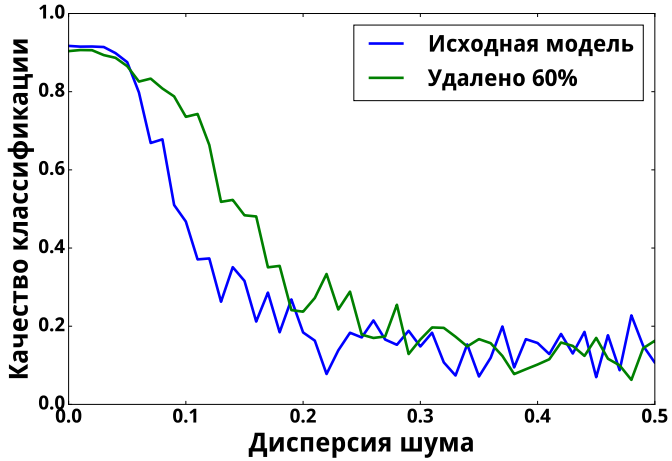
Выбор структуры модели глубокого обучения

Бахтеев Олег

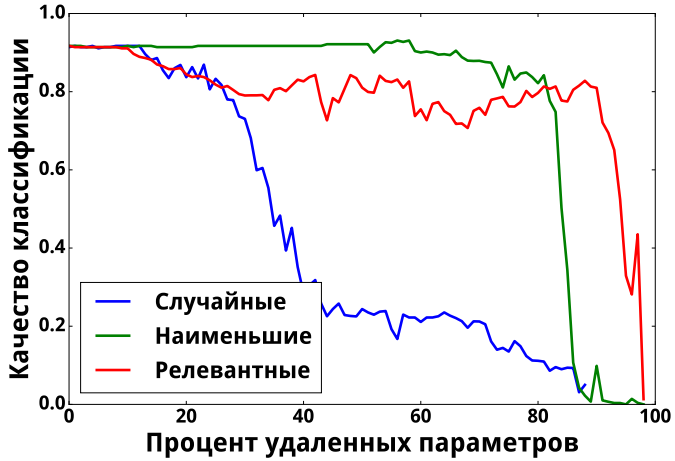
МФТИ

06.02.2019

Сложность модели: зачем?



Устойчивость моделей при возмущении выборки



Качество классификации при удалении параметров

Сложность модели: зачем?

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [59]	299×299	23.8 M	5.72 B	78.0	93.9
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [57]	299×299	55.8 M	13.2 B	80.4	95.3
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [67]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [68]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

Zoph et. al, 2017. Сложность моделей отличается почти в два раза при одинаковом качестве.

Глубокого обучение

Определение

Моделью $\mathbf{f}(\mathbf{w}, \mathbf{x})$ назовем дифференцируемую по параметрам $\mathbf{w}$ функцию из множества признаковых описаний объекта во множество меток:

$$\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{Y},$$

где $\mathbb{W}$ — пространство параметров функции $\mathbf{f}$.

Особенность задачи выбора модели *глубокого обучения* — значительное число параметро в моделях приводит к неприменимости классических методов оптимизации и выбора модели.

Сложность модели:

1. количество параметров;
2. количество суперпозиций внутри модели.

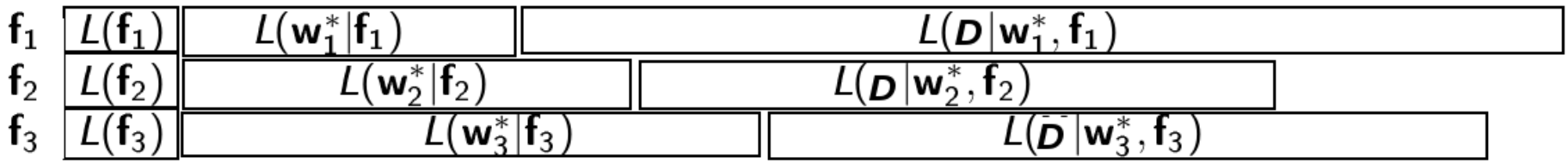
Принцип минимальной длины описания

$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) = L(\mathbf{f}) + L(\mathfrak{D}|\mathbf{f}),$$

где $\mathbf{f}$ — модель, $\mathfrak{D}$ — выборка, L — длина описания в битах.

$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathfrak{D}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$ — оптимальные параметры модели.



MDL и Колмогоровская сложность

Колмогоровская сложность — длина минимального кода для выборки на предварительно заданном языке.

Теорема инвариантности

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависящую от мощности выборки.

Отличия от MDL:

- Колмогоровская сложность невычислима.
- Длина кода может зависеть от выбранного языка. Для небольших выборок теорема инвариантности не дает адекватных результатов.

