

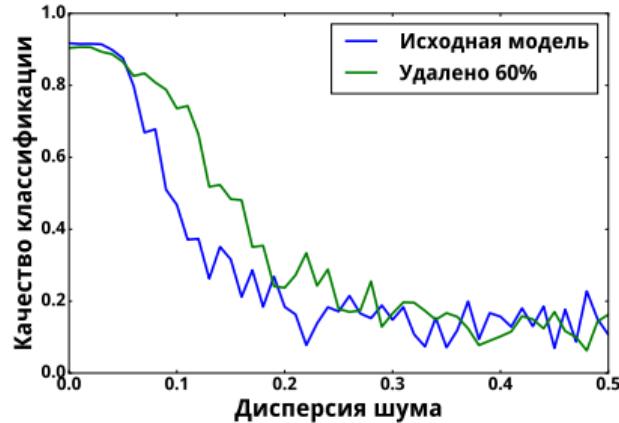
# **Выбор моделей глубокого обучения субпотимальной сложности**

Бахтеев Олег

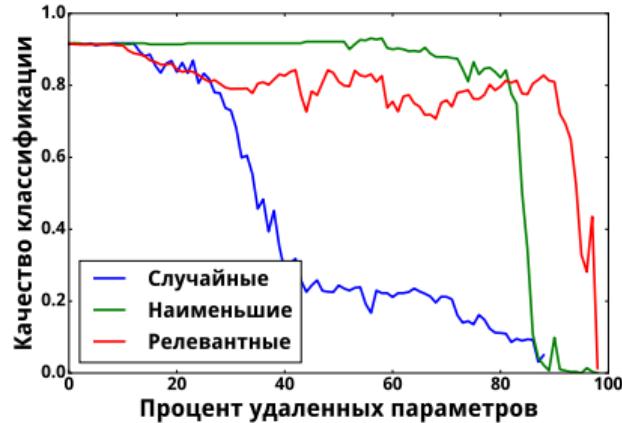
МФТИ

14.03.2018

# Сложность модели: зачем?



Устойчивость моделей при возмущении выборки



Качество классификации при удалении параметров

# Сложность модели: зачем?

Еще мотивация ???

# Принцип минимальной длины описания

$$MDL(f, \mathcal{D}) = L(f) + L(\mathcal{D}|f),$$

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

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

$w^*$  — оптимальные параметры модели.

$f_1$	$L(f_1)$	$L(w_1^* f_1)$	$L(X w_1^*, f_1)$
$f_2$	$L(f_2)$	$L(w_2^* f_2)$	$L(X w_2^*, f_2)$
$f_3$	$L(f_3)$	$L(w_3^* f_3)$	$L(X w_3^*, f_3)$

# 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.

# Оптимальная универсальная модель MDL

Пусть выборка  $\mathfrak{D}$  лежит в некотором конечном множестве.

$$MDL(f, \mathfrak{D}) = L(\mathfrak{D} | w^*(\mathfrak{D}), f) + COMP(f),$$

$$L(\mathfrak{D} | w^*, f) = -\log p(\mathfrak{D} | w^*(\mathfrak{D}), f), \quad COMP = \log \sum_{\mathfrak{D}' \in \mathbb{X}} P(\mathfrak{D}' | w^*(\mathfrak{D}'), f).$$

В случае, если распределение  $p(\mathfrak{D} | w)$  принадлежит экспоненциальному семейству, оценка MDL совпадает с точностью до  $o(1)$  с байесовской оценкой правдоподобия (“Evidence”):

$$p(\mathfrak{D} | f) = \int_w p(\mathfrak{D} | w) p(w) d\mathbf{w},$$

где  $p(w)$  — априорное распределение специального вида:

$$p(w) = \frac{\sqrt{|J(w)|}}{\int_{w'} \sqrt{|J(w')|} d\mathbf{w}'},$$

$J(w)$  — информация Фишера.

# Байесовский подход к сложности

Правдоподобие модели (“Evidence”):

$$p(\mathcal{D}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathbf{f})d\mathbf{w}.$$

