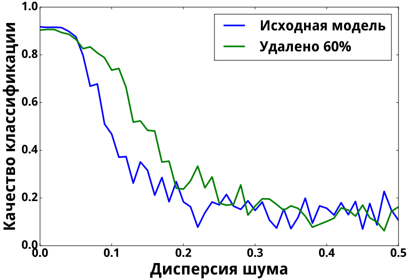
Выбор моделей глубокого обучения субпотимальной сложности

Бахтеев Олег

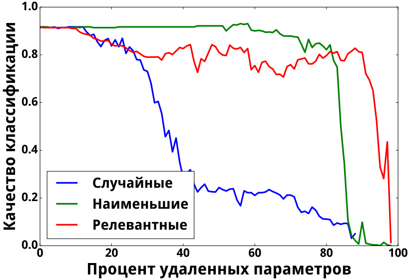
МФТИ

14.03.2018

Сложность модели: зачем?



Устойчивость моделей при возмущении выборки



Качество классификации при удалении параметров

Сложность модели: зачем?

Еще мотивация ???

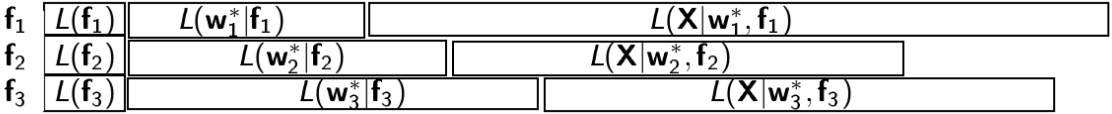
Принцип минимальной длины описания

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

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

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

$\mathbf{w}^*$ — 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.

Оптимальная универсальная модель MDL

Пусть выборка $\mathbf{X}$ лежит в некотором конечном множестве $\mathbb{X}: \mathbf{X} \subset \mathbb{X}$.

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) = L(\mathbf{X}|\mathbf{w}^*(\mathbf{X}), \mathbf{f}) + \mathrm{COMP}(\mathbf{f}),$$

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

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

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w},$$

где $p(\mathbf{w})$ — априорное распределение специанльного вида:

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

$J(\mathbf{w})$ — информация Фишера.

Байесовый подход к сложности

Правдоподобие модели (“Evidence”):

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

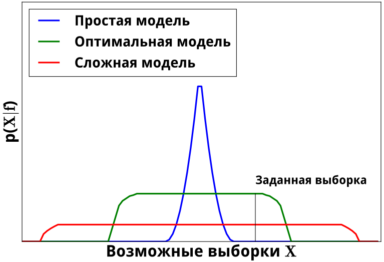
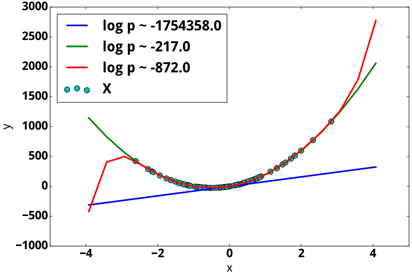


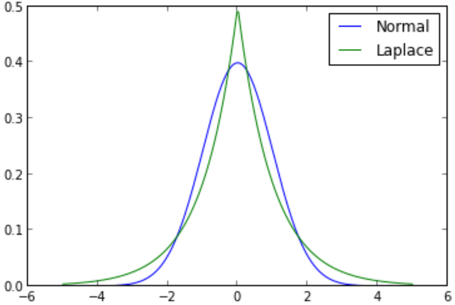
Схема выбора модели по правдоподобию



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 Evidece:

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

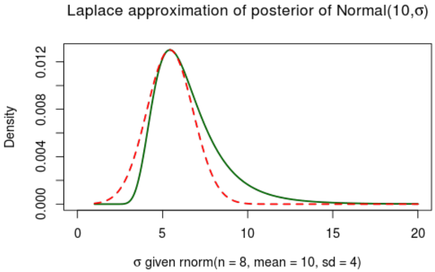
Оценка leave-one-out:

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

Кросс-валидация использует среднее значение последнего члена $p(\mathbf{x}_n|\mathbf{x}_1, \dots, \mathbf{x}_{n-1}, \mathbf{f})$ для оценки сложности.
Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

Методы получения оценок Evidence: метод Лапласа

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



Методы получения оценок Evidence: Метод Монте-Карло

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

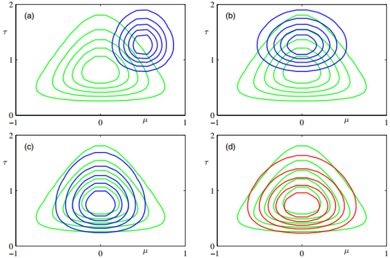
$\mathbf{W}$ — множество векторов параметров мощностью K.