Байесовый подход к сложности

Правдоподобие модели (“Evidence”):

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

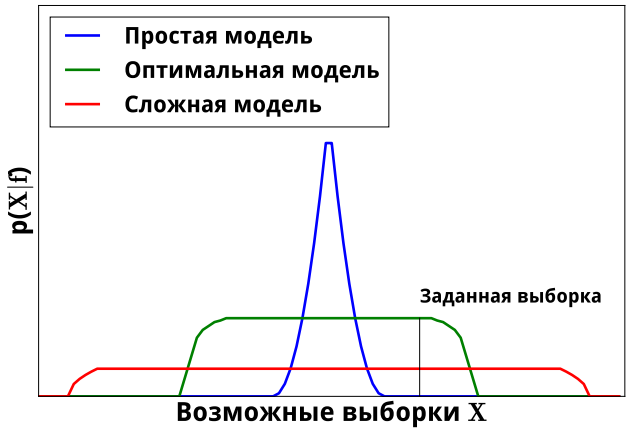
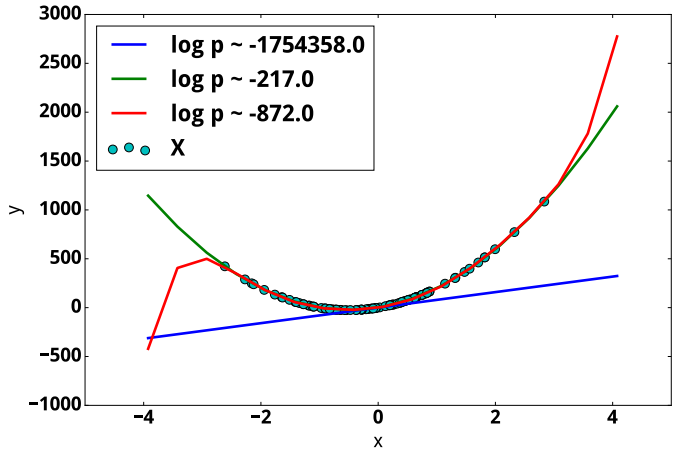


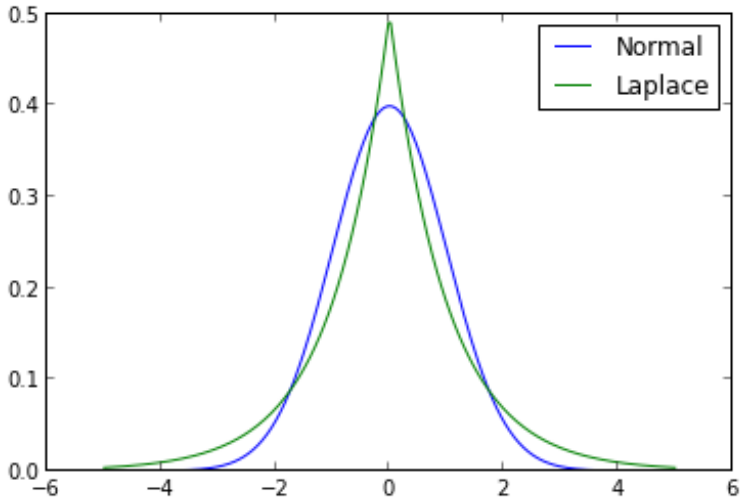
Схема выбора модели по правдоподобию



Пример: полиномы

Evidence vs MDL

Evidence	MDL
Использует априорные знания	Независима от априорных знаний
Основывается на гипотезе о порождении выборки вне зависимости от их природы	Минимизирует длину описания выборки



Оптимальность модели

Определение

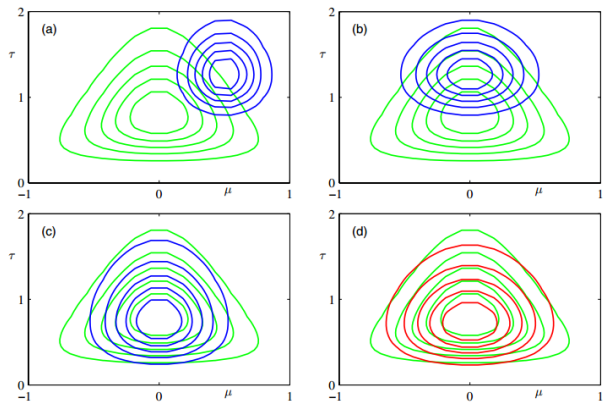
Пусть задано множество моделей M.
Пусть для каждой модели $\mathbf{f}$ задано априорное распределение параметров: $p(\mathbf{w}|\mathbf{h})$, где $\mathbf{h}$ — параметры априорного распределения.
Модель $\mathbf{f}$ назовем оптимальной среди моделей M, если достигается максимум интеграла:

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

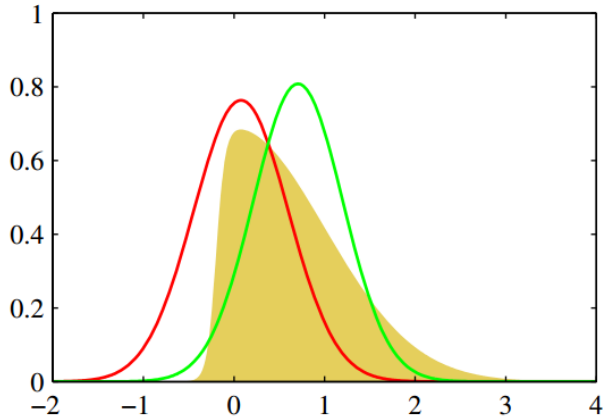
Вариационная оценка, ELBO

Вариационная оценка Evidence, Evidence lower bound — метод нахождения приближенного значения аналитически невычислимого распределения $p(\mathbf{w}|\mathfrak{D}, \mathbf{h})$ распределением $q(\mathbf{w}) \in \mathbf{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D})}{q(\mathbf{w})} d\mathbf{w}.$$