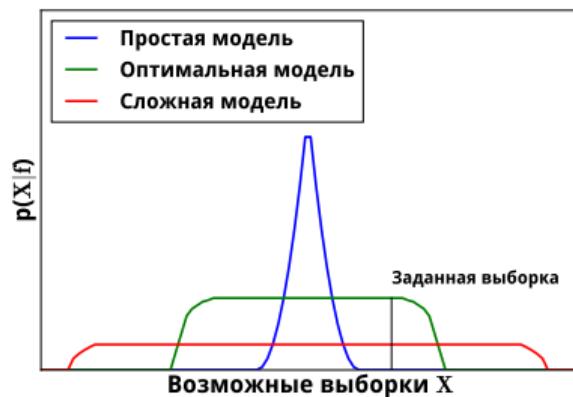
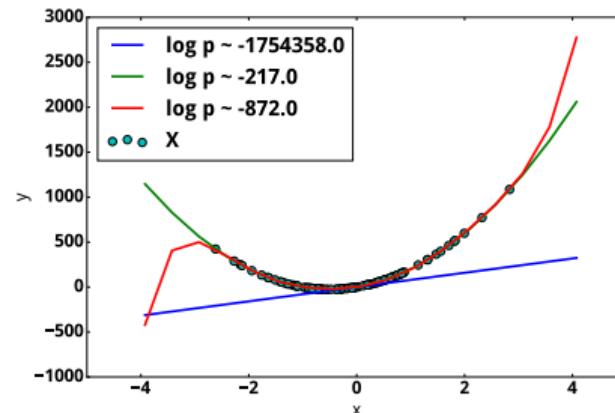


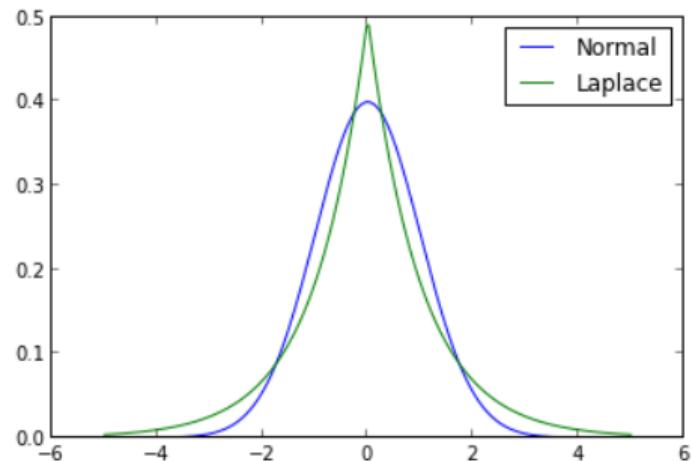
Схема выбора модели по правдоподобию



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



# Evidence vs Кросс-валидация

Оценка Evidence:

$$\log p(\mathcal{D}|\mathbf{f}) = \log p(\mathcal{D}_1|\mathbf{f}) + \log p(\mathcal{D}_2|\mathcal{D}_1, \mathbf{f}) + \dots + \log p(\mathcal{D}_n|\mathcal{D}_1, \dots, \mathcal{D}_{n-1}, \mathbf{f}).$$

Оценка leave-one-out:

$$LOU = E \log p(\mathcal{D}_n|\mathcal{D}_1, \dots, \mathcal{D}_{n-1}, \mathbf{f}).$$

Кросс-валидация использует среднее значение последнего члена  $p(\mathcal{D}_n|\mathcal{D}_1, \dots, \mathcal{D}_{n-1}, \mathbf{f})$  для оценки сложности.

Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

# Оптимальность модели

## Определение

Пусть задано множество моделей  $\mathfrak{F}$ .

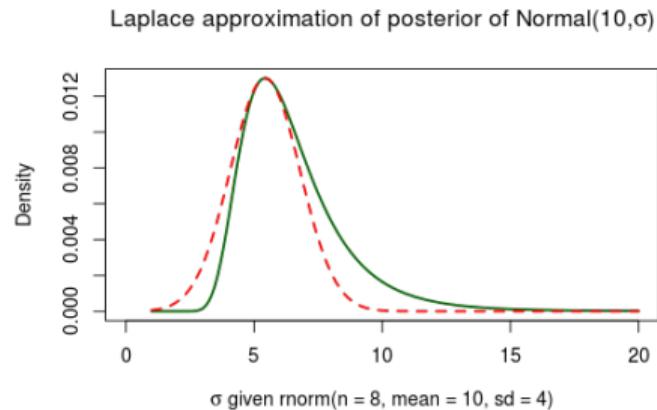
Пусть для каждой модели  $f$  задано априорное распределение параметров:  $p(\mathbf{w}|f)$ .

