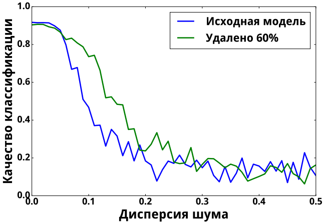
# Выбор модели глубокого обучения

Бахтеев Олег

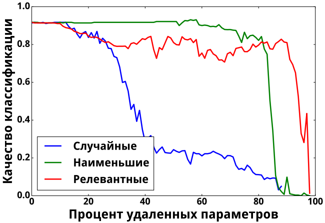
МФТИ

16.10.2019

# 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.

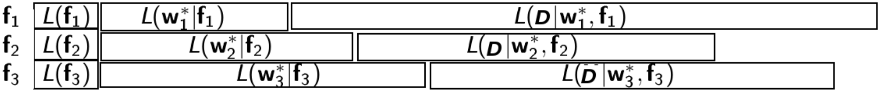
# Принцип минимальной длины описания

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

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

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

$\mathbf{w}^*$ — 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.

# Связанный байесовский вывод

*Первый уровень:* выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

*Второй уровень:* выбираем модель, доставляющую максимум обоснованности модели.

Обоснованность модели ("Evidence"):

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

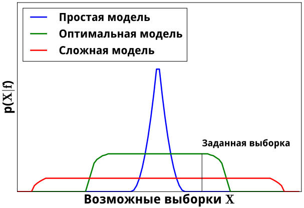
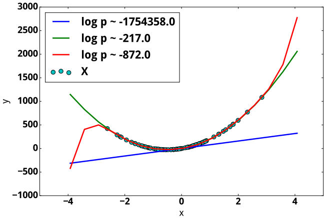


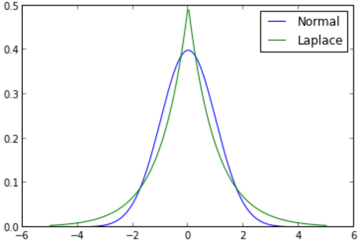

Схема выбора модели




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**

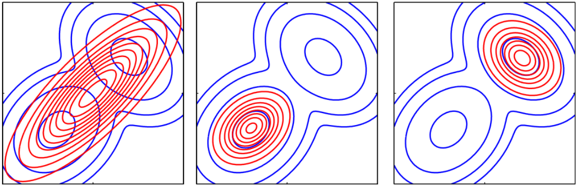
Пусть задано множество моделей $M$.
Пусть для каждой модели $\mathbf{f}$ задано априорное распределение параметров: $p(\mathbf{w}|\mathbf{h})$, где $\mathbf{h}$ — параметры априорного распределения.
Модель $\mathbf{f}$ назовем оптимальной среди моделей $M$, если достигается максимум интеграла:

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

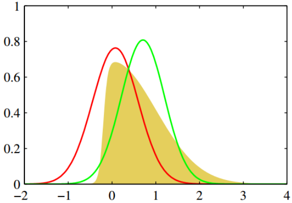
# Вариационная оценка, ELBO

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

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



Вариационный вывод и expectation propogation (Bishop)




Аппроксимация Лапласа и вариационная оценка, зеленая линия (Bishop)

# Получение вариацонной нижней оценки

**Утверждение 1**

Максимизация вариационной нижней оценки

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

эквивалентна минимизации расстояния Кульбака–Лейблера между распределением $q(\mathbf{w}) \in \mathfrak{Q}$ и апостериорным распределением параметров $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})$:

$$\hat{q} = \arg\max_{q \in \mathfrak{Q}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow \hat{q} = \arg\min_{q \in \mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\big),$$

$$\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\big) = \int_{\mathbf{w}} q(\mathbf{w}) \log \left(\frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})}\right) d\mathbf{w}.$$

**Определение**

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

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h})}{q(\mathbf{w})} d\mathbf{w}.$$

# Вариационная оценка и эффективный размер выборки

**Утверждение 2**

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

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \lambda \mathsf{D}_{\mathsf{KL}}\left(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\right)$$

эквивалентна оптимизации вариационной оценки обоснованности для произвольной случайной подвыборки $\hat{\mathbf{y}}, \hat{\mathbf{X}}$ мощности $\frac{m}{\lambda}$ из генеральной совокупности.