Аппроксимация неизвестного распределения нормальным



Аппроксимация Лапласа (красная линия) и вариационная оценка (зеленая линия)

Получение вариационной нижней оценки

$$\log p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D}, \mathbf{h})) \geq$$

$$\geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= -\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})) + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w}, \mathbf{h}) d\mathbf{w},$$

где

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w}.$$

Определение

Модель $\mathbf{f}$ назовем субоптимальной на множестве моделей M, если модель доставляет максимум нижней вариационной оценке:

$$\mathbf{f} = \arg\max_{\hat{\mathbf{f}} \in M} \max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathfrak{D}, \hat{\mathbf{f}})}{q(\mathbf{w})} d\mathbf{w}.$$

$D_{\mathbf{KL}}$

Максимизация вариационной нижней оценки

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w}$$

эквивалентна минимизации дивергенции между распределением распределением $q(\mathbf{w}) \in Q$ и апостериорным распределением параметров $p(\mathbf{w}|\mathfrak{D}, \mathbf{h})$:

$$q = \operatorname{argmax}_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow q = \operatorname{argmin}_{q \in Q} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D}, \mathbf{h})),$$

т.к.

$$\log p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D}, \mathbf{h})) = \text{const}.$$

Использование вариационной нижней оценки

Для чего используют вариационный вывод?

- получение оценок Evidence;
- получение оценок распределений моделей со скрытыми переменными (тематическое моделирование, снижение размерности).

Зачем используют вариационный вывод?

- сводит задачу нахождения апостериорной вероятности к методам оптимизации;
- проще масштабируется, чем аппроксимация Лапласа;
- проще в использовании, чем сэмплирующие методы.

Вариационный вывод может давать сильно заниженную оценку.

ELBO: нормальное распределение

Пусть $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q)$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log\, p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{h}) d\mathbf{w} - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) \simeq$$

$$\sum_{i=1}^{m} \log\, p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathbf{w}}) - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) \to \max_{\mathbf{A_q}, \boldsymbol{\mu_q}}, \quad \hat{\mathbf{w}} \sim q.$$

В случае, если априорное распределение параметров $p(\mathbf{w}|\mathbf{h})$ является нормальным:

$$p(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}),$$

дивергенция $D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})$ вычисляется аналитически:

$$D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) = \frac{1}{2}\big(\mathrm{tr}(\mathbf{A}^{-1}\mathbf{A}_q) + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - n + \ln|\mathbf{A}| - \ln|\mathbf{A}_q|\big).$$

ELBO: нормальное распределение

“Обычная” функция потерь:

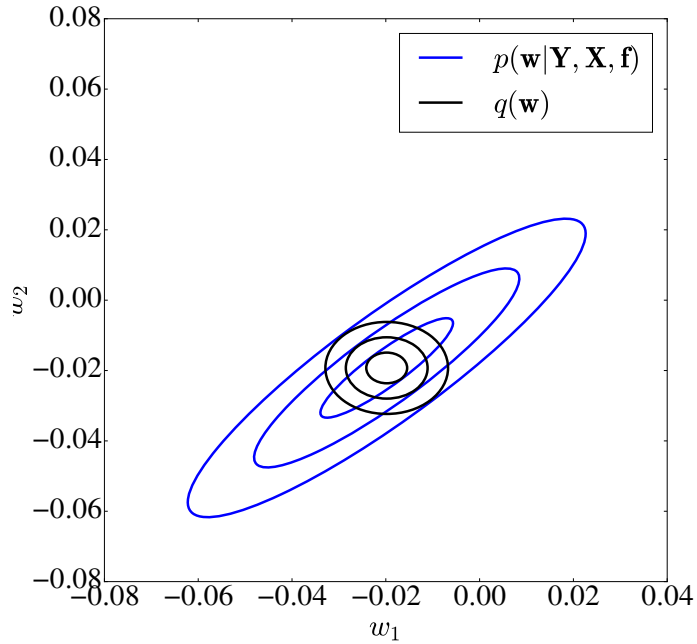
$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} -\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \lambda||\mathbf{w}||_2^2.$$

Пример грубой аппроксимации нормальным диагональным распределением q

Вариационный вывод при $(p(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\mathbf{0},\mathbf{1}))$**:**

$$L = \sum_{\mathbf{x},\mathbf{y}} \log\, p(\mathbf{y}|\mathbf{x},\hat{\mathbf{w}}) +$$

$$+\frac{1}{2}\left(\mathrm{tr}(\mathbf{A}_q) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\mu}_q - \ln\,|\mathbf{A}_q|\right).$$