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

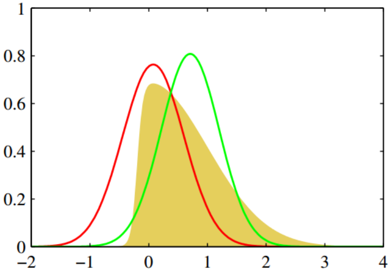
Вариационная оценка

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

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



Аппроксимация неизвестного распределения нормальным



Апроксимация Лапласа и вариационная оценка

Получение вариацонной нижней оценки

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

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

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

где

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

$D_{\mathbf{KL}}$

Максимизация вариационной нижней оценки

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

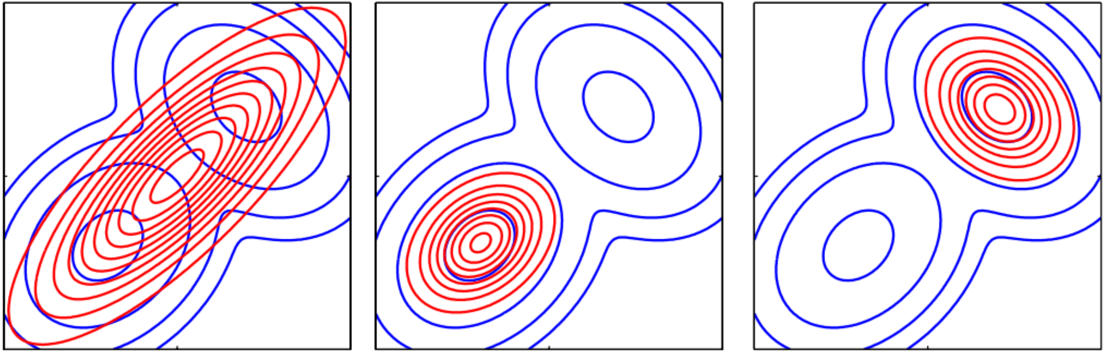
эквивалентна минимизации дивергенции между распределением распределением $q(\mathbf{w}) \in Q$ и апостериорным распределением параметров $p(\mathbf{w}|\mathbf{X}, \mathbf{f})$:

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

т.к.

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

Пример: аппроксимация мультимодального распределения



Использование вариационной нижней оценки

Для чего используют variational inference?

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

Зачем используют variational inference?

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

Variational Inference может давать сильно заниженную оценку.

Получение оценок Evidence

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

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

$$\sum_{i=1}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i) + D_{\mathrm{KL}}\big(q(\mathbf{w}) || p(\mathbf{w}|\mathbf{f})\big) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q},$$

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

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

Graves, 2011

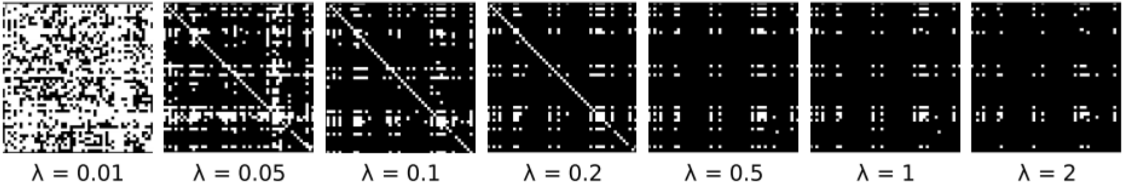
Априорное распределение: $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma\mathbf{I})$.
Вариационное распределение: $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q\mathbf{I})$.
Жадная оптимизация гиперпараметров:

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

Прунинг параметра w_i определяется относительной плотностью:

