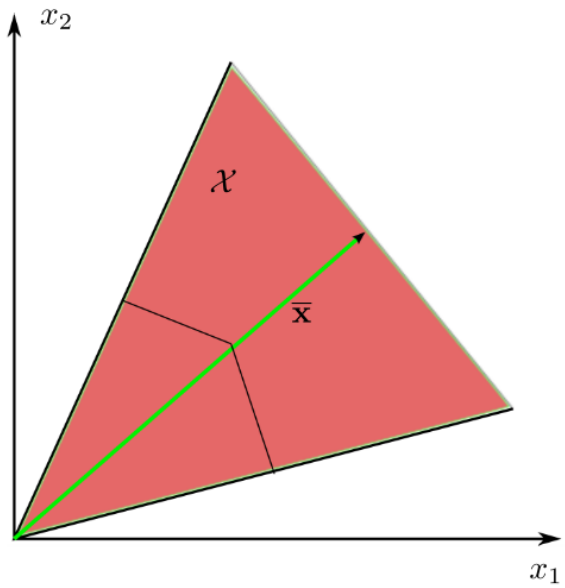
Столбцы матрицы предпочтений $\mathbf{Z}(i, k) = \mathbb{I}[x_i \succeq x_k]$ являются порождающими элементами конуса предпочтений,

$$\mathcal{X} \supset \{\mathbf{Z}\boldsymbol{\lambda} \mid \boldsymbol{\lambda} \in \mathbb{R}^m_+\}.$$

Регуляризация конусной модели

Линейная конусная модель:

$$\mathbf{f}(X) = \sum_{j=1}^{n} \mathbf{Z}_j \boldsymbol{\lambda}_j, \quad \boldsymbol{\lambda}_j \geq \mathbf{0}.$$



Рассмотрим в конусе $\mathcal{X}$ центральную точку $\overline{\mathbf{x}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{z}_j$.

Теорема (о регуляризации конусной модели), [Кузнецов: 2014]

В случае замены каждого конуса $\mathcal{X}_k = \{\sum \lambda_{jk} \mathbf{z}_{jk} \mid \boldsymbol{\lambda}_k \geq \mathbf{0}\}$ его центральной точкой конусная модель представима в виде

$$\mathbf{f}(\mathbf{x}_1, ..., \mathbf{x}_m) = \hat{\mathbf{Z}}\boldsymbol{\lambda}, \quad \hat{\mathbf{Z}} = \sum_{j=1}^{n} w_j \mathbf{Z}_j,$$

при ограничениях $w_j \geq 0$, $\sum^{m} \lambda_k = 1$, $\boldsymbol{\lambda} \geq \mathbf{0}$.

Приближенная оптимальность релаксированной низкоранговой задачи

Теорема (Мотренко)

Пусть $\underline{\mathbf{A}}^R = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \mathbf{a}_2^{(r)} \circ \mathbf{a}_3^{(r)}$ *— построенное низкоранговое приближение матрицы* $\underline{\mathbf{A}}$*. Отклонение функции* $F(\underline{\mathbf{A}})$ *от* $F(\underline{\mathbf{A}}^R)$ *линейно зависит от невязки* $||\underline{\mathbf{A}} - \underline{\mathbf{A}}^R||$*:*

$$\left| F(\underline{\mathbf{A}}) - F\left(\underline{\mathbf{A}}^R\right) \right| = \mathcal{O}\left(||\underline{\mathbf{A}} - \underline{\mathbf{A}}^R|| \right).$$

Теорема (Мотренко)

Пусть $\hat{\underline{\mathbf{A}}} \in [0,1]^{n_1 \times n_2 \times n_3}$ *— некоторая трехиндексная матрица,* $\underline{\mathbf{A}}^{\varepsilon} \in 0,1^{n_1 \times n_2 \times n_3}$ *получена бинаризацией* $\hat{\underline{\mathbf{A}}}$*:*