Оператор оптимизации, Maclaurin et. al, 2015

Определение

Назовем оператором оптимизации алгоритм T выбора вектора параметров $\mathbf{w}'$ по параметрам предыдущего шага $\mathbf{w}$:

$$\mathbf{w}' = T(\mathbf{w}).$$

Определение

Пусть L — дифференцируемая функция потерь.
Оператором градиентного спуска назовем следующий оператор:

$$T(\mathbf{w}) = \mathbf{w} - \beta \nabla L(\mathbf{w}, \mathbf{y}, \mathfrak{D}).$$

Градиентный спуск для оценки правдоподобия

Рассмотрим максимизацию совместного распределения параметров:

$$L = -\log\, p(\mathfrak{D}, \mathbf{w}|\mathbf{h}) = -\sum_{\mathfrak{D}\in\mathfrak{D}} \log\, p(\mathfrak{D}|\mathbf{w}, \mathbf{h}) p(\mathbf{w}|\mathbf{h})$$

Проведем оптимизацию нейросети из r различных начальных приближений $\mathbf{w}_1, \ldots, \mathbf{w}_r$ с использованием градиентного спуска:

$$\mathbf{w}' = T(\mathbf{w}).$$

Векторы параметров $\mathbf{w}_1, \ldots, \mathbf{w}_r$ соответствуют некоторому скрытому распределению $q(\mathbf{w})$.

Энтропия

Формулу вариационной оценки можно переписать с использованием энтропии:

$$\log p(\mathfrak{D}|\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} =$$

$$\mathsf{E}_{q(\mathbf{w})}[\log p(\mathfrak{D}, \mathbf{w}|\mathbf{h})] + \mathsf{S}(q(\mathbf{w})),$$

где $\mathsf{S}(q(\mathbf{w}))$ — энтропия:

$$\mathsf{S}(q(\mathbf{w})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w}.$$

Градиентный спуск для оценки правдоподобия

При достаточно малой длине шага оптимизации β разность энтропии на различных шагах оптимизации вычисляется как:

$$\mathsf{S}(q'(\mathbf{w})) - \mathsf{S}(q(\mathbf{w})) \simeq \frac{1}{r}\sum_{g=1}^{r}\big(-\beta\mathsf{Tr}[\mathbf{H}(\mathbf{w}'^{g})] - \beta^2\mathsf{Tr}[\mathbf{H}(\mathbf{w}'^{g})\mathbf{H}(\mathbf{w}'^{g})]\big).$$

Итоговая оценка на шаге оптимизации τ:

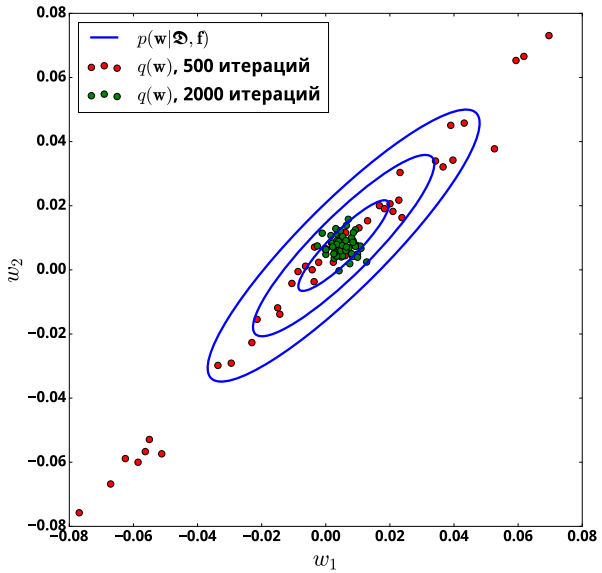
$$\log\hat{p}(\mathbf{Y}|\mathfrak{D},\mathbf{h}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathbf{w}_{\tau}^{g},\mathfrak{D},\mathbf{Y}) + \mathsf{S}(q^0(\mathbf{w}))+$$

$$+\frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\big(-\beta\mathsf{Tr}[\mathbf{H}(\mathbf{w}_b^g)] - \beta^2\mathsf{Tr}[\mathbf{H}(\mathbf{w}_b^g)\mathbf{H}(\mathbf{w}_b^g)]\big),$$

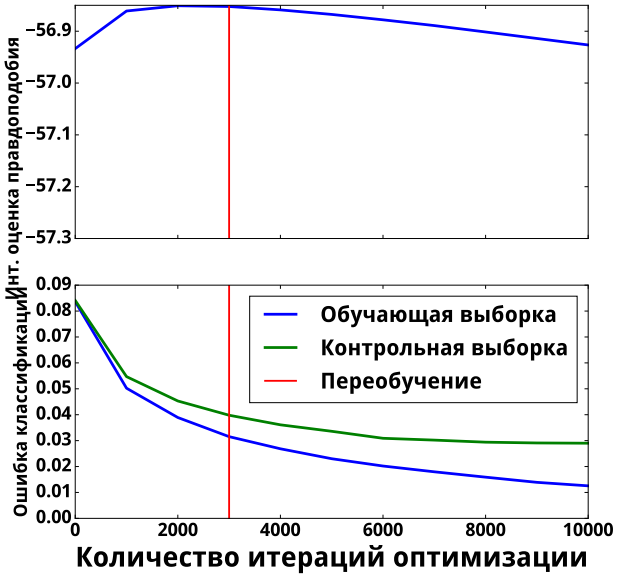
$\mathbf{w}_b^g$ — вектор параметров старта g на шаге b, $\mathsf{S}(q^0(\mathbf{w}))$ — начальная энтропия.