Модель  $f$  назовем оптимальной среди моделей  $\mathfrak{F}$ , если достигается максимум интеграла:

$$p(\mathcal{D}|f) = \int_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|f)d\mathbf{w}.$$

# Методы получения оценок Evidence: метод Лапласа

$$p(\mathfrak{D}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{f}) = \int_{\mathbf{w}} \exp(-S(\mathbf{w})) \sim \exp S(\hat{\mathbf{w}}) \int_{\mathbf{w}} \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T \nabla\nabla S(\hat{\mathbf{w}}) \Delta\mathbf{w}\right).$$



# Методы получения оценок Evidence: Метод Монте-Карло

$$p(\mathcal{D}|\mathbf{f}) \sim \frac{1}{K} \sum_{\mathbf{w} \in W} p(\mathcal{D}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}),$$

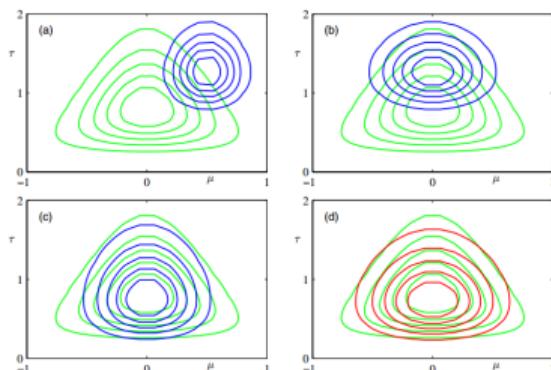
$W$  — множество векторов параметров мощностью  $K$ .

- Плохо работает в пространствах большой размерности
- Существует ряд модификаций, позволяющих преодолеть проклятие размерности
- Может применяться в связке с вариационным выводом

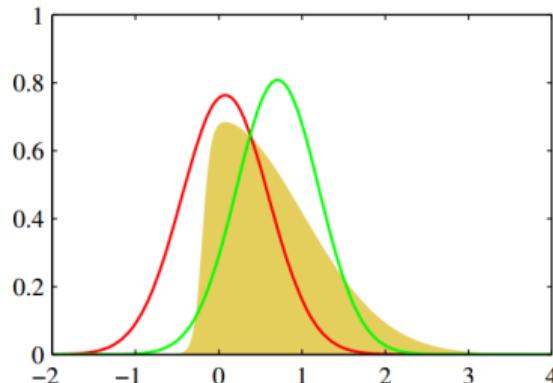
# Вариационная оценка

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

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



Аппроксимация неизвестного  
распределения нормальным



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

# Получение вариационной нижней оценки

$$\begin{aligned}\log p(\mathcal{D}|\mathbf{f}) &= \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathcal{D}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} + D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathcal{D}, \mathbf{f})) \geq \\ &\geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathcal{D}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \\ &= -D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathcal{D}|\mathbf{w}, \mathbf{f}) d\mathbf{w},\end{aligned}$$

где

$$D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) = - \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w}.$$

## Определение

Модель  $\mathbf{f}$  назовем субоптимальной на множестве моделей  $\mathfrak{F}$  по множеству распределений  $Q$ , если модель доставляет максимум нижней вариационной оценке:

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

# $D_{KL}$

Максимизация вариационной нижней оценки

$$\int_w q(w) \log \frac{p(\mathcal{D}, w|\mathbf{f})}{q(w)} dw$$

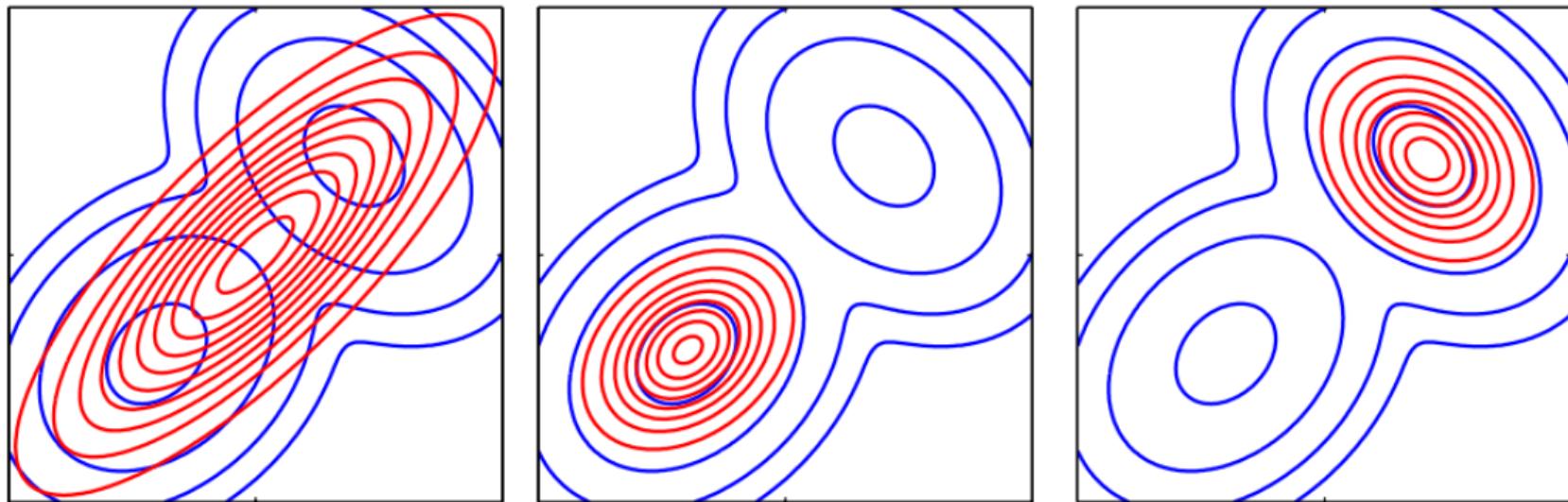
эквивалентна минимизации дивергенции между распределением распределением  $q(w) \in Q$  и апостериорным распределением параметров  $p(w|\mathcal{D}, \mathbf{f})$ :