$$a_{ijk}^{\varepsilon} = [a_{ijk} > \varepsilon].$$

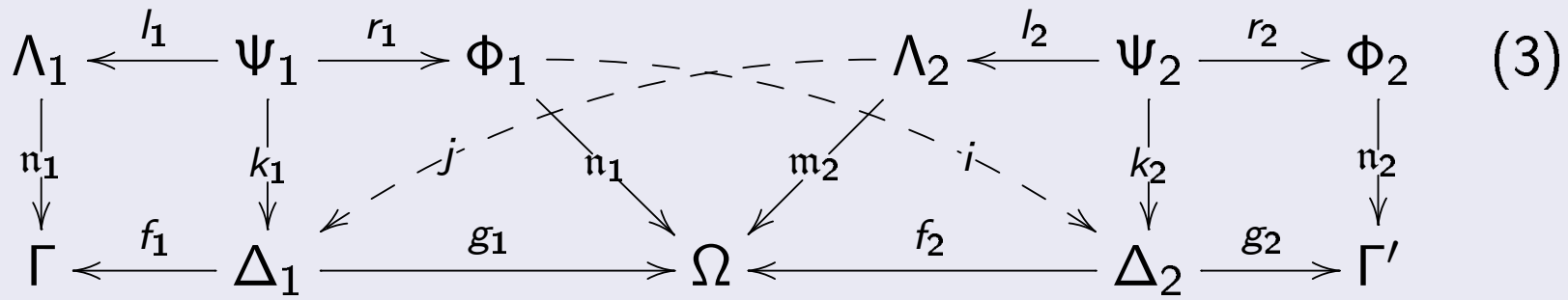
Отклонение функционала $F(\hat{\underline{\mathbf{A}}})$ *от* $F(\underline{\mathbf{A}}^{\varepsilon})$ *линейно зависит от* ε*:*

$$\left| F(\hat{\underline{\mathbf{A}}}) - F\left(\underline{\mathbf{A}}^{\varepsilon}\right) \right| = \mathcal{O}\left(n_1 n_2 n_3 \varepsilon\right).$$

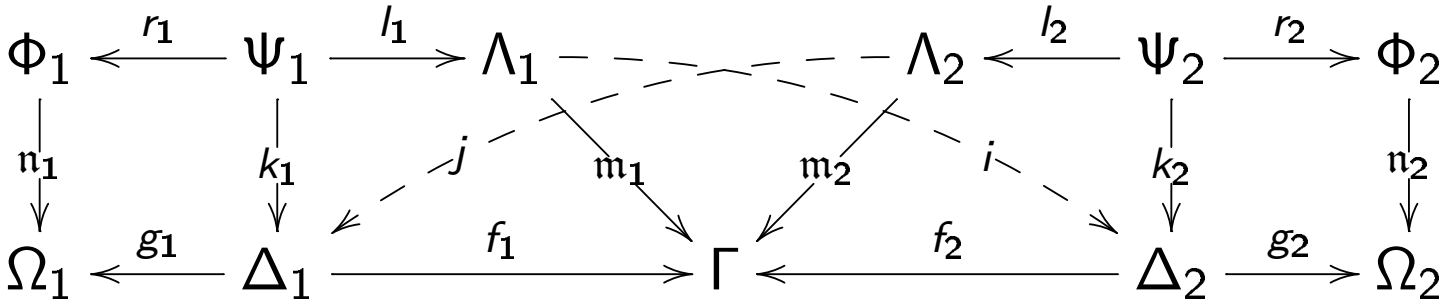
Критерий последовательной независимости

Theorem

Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega$ *и* $\Omega \overset{p_2,m_2}{\to} \Gamma'$ *являются последовательно независимыми, если существуют морфизмы* $i : \Phi_1 \to \Delta_2$ *и* $j : \Lambda_2 \to \Delta_1$, *такие что* $f_2 \circ i = \mathfrak{n}_1$ *и* $g_1 \circ j = \mathfrak{m}_2$*:*



Доказательство критерия параллельной независимости



Рассмотрим варианты принадлежности $\mathfrak{m}_1(v)$.

1. Множество $\mathfrak{m}_1(v) \notin \mathfrak{m}_2(\Lambda_2)$. Все вершины графа Γ являются образами при применении отображений $\mathfrak{m}_2$ или f_2. Отсюда $\mathfrak{m}_1(v) \in f_2(\Delta_2)$.
2. Множество $\mathfrak{m}_1(v) \in \mathfrak{m}_2(\Lambda_2)$. Тогда $\mathfrak{m}_1(v) \in \mathfrak{m}_1(\Lambda_1) \cap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{m}_1(l_1(\Psi_1)) \cap \mathfrak{m}_2(l_2(\Psi_2))$. При этом из коммутативной диаграммы следует, что $\mathfrak{m}_2(l_2(\Psi_2)) = f_2(k_2(\Psi_2))$. Отсюда $\mathfrak{m}_1(v) \in f_2(\Delta_2)$.