Переобучение, Maclaurin et. al, 2015

Градиентный спуск не минимизирует дивергенцию $\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D},\mathbf{h}))$. При приближении к моде распределения снижается оценка Evidence, что интерпретируется как переоубчение модели.



Схождение распределения к моде



Оценка начала переобучения

Задача оптимизации гиперпараметров

Задана дифференцируемая по параметрам модель, приближающая зависимую переменную y:

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) : \mathbb{W} \times \mathbb{X} \to \mathbb{Y}, \quad \mathbf{w} \in \mathbb{W}$$

Пусть $\boldsymbol{\theta} \in \mathbb{R}^u$ — вариационные параметры распределения.
$L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y})$ — дифференцируемая функция потерь по которой производится оптимизация функции f.
$Q(\mathbf{h}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$ — дифференцируемая функция определяющая итоговое качество модели f.
Требуется найти параметры $\boldsymbol{\theta}^*$ и гиперпараметры $\mathbf{h}$ модели, доставляющие минимум следующему функционалу:

$$\mathbf{h}^* = \arg\min_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{y}),$$

$$\boldsymbol{\theta}^*(\mathbf{h}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^u} L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y}),$$

Байесовский вывод

Первый уровень:

$$\theta^* = \arg\max\bigl(-L\bigr) = p(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathbf{y}|\mathbf{X},\mathbf{h})}.$$

Второй уровень:

$$p(\mathbf{h}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{h})p(\mathbf{h}).$$

Полагая распределение параметров $p(\mathbf{h})$ равномерным на некоторой большой окрестности, получим задачу оптимизации гиперпараметров:

$$p(\mathbf{y}|\mathbf{X},\mathbf{h}) = \int_{\mathbf{w}\in\mathbb{R}^{\boldsymbol{u}}} p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\mathbf{h}) = -Q \to \max_{[\alpha_{\mathbf{1}},\ldots,\alpha_{\boldsymbol{n}}]\in\mathbb{R}^{\boldsymbol{n}}}. \qquad (1)$$

Другие примеры L, Q

- Кросс-валидация (L — ошибка на обучении, Q — на контроле);
- вариационная оценка ($L = Q = \text{ELBO}$).

Формальная постановка задачи: градиентная оптимизация

Определение

Пусть задан оператор T, проводящий η шагов оптимизации по функции L:

$$\boldsymbol{\theta}^* = T \circ T \circ \cdots \circ T(\boldsymbol{\theta}^0, \mathbf{h}) = T^{\eta}(L, \boldsymbol{\theta}_0, \mathbf{h}), \tag{2}$$

где β — длина шага градиентного спуска, $\boldsymbol{\theta}^0$ — начальное значение параметров $\boldsymbol{\theta}$.

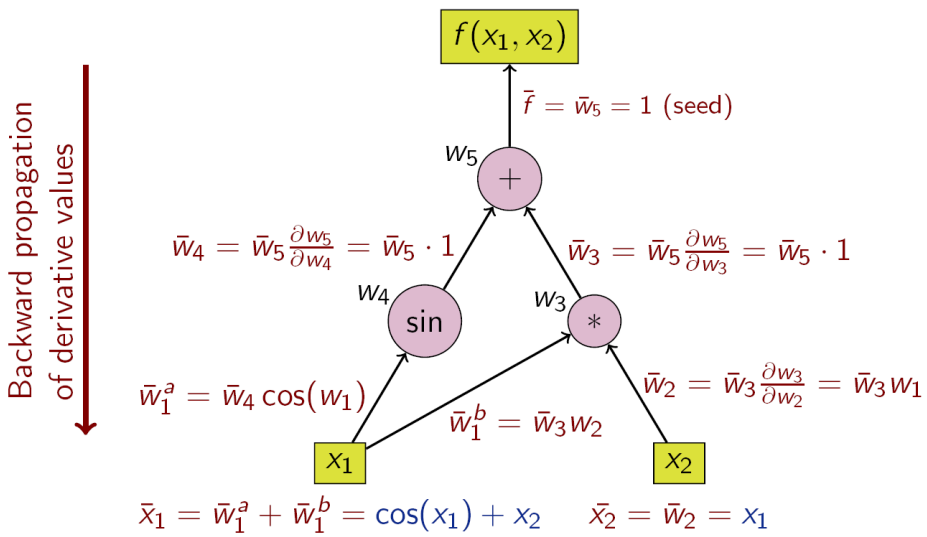
Перепишем итоговую задачу оптимизации:

$$\mathbf{h}^* = \arg\min_{\mathbf{h} \in \mathbb{R}^{\boldsymbol{h}}} Q(T^{\eta}(L, \boldsymbol{\theta}_0, \mathbf{h})).$$