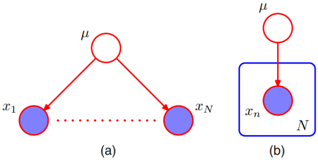
См. также [Alemi et al., 2017, Fixing Broken ELBO].

# Plate notation

Plate notation — формат представления вероятностных моделей, альтернативный вероятностным графам.

Элементы:

- Белые кружки (случайные величины);
- Серые кружки (наблюдаемые реализации случайной величины);
- Маленькие кружки (неслучайные величины);
- Плитки (дублирование вероятностного вывода).




DAG и Plate notation (Bishop)

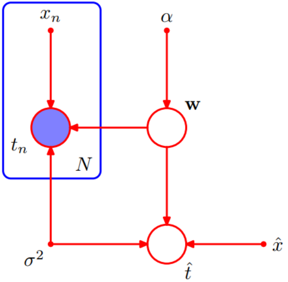
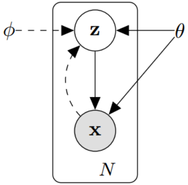



Plate notation для модели регрессии (Bishop)

# Вариационный автокодировщик

Пусть объекты выборки $\mathbf{X}$ порождены при условии скрытой переменной $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:




$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{h}, \mathbf{w}).$$

$p(\mathbf{h}|\mathbf{x}, \mathbf{w})$ — неизвестно.
Будем максимизировать вариационную оценку правдоподобия выборки:

$$\log p(\mathbf{x}|\mathbf{w}) \geq \mathsf{E}_{q_\phi(\mathbf{h}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{h}, \mathbf{w}) - D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{x}) || p(\mathbf{h})) \to \max .$$

Распределения $q_\phi(\mathbf{h}|\mathbf{x})$ и $p(\mathbf{x}|\mathbf{h}, \mathbf{w})$ моделируются нейросетью:

$$q_\phi(\mathbf{h}|\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{h}, \mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_w(\mathbf{h}), \boldsymbol{\sigma}^2_w(\mathbf{h})),$$

где функции $\boldsymbol{\mu}$,$\boldsymbol{\sigma}$ — выходы нейросети.

# 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:

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

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

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

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

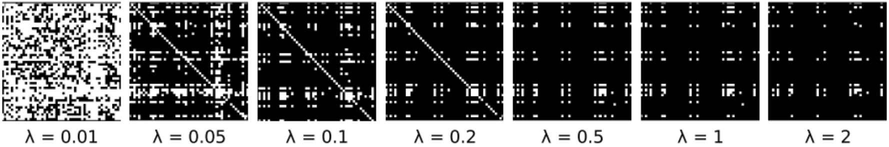
# Graves, 2011

**Априорное распределение:** $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma\mathbf{I})$.
**Вариационное распределение:** $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q\mathbf{I})$.
Жадная оптимизация гиперпараметров:

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

Прунинг параметра $w_i$ определяется относительной плотностью:

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



# ELBO: нормальное распределение

**“Обычная” функция потерь:**

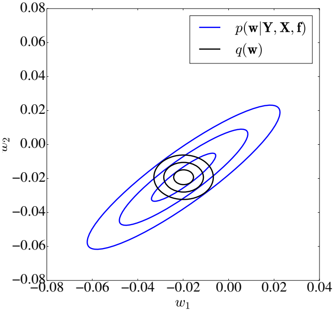
$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} -\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \lambda||\mathbf{w}||_2^2.$$

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

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

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

Пример грубой аппроксимации нормальным диагональным распределением $q$



# MCMC и вариационный вывод

**Идея MCMC:** Порождаем сэмплы из простого распределения и принимаем их, если заданное отношение больше порога:

$$\min\left(1, \frac{p(\mathbf{w}^\tau|\mathbf{y}, \mathbf{X}, \mathbf{h})}{p(\mathbf{w}^{\tau-1}|\mathbf{y}, \mathbf{X}, \mathbf{h})}\right),$$