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

Выбор моделей: Graves, 2011



Louizos et. al, 2017

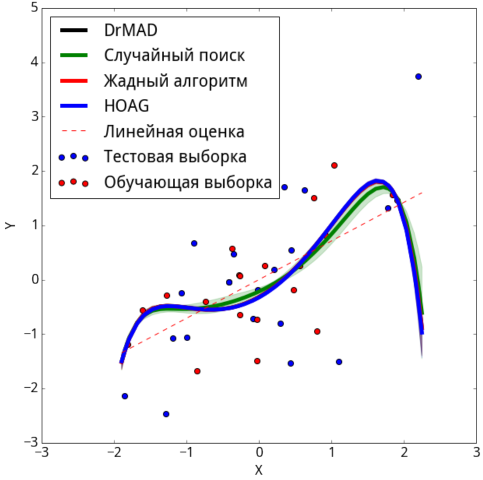
Априорное распределение задается для каждого нейрона отдельно:
$p(w_{ij}|\sigma) \sim \mathcal{N}(0, z), \quad p(z_i) \propto |z_i|^{-1}$.

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

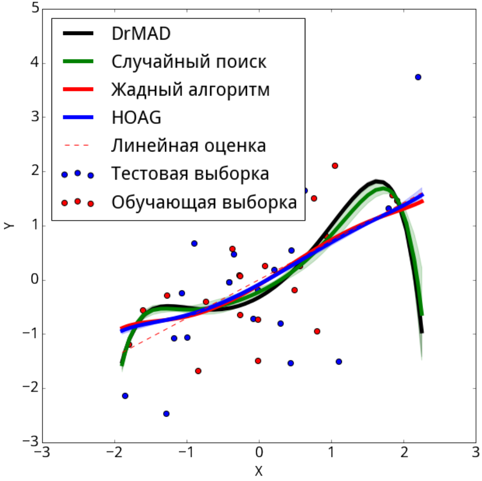
Вариационное распределение: $q(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu^z}_q, \boldsymbol{\sigma^z}_q \mathbf{I}), \quad q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q \mathbf{I})$.
Прунинг нейронов $\mathbf{w}_i$ определяется величиной

$$\frac{\boldsymbol{\sigma^z}^2_{q,i}}{\boldsymbol{\mu^z}^2_{q,i}}.$$

Кросс-Валидация vs. Evidence: отбор признаков



Кросс-валидация



Evidence

Градиентный спуск для оценки правдоподобия

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

$$\mathbf{w}' = \mathbf{w} - \alpha \nabla \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}, \mathbf{w}|\mathbf{f}) = \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}).$$

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:

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

$$\mathsf{E}_{q(\mathbf{w})}[\log p(\mathbf{X}, \mathbf{w}|\mathbf{f})] - \mathsf{S}(q(\mathbf{w})),$$

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 α разность энтропии на различных шагах оптимизации вычисляется как:

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

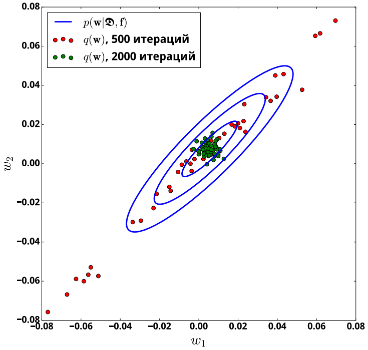
Итоговая оценка на шаге оптимизации τ:

$$\log \hat{p}(\mathbf{Y}|\mathbf{X},\mathbf{f}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathbf{w}^{g}_{\tau},\mathbf{X},\mathbf{Y}) + \mathsf{S}(q^0(\mathbf{w})) + \frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\bigl(-\alpha\mathrm{Tr}[\mathbf{H}(\mathbf{w}^{g}_{b})] - \alpha^2\mathrm{Tr}[\mathbf{H}(\mathbf{w}^{g}_{b})\mathbf{H}(\mathbf{w}^{g}_{b})]\bigr),$$

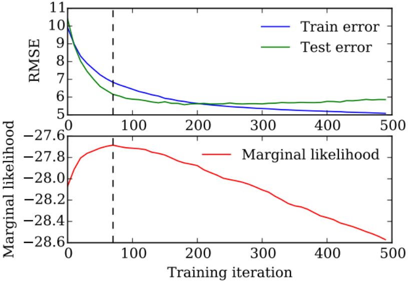
$\mathbf{w}^{g}_{b}$ — вектор параметров старта g на шаге b, $\mathsf{S}(q^0(\mathbf{w}))$ — начальная энтропия.