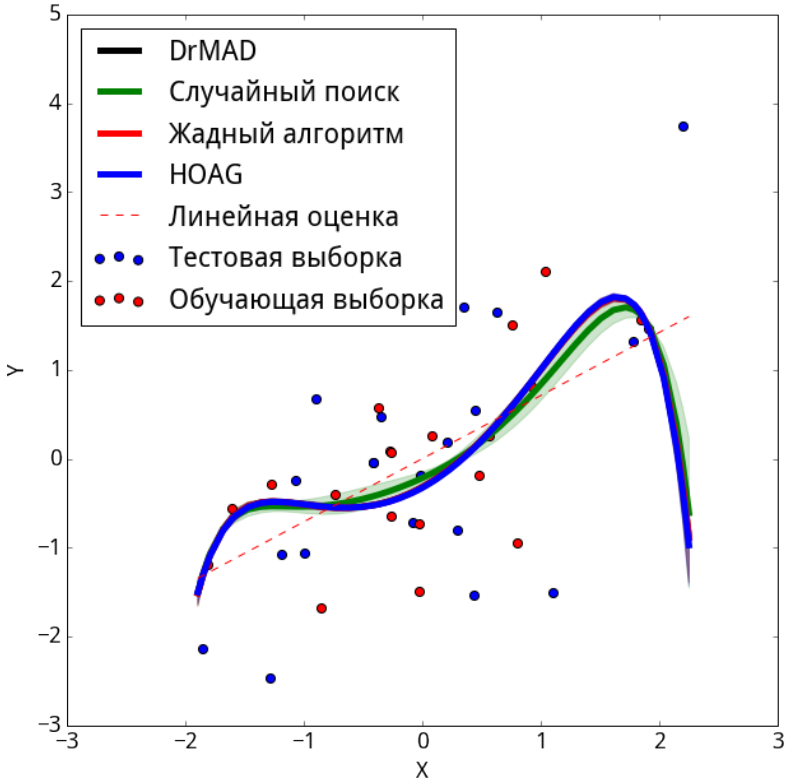
RMAD, Maclaurin et. al, 2015

1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{A}^{-1})$.
2. Положим $\hat{\nabla}\mathbf{A}^{-1} = \nabla^{-1}_{\mathbf{A}} Q(\boldsymbol{\theta}, \mathbf{A}^{-1})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}^{\tau} - \gamma \mathbf{v}^{\tau}$.
6. $\mathbf{v}^{\tau-1} = \mathbf{v}^{\tau} + \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
7. $d\mathbf{v} = \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
8. $\hat{\nabla}\mathbf{A}^{-1} = \hat{\nabla}\mathbf{A}^{-1} - d\mathbf{v} \nabla_{\mathbf{A}^{-1}} \nabla_{\boldsymbol{\theta}} Q$.
9. $\hat{\nabla}\boldsymbol{\theta} = \hat{\nabla}\boldsymbol{\theta} - d\mathbf{v} \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} Q$.

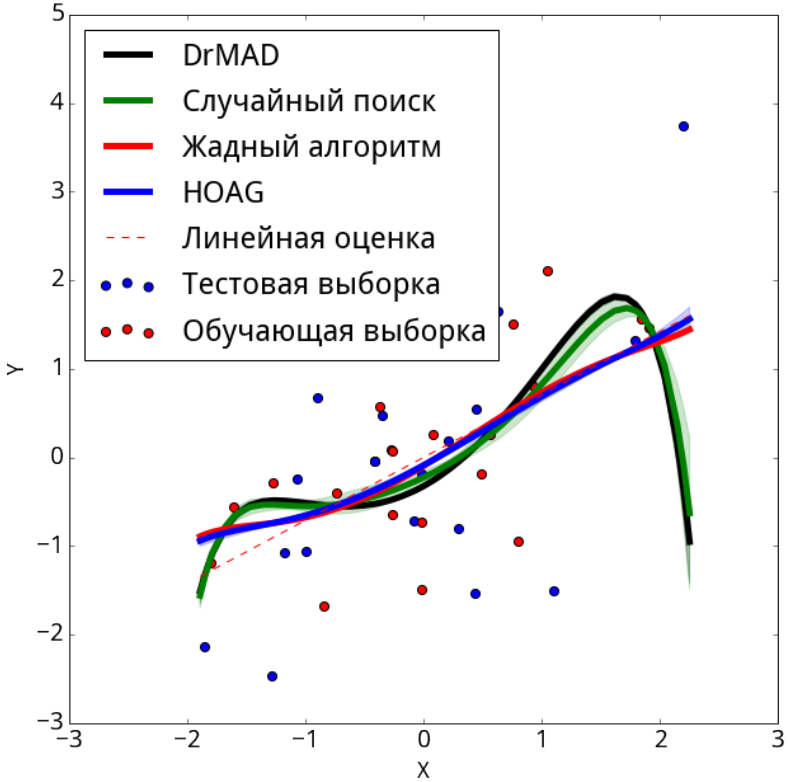
Алгоритм RMAD основывается на Reverse-mode differentiation.



Эксперименты: полиномы



Кросс-валидация



Evidence

Задача выбора структуры модели

Однослойная нейросеть:

$$\mathbf{f}(\mathbf{x}) = \text{softmax}\left(\mathbf{W}_0^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})\right), \quad f(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_{0,1}^1\mathbf{g}_{0,1}^1(\mathbf{x}) + \cdots + \gamma_{0,1}^K\mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1\boldsymbol{\sigma}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) + \cdots + \gamma_{0,1}^K\boldsymbol{\sigma}(\mathbf{W}_K^{\mathsf{T}}\mathbf{x}),$$

где $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_K]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^i\}_{i=1}^K$ — базовые функции скрытого слоя нейросети.