где $\mathbf{w}^\tau$ выбирается на основе предыдущего сэмпла:

$$\mathbf{w}^\tau = T(\mathbf{w}^{\tau-1}).$$

**Salimans et al., 2014:** будем интерпретировать последовательность применения оператора $T$ как оптимизацию вариационной оценки:

$$T^1 \circ \dots T^\eta(\mathbf{w}) \to p(\mathbf{w}^\tau|\mathbf{y}, \mathbf{X}, \mathbf{h}).$$

**Maclaurin et. al, 2015:** в качестве оператора $T$ будем рассматривать оператор оптимизации. Откажемся от отклонения сэмплов по порогу.

# 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 $\mathbf{w}_1, \dots, \mathbf{w}_r$ 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:

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

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

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

**Утверждение 3**

Пусть $L$ — липшицева функция, оператор оптимизации — биекция. Тогда разность энтропии на различных шагах оптимизации вычисляется как:

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

Итоговая оценка на шаге оптимизации $\tau$:

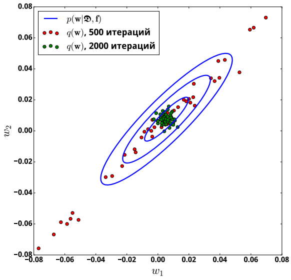
$$\log \hat{p}(\mathbf{Y}|\mathfrak{D}, \mathbf{h}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathbf{w}_{\tau}^{g}, \mathfrak{D}, \mathbf{Y}) + \mathsf{S}(q^0(\mathbf{w})) +$$

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

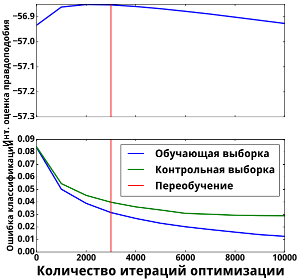
$\mathbf{w}_b^g$ — вектор параметров старта $g$ на шаге $b$, $\mathsf{S}(q^0(\mathbf{w}))$ — начальная энтропия.

# 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

# Стохастическая динамика Ланжевена

Модификация стохастического градиентного спуска:

$$T = \mathbf{w} - \beta \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

где шаг оптимизации $\alpha$ изменяется с количеством итераций:

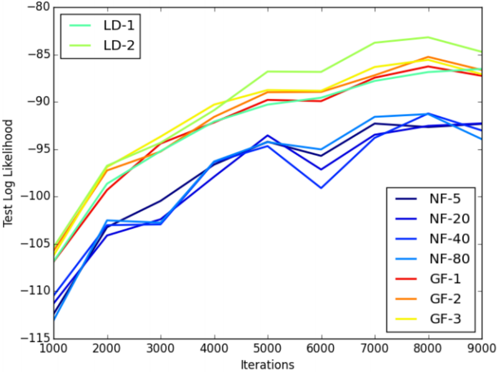
$$\sum_{\tau=1}^{\infty} \beta_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \beta_\tau^2 < \infty.$$

**Утверждение [Welling, 2011].** Распределине $q^\tau(\mathbf{w})$ сходится к апостериорному распределению $p(\mathbf{w}|\mathbf{X}, \mathbf{f})$.
Изменение энтропии с учетом добавленного шума:

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

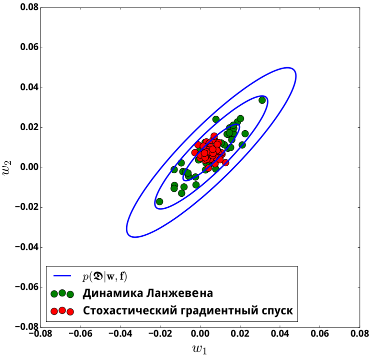
# Стохастическая динамика Ланжевена в генеративных моделях

**Altieri et al., 2015:** будем сэмплировать скрытую переменную **z** и приближать его распределение к максимуму вариационной оценки с использованием динамики Ланжевена.