В обоих случаях $\mathfrak{m}_1(x) \in f_2(\Delta_2)$, из инъективности f_2: $i(x) = f_2^{-1} \circ \mathfrak{m}_1(x)$. Аналогично, j из условия $f_1 \circ j = \mathfrak{m}_2$.

Оценка апостериорного распределения параметров $\boldsymbol{\theta}$ и $\boldsymbol{\alpha}$

Теорема [Кузьмин, 2016]

Апостериорное распределение параметров q из класса $q(\boldsymbol{\theta})q(\mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})$, заданное оптимальными оценками факторов $q^\star(\boldsymbol{\theta})$ и $q^\star(\mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})$, в которых используется верхняя оценка правдоподобия имеет вид

$$q = \mathcal{N}(\boldsymbol{\alpha}_0, a^{-1}\mathbf{I}) \prod_{k=1}^{k_h} \mathcal{N}(\mathbf{m}'_{0k}, (\nu'\mathbf{V}_k)^{-1}) \mathcal{N}(\mathbf{m}_{0k}, (b'\mathbf{V}_k)^{-1}) \mathcal{W}(\mathbf{W}_k, \nu'),$$

$$\mathbf{m}'_{0k} = \mathbf{m}_{0k} + \frac{1}{\nu'}(\mathbf{W}_k^{-1})^{\mathsf{T}} \mathbf{M}_k^{\mathsf{T}} \mathsf{E}_{\boldsymbol{\alpha}}[\boldsymbol{\Lambda}] \sum_{n=1}^{N} \mathbf{x}_n \left(z_{nk} - \frac{\exp(\xi_{nk})}{g(\boldsymbol{\xi}_n)} \right),$$

$$\boldsymbol{\alpha}_0 = \frac{1}{a} \sum_{n=1}^{N} \sum_{m=1}^{|W|} x_{nm} \boldsymbol{\iota}_m \sum_{k=1}^{k_h} (\mathbf{M}_k \mathsf{E}[\boldsymbol{\theta}_k])_m \left(z_{nk} - \frac{\exp(\xi_{nk})}{g(\boldsymbol{\xi}_n)} \right),$$

$$\mathbf{W}_k^{-1} = (b+1)\mathbf{m}_{0k}\mathbf{m}_{0k}^{\mathsf{T}} + b\mathbf{m}_0\mathbf{m}_0^{\mathsf{T}} + \mathsf{E}\left[\boldsymbol{\theta}_k\boldsymbol{\theta}_k^{\mathsf{T}}\right] + \mathbf{W}^{-1},$$

$$\mathbf{m}_{0k} = \frac{\mathsf{E}[\boldsymbol{\theta}_k] + b\mathbf{m}_0}{1+b}, \quad b' = b+1, \quad \nu' = \nu+1.$$

Сравнение точечной оценки вероятности с оценкой обоснованности

Операторы релевантности:

R_{MAP} ранжирует кластеры по точечной оценке $p(\tilde{z}_{tk}|\boldsymbol{\theta}_k^{\text{MAP}}, \boldsymbol{\alpha}^{\text{MAP}})$,

$$\boldsymbol{\theta}_k^{\text{MAP}}, \boldsymbol{\alpha}^{\text{MAP}} = \arg\max_{\boldsymbol{\theta},\boldsymbol{\alpha}} q(\boldsymbol{\theta}, \boldsymbol{\alpha}),$$