Структурные параметры: $\boldsymbol{\Gamma} = [\gamma_{0,1}]$.
Структура модели задается вершиной K-мерного симплекса.

Задача выбора структуры модели: два скрытых слоя

Двухслойная нейросеть:

$$\mathbf{f}(\mathbf{x}) = \text{softmax}\left(\mathbf{W}^{\mathsf{T}}\mathbf{f}_2(\mathbf{x})\right), \quad f(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

$$\mathbf{f}_2(\mathbf{x}) = \gamma_{1,2}^1 \mathbf{g}_{1,2}^1(\mathbf{f}_1(\mathbf{x})) + \cdots + \gamma_{1,2}^K \mathbf{g}_{1,2}^K(\mathbf{f}_1(\mathbf{x})) = \gamma_{1,2}^1 \sigma(\mathbf{W}_{K+1}^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})) + \cdots + \gamma_{1,2}^K \sigma(\mathbf{W}_{2K}^{\mathsf{T}}\mathbf{f}_1(\mathbf{x}))$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_{0,1}^1 \mathbf{g}_{0,1}^1(\mathbf{x}) + \cdots + \gamma_{0,1}^K \mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1 \sigma(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) + \cdots + \gamma_{0,1}^K \sigma(\mathbf{W}_K^{\mathsf{T}}\mathbf{x}),$$

где $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{2K}]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^i, \mathbf{g}_{1,2}^i\}_{i=1}^K$ — базовые функции скрытых слоев нейросети.

Структурные параметры: $\boldsymbol{\Gamma} = [\gamma_{0,1}, \gamma_{1,2}]$.
Структура модели задается вершинами двух K-мерных симплексов.

Графовое представление модели глубокого обучения

Определение

Задан граф (V, E). Для каждого ребра $(j, k) \in E$ определен вектор базовых функций мощности $K^{j,k}$:

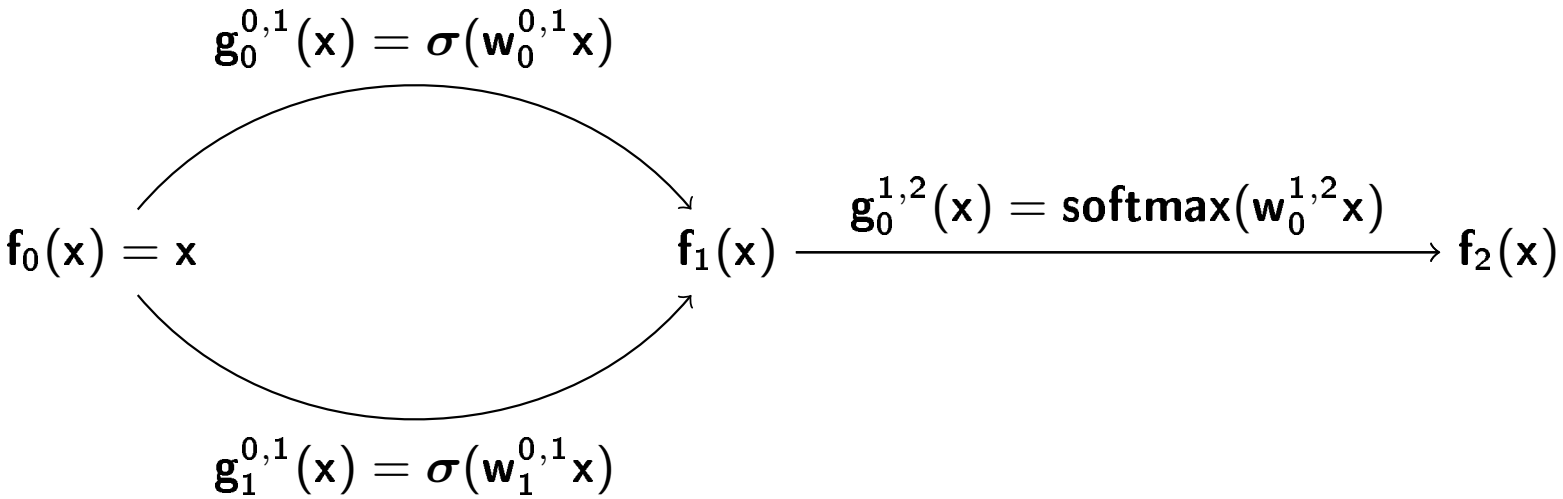
$$\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \ldots, \mathbf{g}_{Kj,k}^{j,k}]$$

. Пусть для каждой вершины $v \in V$ определена функция агрегации $\mathbf{agg}_v$. Граф (V, E) в совокупности со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и множеством функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ называется *параметрическим семейством моделей* $\mathfrak{F}$, если функция, задаваемая как