# Стохастическая динамика Ланжевена

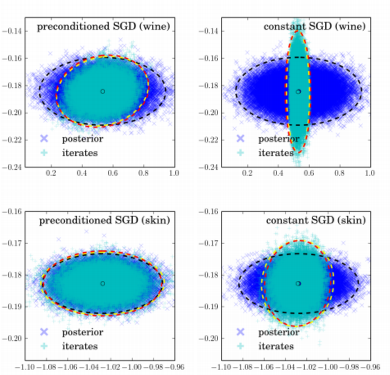
Распределения параметров после 2000 итераций:



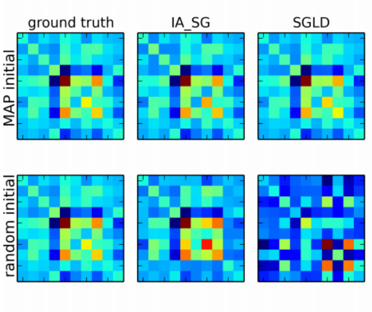

**Проблема:** медленная сходимость динамики.

# SGD с оптимизацией длины шага

**Mandt et al., 2017:** вблизи точки экстремума градиентный спуск приближает апостериорное распределение параметров модели. Существуют оценки на длину шага градиентного спуска.




SGD с разным типом длин шагов




Сравнение с динамикой Ланжевена

# Список источников


- Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V., 2018. Learning transferable architectures for scalable image recognition
- David J. C. MacKay, Information Theory, Inference & Learning Algorithms
- Peter Grunwald, A tutorial introduction to the minimum description length principle
- Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation
- Christopher Bishop, Pattern Recognition and Machine Learning
- Diederik P Kingma, Max Welling, Auto-Encoding Variational Bayes
- Dougal Maclaurin, David Duvenaud, Ryan P. Adams, Early Stopping is Nonparametric Variational Inference
- Max Welling, Yee Whye Teh, Bayesian Learning via Stochastic Gradient Langevin Dynamics

# Список источников


- A. Graves, Practical Variational Inference for Neural Networks
- Salimans, Tim, Diederik Kingma, and Max Welling, 2015. Markov chain monte carlo and variational inference: Bridging the gap
- Altieri: http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf
- Stephan Mandt, Matthew D. Hoffman, David M. Blei, 2017. Stochastic Gradient Descent as Approximate Bayesian Inference
- О. Ю. Бахтеев, В. В. Стрижов, “Выбор моделей глубокого обучения субоптимальной сложности”
- А. Н. Смердов, О. Ю. Бахтеев, В. В. Стрижов, “Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза”

# ДЗ: выбор задания

**Дедлайн: 23 октября, 0 часов.**

```
from zlib import crc32

theory = crc32('фамилия кириллицей'.lower().encode('utf-8'))%3+1

practice = crc32('фамилия латиницей'.lower().encode('utf-8'))%3+1
```

Задания заливаются на github:
*https://github.com/Intelligent-Systems-Phystech/model_selection/фамилия латиницей*

# ДЗ: теория

**Формат: tex + pdf.**

1. Доказать утверждение 1;
   - Воспользоваться Bishop.
2. Доказать утверждение 2;
   - Воспользоваться УЗБЧ.
3. Доказать утверждение 3;
   - Воспользоваться разложением по Тейлору и свойством энтропии распределения под действием биекции (https://en.wikipedia.org/wiki/Differential_entropy)

# ДЗ: практика

**Формат: ipynb.**

1. Реализовать пример выбора модели с аппроксимацией Лапласа (Bishop/McKay).
2. Реализовать пример выбора модели с вариационным нормальным распределением (Graves).
3. Реализовать пример выбора модели с распределением под дейвтием градиентного спуска (Maclaurin).

При оценивании будут учитываться аккуратность кода ноутбуков и наглядность примера.

Пример должен быть выполнен на **простых** игрушечных синтетических данных.