$$q = \operatorname{argmax}_{q \in Q} \int_w q(w) \log \frac{p(\mathcal{D}, w|\mathbf{f})}{q(w)} dw \Leftrightarrow q = \operatorname{argmin}_{q \in Q} D_{KL}(q(w)||p(w|\mathcal{D}, \mathbf{f})),$$

т.к.

$$\log p(\mathcal{D}|\mathbf{f}) = \int_w q(w) \log \frac{p(\mathcal{D}, w|\mathbf{f})}{q(w)} dw + D_{KL}(q(w)||p(w|\mathcal{D}, \mathbf{f})) = \text{const.}$$

# Пример: аппроксимация мультимодального распределения



# 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.**

# Evidence: нормальное распределение

Пусть  $q \sim \mathcal{N}(\mu_q, \mathbf{A}_q)$ .

Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{Y}|\mathfrak{D}, \mathbf{w}, \mathbf{f}) d\mathbf{w} + D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) \simeq$$
$$\sum_{i=1}^m \log p(\mathbf{y}_i|\mathfrak{D}_i, \mathbf{w}_i) + D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) \rightarrow \max_{\mathbf{A}_q, \mu_q},$$

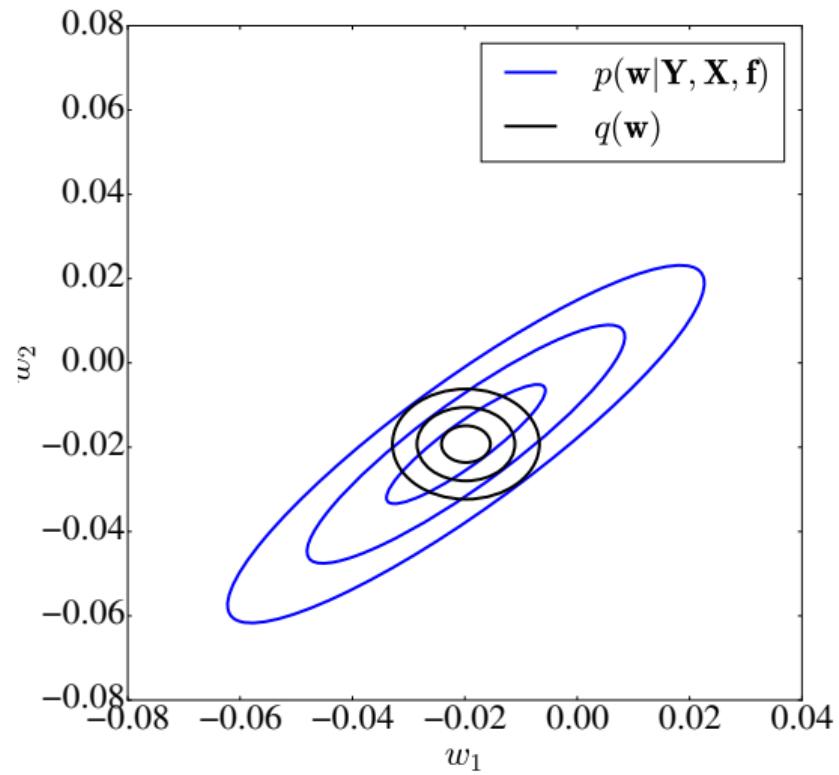
В случае, если априорное распределение параметров  $p(\mathbf{w}|\mathbf{f})$  является нормальным:

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

дивергенция  $D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f}))$  вычисляется аналитически:

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

# Evidence: нормальное распределение



# Graves, 2011

Априорное распределение:  $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\mu, \sigma\mathbf{I})$ .