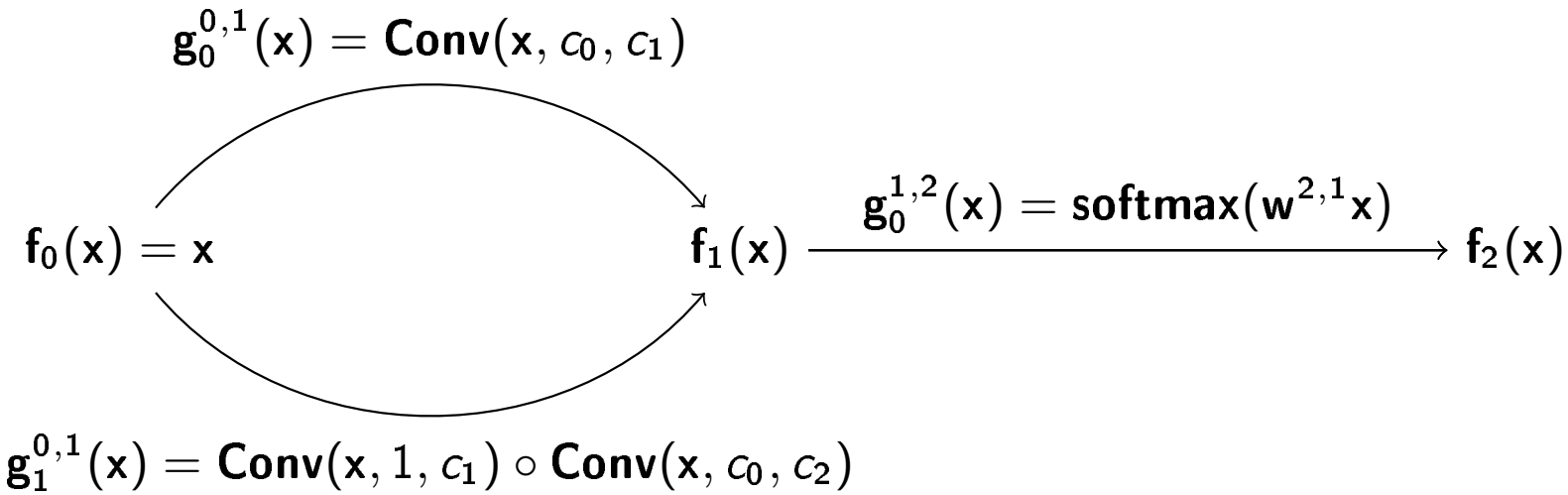
$$\mathbf{f}_k(\mathbf{x}) = \mathbf{agg}_k\left(\{\langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k}\rangle(\mathbf{f}_j(\mathbf{x})) \,|j \in \mathrm{Adj}(v_k)\}\right), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \tag{3}$$

является моделью при любых значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0, 1]^{K^{j,k}}$.

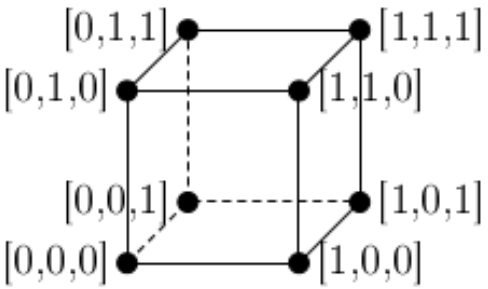
Пример: двуслойная нейросеть



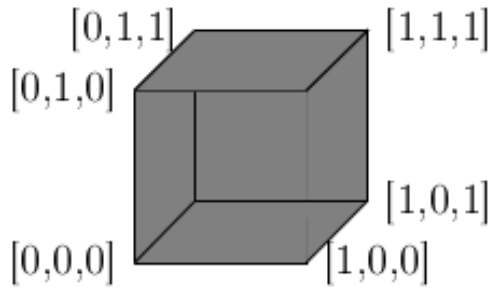
Пример: сверточная сеть



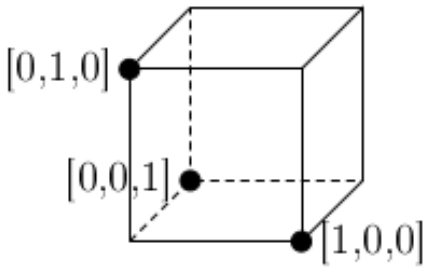
Ограничения на структурные параметры



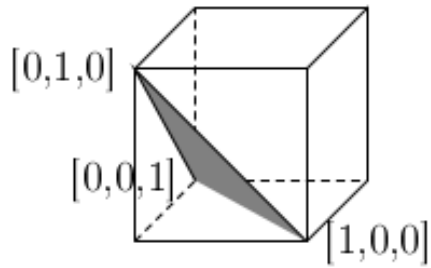
(а)



(б)



(в)



(г)

Статистические критерии качества модели

Параметрическая сложность — наименьшая дивергенция между априорным распределением параметров и апостериорным распределением параметров:

$$C_{\text{param}} = \min_{\mathbf{h}} \mathrm{D}_{\mathrm{KL}}(p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X})||p(\mathbf{W}, \boldsymbol{\Gamma}||\mathbf{h})).$$

Структурная сложность модели — энтропия апостериорного распределения структуры модели:

$$C_{\text{struct}} = -\mathrm{E}_p \log\, p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}).$$