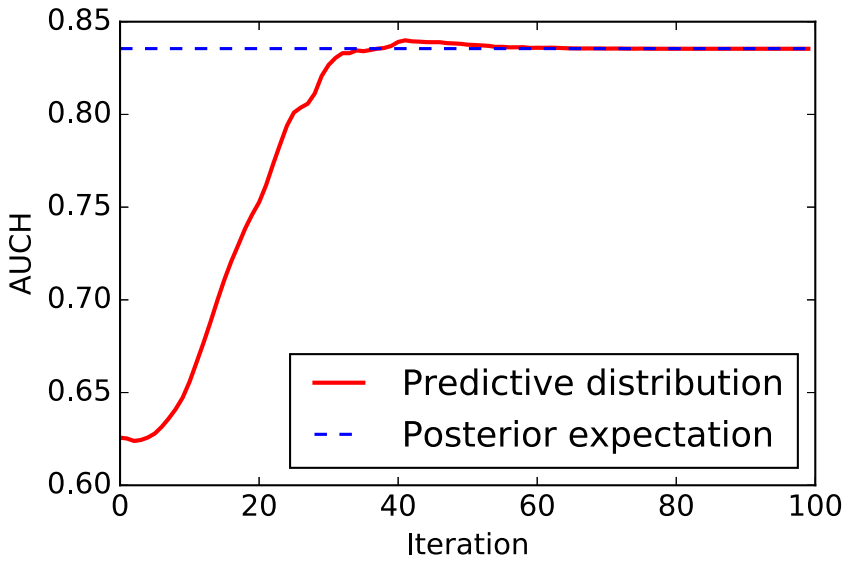
$\hat{R}_{evidence}$ ранжирует кластеры по оценке обоснованности

$$p(\tilde{z}_{kt}|\tilde{\mathbf{x}}_t) = \int p(\tilde{z}_{kt}|\boldsymbol{\theta}, \boldsymbol{\alpha})_k q(\boldsymbol{\theta}, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\alpha} \approx \hat{p}(\tilde{\mathbf{x}}_t, \tilde{\boldsymbol{\xi}}_t, \psi_{tk}) \to \max_{\psi_{tk}} \min_{\tilde{\boldsymbol{\xi}}_t},$$

$\tilde{\boldsymbol{\xi}}_t, \psi_{tk}$ – вариационные параметры оценки.

Теорема [Кузьмин, 2016]

При оптимальных значениях вариационных параметров $\tilde{\boldsymbol{\xi}}_t$ и $\boldsymbol{\psi}_t$, значение критерия качества AUCH(R) для операторов R_{MAP} и $\hat{R}_{evidence}$ совпадают.



Требования к функции сходства s

Корректная функция сходства s должна быть

1. определена в случае несовпадения носителей,
2. $s(g_1,\ g_2) \leq s(g_1,\ g_1)$,
3. $s \in [0,\ 1]$,
4. $s(g_1,\ g_1) = 1$,
5. близка к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение,
6. симметрична, $s(g_1,\ g_2) = s(g_2,\ g_1)$.

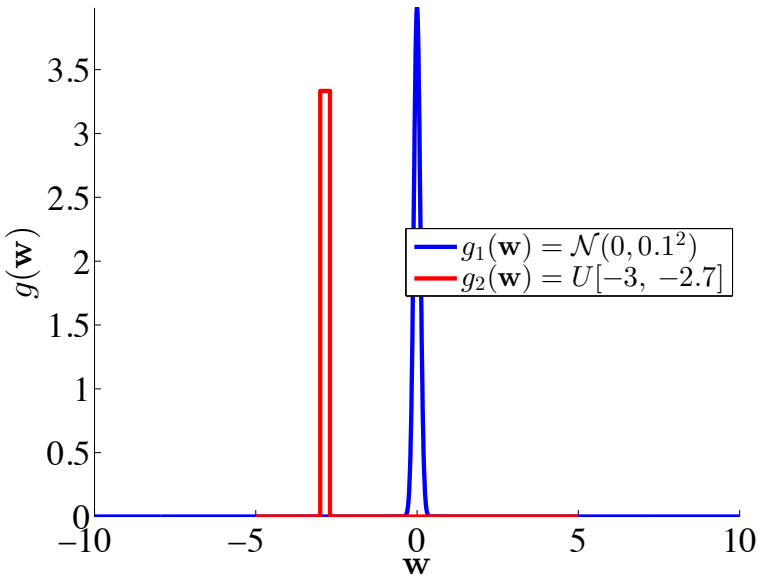
Теорема 3 (Адуенко, 2014)