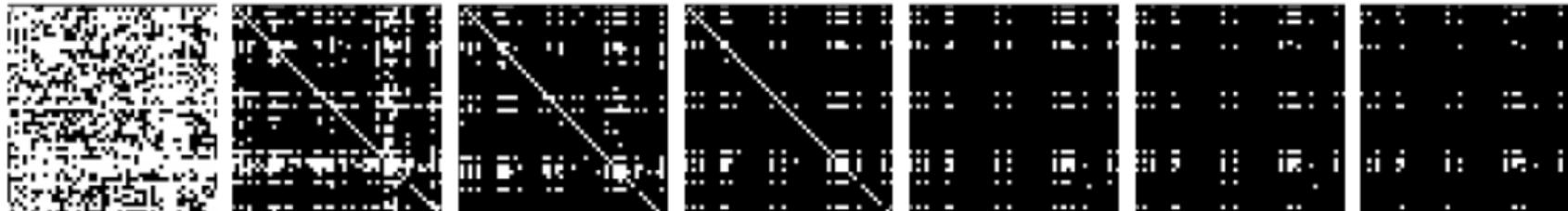
Вариационное распределение:  $q(\mathbf{w}) \sim \mathcal{N}(\mu_q, \sigma_q\mathbf{I})$ .

Жадная оптимизация гиперпараметров:

$$\mu = \hat{\mathbf{E}}\mathbf{w}, \quad \sigma = \hat{\mathbf{D}}\mathbf{w}.$$

Прунинг параметра  $w_i$  определяется относительной плотностью:

$$\lambda = \frac{q(\mathbf{0})}{q(\mu_{i,q})} = \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right).$$



$\lambda = 0.01$

$\lambda = 0.05$

$\lambda = 0.1$

$\lambda = 0.2$

$\lambda = 0.5$

$\lambda = 1$

$\lambda = 2$

**Априорное распределение** задается для каждого нейрона отдельно:

$$p(w_{ij}|\sigma) \sim \mathcal{N}(0, z), \quad p(z_i) \propto |z_i|^{-1}.$$

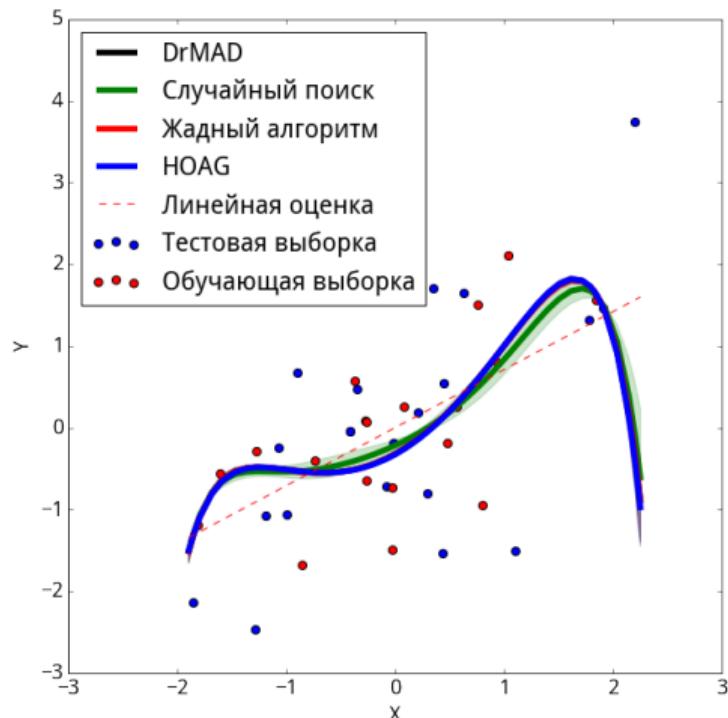
$$p(\mathbf{w}, \mathbf{z}) \propto \prod_i \frac{1}{|z_i|} \prod_j \mathcal{N}(w_{i,j}|0, z_i^2).$$

**Вариационное распределение:**  $q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_q}, \boldsymbol{\sigma}_{\mathbf{z}_q} \mathbf{I}), \quad q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q \mathbf{I}).$