Переобучение

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



Схождение распределения к моде



Оценка начала переобучения, (Maclaurin et. al, 2015)

Стохастическая динамика Ланжевена

Модификация стохастического градиентного спуска:

$$\Delta \mathbf{w} = \alpha \nabla(\log p(\mathbf{w}) + \frac{m}{\hat{m}} \log p(\hat{\mathbf{X}}|\mathbf{w})) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

где $\hat{m}$ — размер подвыборки, $\hat{\mathbf{X}} \subset \mathbf{X}$ — подвыборка, шаг оптимизации α изменяется с количеством итераций:

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

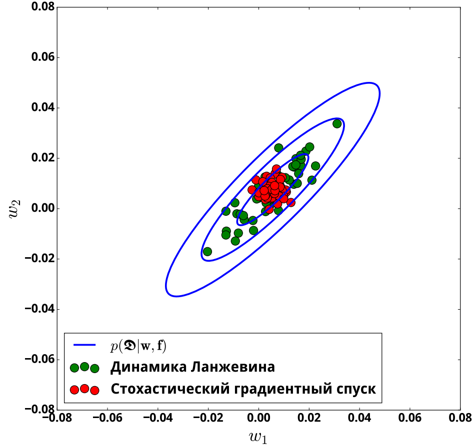
Утверждение [Welling, 2011]. Распределине $q^\tau(\mathbf{w})$ сходится к апостериорному распределению $p(\mathbf{w}|\mathbf{X}, \mathbf{f})$.

Изменение энтропии с учетом добавленного шума:

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

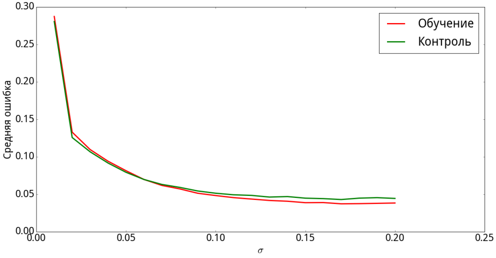
Стохастическая динамика Ланжевена

Распределения параметров после 2000 итераций:

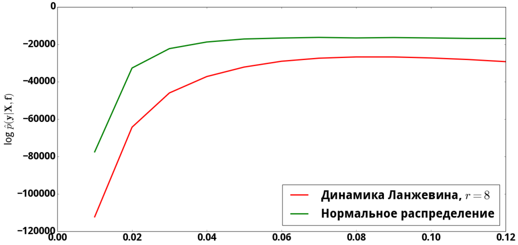


Выбор константы регуляризации

Выборка MNIST, 50 нейронов на скрытом слое.

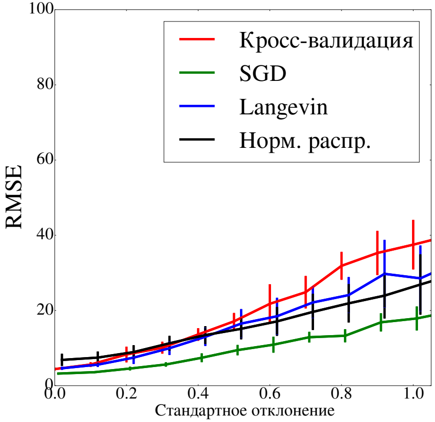


Кросс-валидация

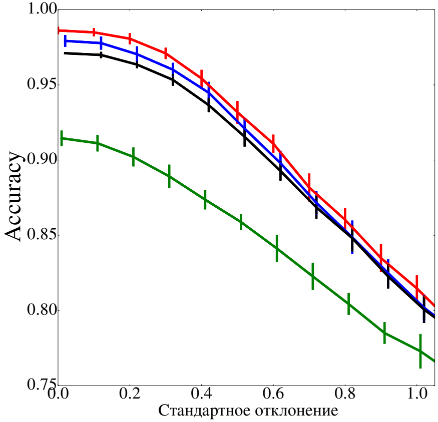


Оценка Evidence

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
2. Peter Grunwald, A tutorial introduction to the minimum description length principle
3. Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation
4. Christopher Bishop, Pattern Recognition and Machine Learning
5. Christos Louizos, Karen Ullrich, Max Welling, Bayesian Compression for Deep Learning
6. Dougal Maclaurin, David Duvenaud, Ryan P. Adams, Early Stopping is Nonparametric Variational Inference
7. Max Welling, Yee Whye Teh, Bayesian Learning via Stochastic Gradient Langevin Dynamics
8. A. Graves, Practical Variational Inference for Neural Networks