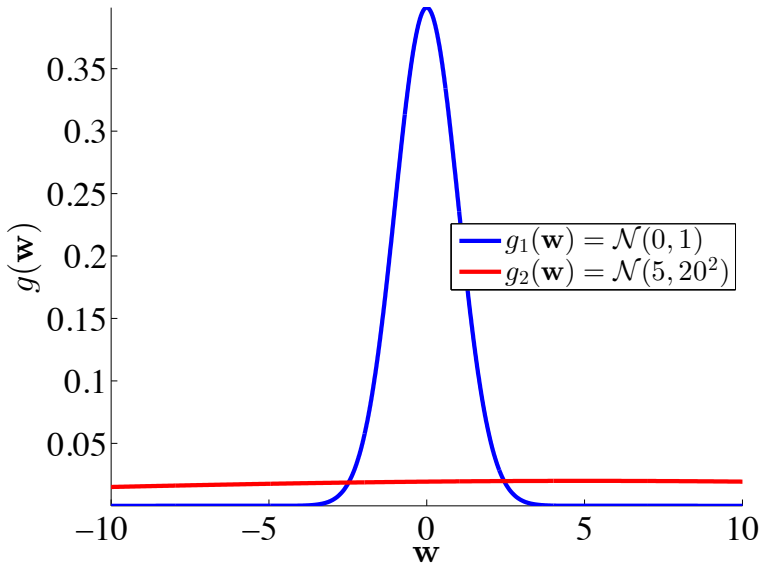
Функции сходства, порожденные расстояниями Кульбака-Лейблера, Дженсона-Шеннона, Хеллингера, Бхаттачарайа, не являются корректными.

Иллюстрация требований к функции сходства

Важно, чтобы значение функции s

было близко к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение.





Теорема 4 (Адуенко, 2014)

Функции сходства, порожденные дивергенциями Брегмана, симметризованными дивергенциями Брегмана и f-дивергенциями, не являются корректными.

Предлагаемая функция сходства

В качестве меры сходства распределения предлагается мера сходства s-score:

$$s(g_1,\ g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w}-\mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}.$$

Теорема 5 (Адуенко, 2014). Предлагаемая функция сходства является корректной.

Примеры:

$g_1(\mathbf{w})$	$g_2(\mathbf{w})$	$s(g_1,\ g_2)$
$U[0,\ 1]$	$U[0.5,\ 1.5]$	0.5
$U[0,\ 1]$	$U[0,\ 1]$	1
$\mathcal{N}(0,\ 1)$	$\mathcal{N}(10, 10^{10})$	1

Выражение для $s(g_1,\ g_2)$ для пары нормальных распределений

Определение 6. *Обобщенно-линейной моделью с натуральной функцией связи* и априорным распределением на вектор параметров $p(\mathbf{w}|\mathbf{A})$ называется вероятностная модель с совместным правдоподобием

$$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A}),\ \text{где}\ p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i,\ \mathbf{w}),$$

$$p(y_i|\mathbf{x}_i,\ \mathbf{w}) = c(y_i)\exp\big(\theta_i y_i - b(\theta_i)\big),\ \text{где}\ \theta_i = \mathbf{w}^{\mathsf{T}}\mathbf{x}_i.$$

Теорема 6 (Адуенко, 2014).
Пусть $g_1 = \mathcal{N}(\mathbf{v}_1,\ \mathbf{\Sigma}_1),\ g_2 = \mathcal{N}(\mathbf{v}_2,\ \mathbf{\Sigma}_2)$. Тогда выражение для $s(g_1,\ g_2)$ имеет вид

$$s(g_1,\ g_2) = \exp\left(-\tfrac{1}{2}(\mathbf{v}_1 - \mathbf{v}_2)^{\mathsf{T}}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{v}_1 - \mathbf{v}_2)\right).$$

Следствие 1. В случае $\mathbf{\Sigma}_2 = \mathbf{0}$ выражение для s-score

$$s(g_1,\ g_2) = \exp\left(-\tfrac{1}{2}(\mathbf{v}_2 - \mathbf{v}_1)^{\mathsf{T}}\mathbf{\Sigma}_1^{-1}(\mathbf{v}_2 - \mathbf{v}_1)\right).$$

Распределение s-score в условии истинности гипотезы о совпадении моделей

Рассматриваем далее пару обобщенно-линейных моделей с натуральной функцией связи. Обозначим $\mathbf{H}_m(\mathbf{w}) = \mathbf{H}_m$ и введем $O_m^{\delta}(\mathbf{w}) = \{\mathbf{v} : \|\mathbf{H}_m^{T/2}(\mathbf{v} - \mathbf{w})\| \leq \delta\}$.