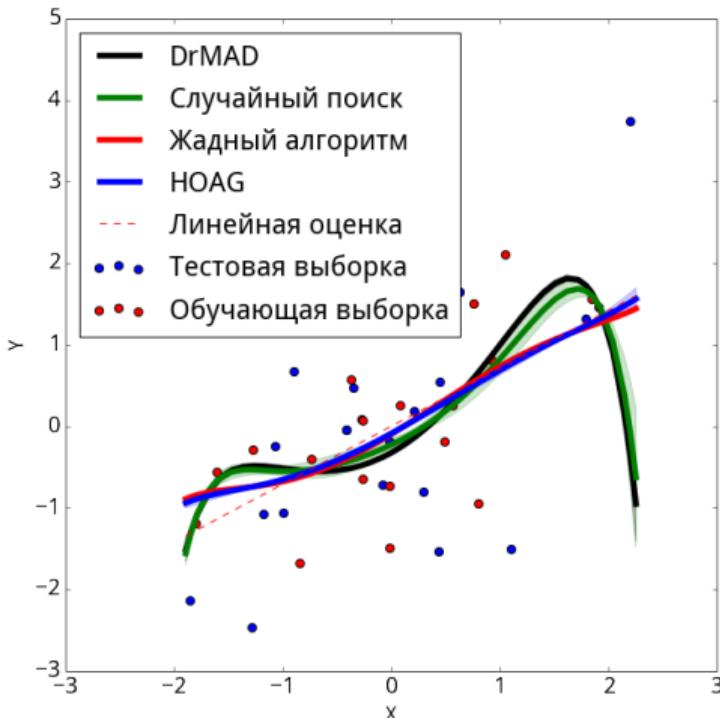
Прунинг нейронов  $\mathbf{w}_i$  определяется величиной

$$\frac{\sigma_{q,i}^2}{\mu_{q,i}^2}.$$

# Кросс-Валидация vs. Evidence: отбор признаков



Кросс-валидация



Evidence

# 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:

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

# Градиентный спуск для оценки правдоподобия

Рассмотрим максимизацию совместного распределения параметров:

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

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

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

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

# Энтропия

Формулу вариационной оценки можно переписать с использованием энтропии:

$$\log p(\mathcal{D}|\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathcal{D}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \\ E_{q(\mathbf{w})}[\log p(\mathcal{D}, \mathbf{w}|\mathbf{f})] - S(q(\mathbf{w})),$$

где  $S(q(\mathbf{w}))$  — энтропия:

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

# Градиентный спуск для оценки правдоподобия

При достаточно малой длине шага оптимизации  $\gamma$  разность энтропии на различных шагах оптимизации вычисляется как:

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

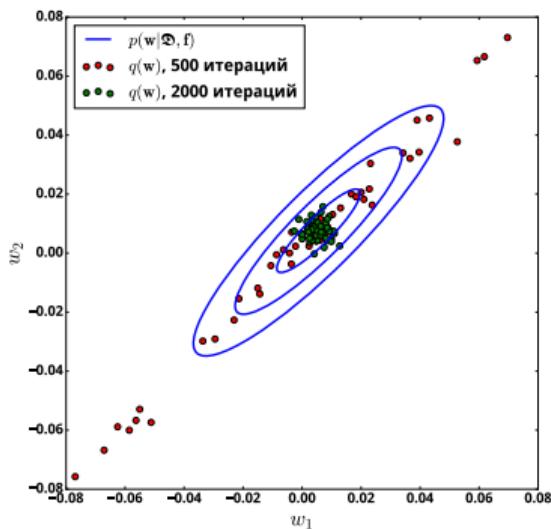
Итоговая оценка на шаге оптимизации  $\tau$ :

$$\log \hat{p}(\mathbf{Y}|\mathcal{D}, \mathbf{f}) \sim \frac{1}{r} \sum_{g=1}^r L(\mathbf{w}_\tau^g, \mathcal{D}, \mathbf{Y}) + S(q^0(\mathbf{w})) + \frac{1}{r} \sum_{b=1}^\tau \sum_{g=1}^r (-\gamma \text{Tr}[\mathbf{H}(\mathbf{w}_b^g)] - \gamma^2 \text{Tr}[\mathbf{H}(\mathbf{w}_b^g) \mathbf{H}(\mathbf{w}_b^g)]),$$

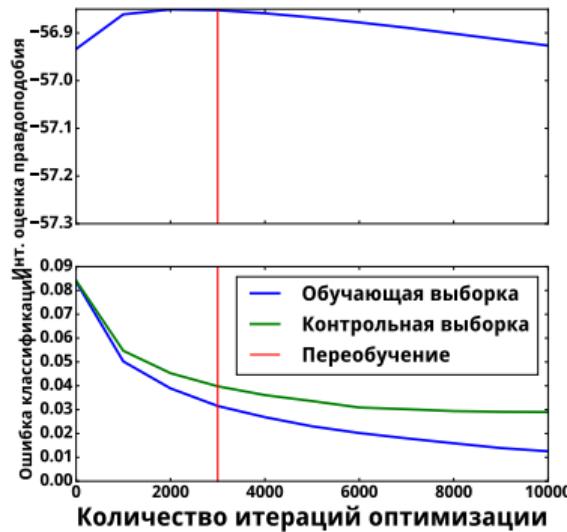
$\mathbf{w}_b^g$  — вектор параметров старта  $g$  на шаге  $b$ ,  $S(q^0(\mathbf{w}))$  — начальная энтропия.