Теорема 7 (Адуенко, 2016). Пусть

- Модели f_1 и f_2 совпадают: $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$, $\mathbf{w}_2$ известно;
- Априорное распределение: $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0},\ \mathbf{A}_{m^k}^{-1}),\ k = 1,\ 2$;
- $\exists\ c_0^k < \infty :\ \ \|\mathbf{A}_{m^k}^k\| < c_0^k\ \ \forall\ m^k,\ \ k = 1,\ 2$;
- $\sum_{i=1}^{m^1} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}$ имеет полный ранг для $m^1 \geq m_0$;
- $\lambda_{\min}(\mathbf{H}_{m^1}) \to \infty$ при $m^1 \to \infty$;
- $\forall\, \delta > 0 \ \max\limits_{\mathbf{v} \in O_{m^1}^{\delta}(\mathbf{w})} \left\|\mathbf{H}_{m^1}^{-\frac{1}{2}} \mathbf{H}_{m^1}(\mathbf{v}) \mathbf{H}_{m^1}^{-\frac{\mathsf{T}}{2}} - \mathbf{I}\right\| \to 0$ при $m^1 \to \infty$.

Тогда $-2 \log \text{s-score} = (\hat{\mathbf{w}} - \mathbf{w})^{\mathsf{T}} \tilde{\mathbf{H}}_{m^1}(\hat{\mathbf{w}})(\hat{\mathbf{w}} - \mathbf{w}) \xrightarrow{d} \chi^2(n)$.

Следствие 1. Для случая $n = 2$ s-score имеет асимптотически равномерное распределение на отрезке $[0,\ 1]$.

Обобщающая задача

Задачу выбора модели $\mathbf{h}^*, \boldsymbol{\theta}^*$ назовем обобщающей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, если выполнены условия:

1. Для каждого $\mathbf{h} \in U_h$ и каждого $\boldsymbol{\lambda} \in U_\lambda$ решение $\boldsymbol{\theta}^*$ определено однозначно.
2. **Условие максимизации правдоподобия выборки:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$: матожидания правдоподбия выборок: $\mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) > \log \mathrm{E}_q\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$.
3. **Условие минимизации сложности модели:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, $\mathrm{E}_q \log\, p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}, \mathbf{f}) = \log \mathrm{E}_q\, p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}}, \mathbf{f})$, количество ненулевых параметров у первой модели меньше, чем у второй.
4. **Условие максимизации обоснованности модели:** существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:
$$\mathbf{h}^* = \arg\max p(\mathbf{y}|\mathbf{X}, \mathbf{h}', \lambda_{\text{temp}}, \mathbf{f}), \quad \boldsymbol{\theta}^* = \arg\min D_{\mathrm{KL}}(q|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}}, \mathbf{f})).$$
5. **Условие перехода между структурами:** Существует константа K_3, такая что для любых двух векторов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих векторов $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*$: $D_{\mathrm{KL}}(q_{\boldsymbol{\Gamma}_2}, q_{\boldsymbol{\Gamma}_1}) > K_3, D_{\mathrm{KL}}(q_{\boldsymbol{\Gamma}_1}, q_{\boldsymbol{\Gamma}_2}) > K_3$: 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. **Условие непрерывности:** $\mathbf{h}^*, \boldsymbol{\theta}^*$ непрерывны по метапараметрам.

Анализ задач выбора моделей

Теорема [Бахтеев, 2019]

Следующие задачи выбора модели не являются обобщающими:

1. метод максимума правдоподобия: $\max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f})$;
2. метод максимума апостериорной вероятности $\max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h}, \lambda_{\text{temp}}) p(\mathbf{h}|\mathbf{f})$;
3. метод максимума вариационной оценки обоснованности модели $\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} \mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{f}) - \mathrm{D}_{KL}\big(q(\mathbf{w}, \boldsymbol{\Gamma})||p(\mathbf{w}, \boldsymbol{\Gamma}, \lambda_{\text{temp}})\big) p(\mathbf{h}|\mathbf{f})$;
4. кросс-валидация $\max_{\mathbf{h}} \mathrm{E}_q \log p(\mathbf{y}_{\text{valid}}|\mathbf{X}_{\text{valid}}, \boldsymbol{\theta}^*, \lambda_{\text{temp}}, \mathbf{f}) p(\mathbf{h}|\mathbf{f})$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h})$.
5. AIC: $\max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - |\theta_i : \theta_i \neq 0|$;
6. BIC: $\max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - \frac{1}{2}\log(m)|\theta_i : \theta_i \neq 0|$;
7. перебор структуры модели: $\max \boldsymbol{\Gamma}' \max_{\boldsymbol{\theta}} \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) \mathbb{I}(\boldsymbol{\Gamma} = \boldsymbol{\Gamma}')$.