# Переобучение, Maclaurin et. al, 2015

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



Схождение распределения к моде



Оценка начала переобучения

# Стохастическая динамика Ланжевена

Модификация стохастического градиентного спуска:

$$T(\mathbf{w}) = \mathbf{w} - \gamma \nabla(\log p(\mathbf{w}) + \frac{m}{\hat{m}} \log p(\hat{\mathcal{D}}|\mathbf{w})) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\gamma}{2})$$

где  $\hat{m}$  — размер подвыборки,  $\hat{\mathcal{D}} \subset \mathcal{D}$  — подвыборка, шаг оптимизации  $\gamma$  изменяется с количеством итераций:

$$\sum_{\tau=1}^{\infty} \gamma_{\tau} = \infty, \quad \sum_{\tau=1}^{\infty} \gamma_{\tau}^2 < \infty.$$

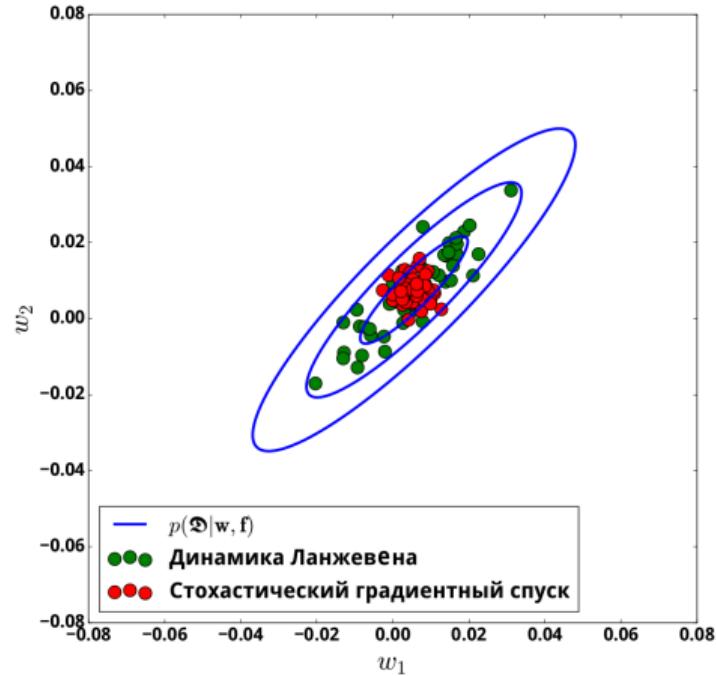
**Утверждение [Welling, 2011].** Распределение  $q^{\tau}(\mathbf{w})$  сходится к апостериорному распределению  $p(\mathbf{w}|\mathcal{D}, \mathbf{f})$ .

Изменение энтропии с учетом добавленного шума:

$$\hat{S}(q^{\tau}(\mathbf{w})) \geq \frac{1}{2} |\mathbf{w}| \log \left( \exp \left( \frac{2S(q^{\tau}(\mathbf{w}))}{|\mathbf{w}|} \right) + \exp \left( \frac{2S(\epsilon)}{|\mathbf{w}|} \right) \right).$$

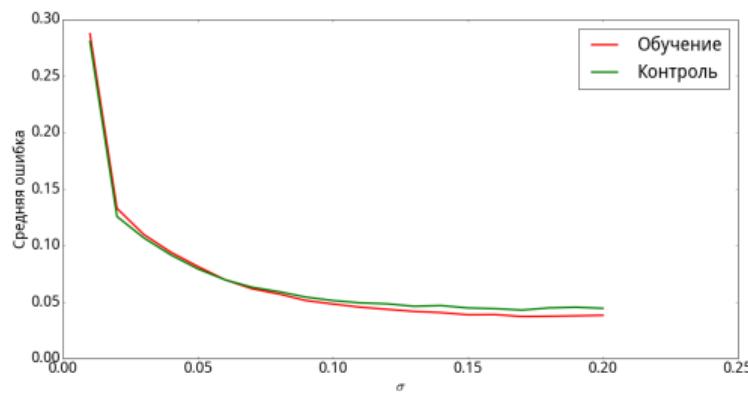
# Стохастическая динамика Ланжевена

Распределения параметров после 2000 итераций:

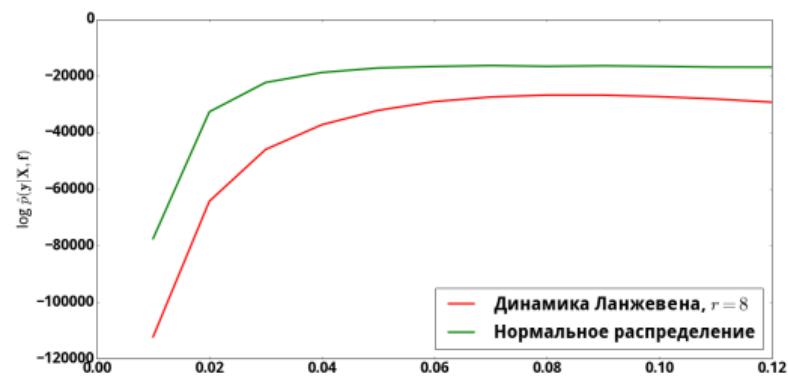


# Выбор константы регуляризации

Выборка MNIST, 50 нейронов на скрытом слое.

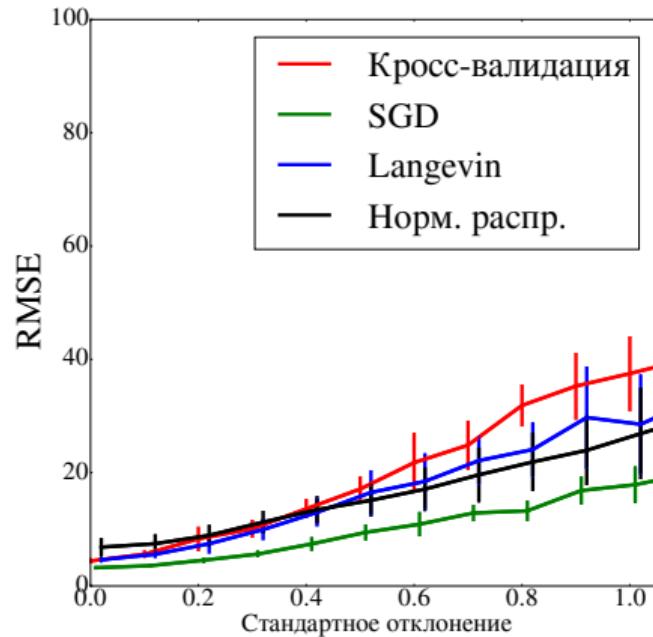


Кросс-валидация

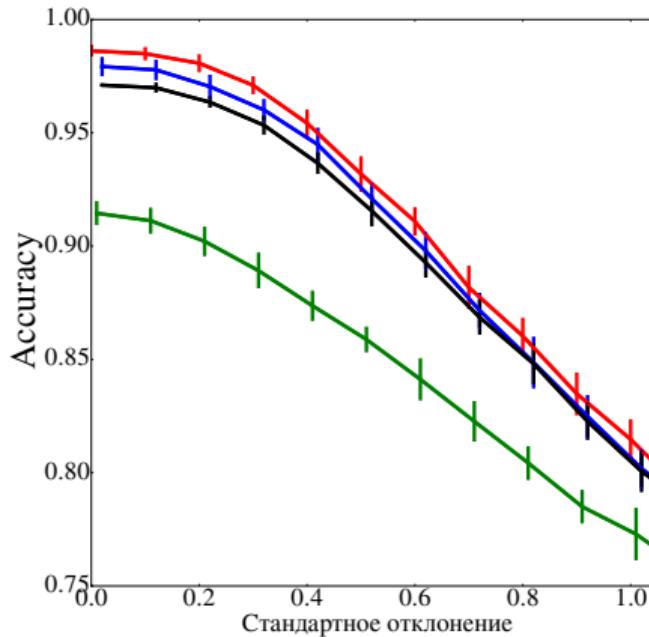


Оценка Evidence

# Качество моделей при возмущении параметров



Boston: 3-слойная нейросеть



MNIST (50-dim PCA): 3-слойная нейросеть

# Используемые материалы

- ① David J. C. MacKay, Information Theory, Inference & Learning Algorithms, 2003
- ② Peter Grunwald, A tutorial introduction to the minimum description length principle, 2004
- ③ Christopher Bishop, Pattern Recognition and Machine Learning, 2006
- ④ Welling W., Teh Y.W., Bayesian Learning via Stochastic Gradient Langevin Dynamics, 2011
- ⑤ Alex Graves, Practical Variational Inference for Neural Networks, 2011
- ⑥ Maclaurin D., Duvenaud D., Adams R.P., Early Stopping is Nonparametric Variational Inference, 2015
- ⑦ Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation, 2016
- ⑧ Louizos C., Ullrich K., Welling M., Bayesian Compression for Deep Learning, 2017
- ⑨ Бахтеев О.Ю., Стрижов В.В., Выбор моделей глубокого обучения субоптимальной сложности, 2018