Предлагаемая задача оптимизации

Теорема [Бахтеев, 2018]

Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на некоторой области U. Тогда следующая задача является обобщающей на U.

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q = \qquad (Q^*)$$

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

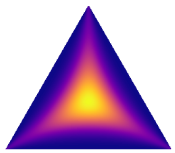
$$- \lambda^{\mathrm{prior}}_{\mathrm{Q}} \mathrm{D}_{KL}\left(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f})\right) -$$

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

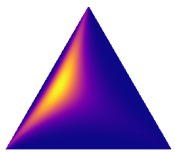
где

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

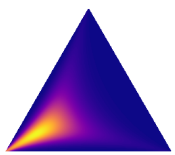
$$- \lambda^{\mathrm{prior}}_{\mathrm{L}} \mathrm{D}_{KL}\left(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f})\right).$$



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [0; 0; 0].$



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [1; 0; 0].$



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [1; 1; 0].$

Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.

Адекватность задачи оптимизации

Теорема, [Бахтеев, 2018]

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

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

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:

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

Анализ обобщающей задачи оптимизации

Теорема, [Бахтеев, 2018]

Пусть $\lambda^L_{\text{prior}} > 0, m \gg 0, \frac{m}{\lambda^L_{\text{prior}}} \in \mathbb{N}$. Тогда оптимизация функции

$$L = \mathrm{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - \lambda^L_{\text{prior}} \mathrm{D}_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp,f}})))$$

эквивалентна минимизации $\mathrm{E}_{\hat{\mathbf{X}},\hat{\mathbf{y}}\sim p(\mathbf{X},\mathbf{y})} \mathrm{D}_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\hat{\mathbf{X}}, \hat{\mathbf{y}}, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}))$, где $\hat{\mathbf{X}}, \hat{\mathbf{y}}$ — случайные подвыборки мощностью $\frac{m}{\lambda^L_{\text{prior}}}$ из генеральной совопкупности.

Определение

Параметрической сложностью модели назовем минимальную дивергенцию между априорным и вариационным распределением:

$$C_p = \min_{\mathbf{h}} D_{\mathrm{KL}}(q||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})).$$

Теорема, [Бахтеев, 2018]

При устремлении параметрической сложности модели к нулю относительная плотность параметров модели стремится к единице:

$$C_p \to 0 \implies \boldsymbol{\rho}(\mathbf{w}) \to 1, \quad \rho(w) = \frac{q(0)}{q(w)} = \exp\left(-\frac{\mu^2}{\sigma^2}\right).$$

Оптимизация параметрической сложности

Теорема, [Бахтеев, 2018]

Пусть $\lambda^Q_{\text{likelihood}} = \lambda^L_{\text{prior}} > 0, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Тогда предел оптимизации

$$\lim_{\lambda^{\mathbf{Q}}_{\mathbf{prior}} \to \infty} \lim_{\eta \to \infty} T^\eta\big(Q, \mathbf{h}, T^\eta(L, \theta_0, \mathbf{h})\big)$$

доставляет минимум параметрической сложности. Существует компактная область U, такая что для любой точки $\theta_0 \in U$ предел данной оптимизации доставляет нулевую параметрическую сложность: $C_p = 0$.

Теорема, [Бахтеев, 2018]

Пусть $\lambda^L_{\text{likelihood}} = 1, \boldsymbol{\lambda}^Q_{\text{struct}} = \mathbf{0}$. Пусть $\mathbf{f}_1, \mathbf{f}_2$ — результаты градиентной оптимизации при разных значениях гиперпараметров $\lambda^{Q,1}_{\text{prior}}, \lambda^{Q,2}_{\text{prior}}, \lambda^{Q,1}_{\text{prior}} < \lambda^{Q,2}_{\text{prior}}$, полученных при начальном значении вариационных параметров θ_0 и гиперпараметров $\mathbf{h}_0$. Пусть $\theta_0, \mathbf{h}_0$ принадлежат области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда:

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