

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ  
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

На правах рукописи  
УДК 519.254

Бахтеев Олег Юрьевич

ПОСЛЕДОВАТЕЛЬНОЕ ПОРОЖДЕНИЕ МОДЕЛЕЙ  
ГЛУБОКОГО ОБУЧЕНИЯ ОПТИМАЛЬНОЙ СЛОЖНОСТИ

05.13.17 — Теоретические основы информатики

Диссертация на соискание ученой степени  
кандидата физико-математических наук

Научный руководитель:  
д.ф.-м.н. В. В. Стрижов

Москва — 2018

## Оглавление

	Стр.
Введение . . . . .	3
Глава 1. Постановка задачи	7
Постановка задачи . . . . .	8
1.1. Статистическая сложность модели . . . . .	8
1.2. Эвристические и прикладные методы . . . . .	8
1.3. Метаоптимизация . . . . .	8
1.4. Адаптивное обучение . . . . .	8
1.5. Байесовские методы порождения и выбора моделей . . . . .	8
1.6. Способы прогнозирования графовых структур . . . . .	8
Глава 2. Методы порождения и выбора модели	8
Глава 3. Оценка параметров модели	8
Глава 4. Анализ прикладных задач порождения и выбора моделей глубокого обучения	8
Заключение . . . . .	8
Список иллюстраций . . . . .	9
Список таблиц . . . . .	10
Список литературы . . . . .	11
Список использованных источников	11

## Введение

**Актуальность темы.** В работе рассматривается задача автоматического построения моделей глубокого обучения.

Под сложностью модели понимается *минимальная длина описания* [1], т.е. минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Вычисление минимальной длины описания модели является вычислительно сложной процедурой. В работе предлагается получение ее приближенной оценки, основанной на связи минимальной длины описания и *правдоподобия модели* [1]. Для получения оценки правдоподобия используются вариационные методы получения оценки правдоподобия [2], основанные на аппроксимации неизвестного другим заданным распределением. Под субоптимальной сложностью понимается вариационная оценка правдоподобия модели.

Одна из проблем построения моделей глубокого обучения — большое количество параметров моделей [3, 4]. Поэтому задача выбора моделей глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. В работе [5] приводятся теоретические оценки построения нейросетей с использованием , при которых построение модели производится итеративно последовательным увеличением числа нейронов в сети. В работе [6] предлагается жадная стратегия выбора модели нейросети с использованием релевантных априорных распределений, т.е. параметрических распределений, оптимизация параметров которых позволяет удалить часть параметров из модели. Данный метод был к задаче построения модели метода релевантных векторов [7]. Альтернативой данным алгоритмам построения моделей являются методы, основанные на прореживании сетей глубокого обучения [8, 9, 10], т.е. последовательного удаления параметров, не дающих существенного прироста качества модели. В работах [11, 12] рассматривается послойное построение модели с отдельным критерием оптимизации для каждого слоя. В работах [13, 14, 15] предлагается декомпозиция модели на порождающую и разделяющую, оптимизируемых последовательно. В работе [16] предлагается метод автоматического построения сети, основанный на бустинге. В качестве оптимизируемого функционала предлагается линейная комбинация функции правдоподобия выборки и сложности модели по Радемахеру.

В качестве порождающих моделей в сетях глубокого обучения выступают ограниченные машины Больцмана [3] и автокодировщики [17]. В работе [18] рассматриваются некоторые типы регуляризации автокодировщиков, позволяющие формально рассматривать данные модели как порождающие модели с использованием байесового вывода. В работе [19] также рассматриваются регуляризованные автокодировщики и свойства оценок их правдоподобия. В работе [20] предлагается обобщение автокодировщика с использованием вариационного байесовского вывода [2]. В работе [21] рассматриваются модификации вариационного автокодировщика и ступенчатых сетей (англ. ladder network) [22]

для случая построения многослойных порождающих моделей.

В качестве критерия выбора модели в ряде работ [23, 2, 24, 25, 26, 27] выступает правдоподобие модели. В работах [24, 25, 26, 27] рассматривается проблема выбора модели и оценки гиперпараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [1], являющаяся показателем статистической сложности модели и заданной выборки. В работе [28] рассматривается перечень критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [29] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [30] анализируется показатель избыточности параметров сети. Утверждается, что по небольшому набору параметров в глубокой сети с большим количеством избыточных параметров можно спрогнозировать значения остальных. В работе [31] рассматривается показатель робастности моделей, а также его взаимосвязь с топологией выборки и классами функций, в частности рассматривается влияние функции ошибки и ее липшицевой константы на робастность моделей. Схожие идеи были рассмотрены в работе [32], в которой исследуется устойчивость классификации модели под действием шума.

Одним из методов получения приближенного значения интеграла правдоподобия является вариационный метод получения нижней оценки интеграла [2]. В работе [33] рассматривается стохастическая версия вариационного метода. В работе [34] рассматривается алгоритм получения вариационной нижней оценки правдоподобия для оптимизации гиперпараметров моделей глубокого обучения. В работе [35] рассматривается получение вариационной нижней оценки интеграла с использованием модификации методов Монте-Карло. В работе [36] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. Схожий подход рассматривается в работе [37], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [38] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [39, 24]. Проблемой такого подхода является возможная высокая вычислительная сложность [40, 41]. В работах [42, 43] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании  $k$ -fold метода скользящего контроля, при котором выборка делится на  $k$ -частей с обучением на  $k - 1$  части и валидацией результата на оставшейся части выборки.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров [23, 2]. В работе [24] рассматривается оптимизация гиперпараметров с использованием метода скользящего контроля и методов оптимизации интеграла правдоподобия моделей, отмечается низкая скорость схо-

димости гиперпараметров при использовании метода скользящего контроля. В ряде работ [44, 45] рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе [44] предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемого функционала рассматривается ошибка на валидационной части выборки. В работе [46] предлагается метод аппроксимации градиента функции потерь по гиперпараметрам, позволяющий использовать градиентные методы в задаче оптимизации гиперпараметров на больших выборках. В работе [47] предлагается упрощенный метод оптимизации гиперпараметров с градиентным спуском: вместо всей истории обновлений параметров для оптимизации используется только последнее обновление. В работе [37] рассматривается задача оптимизации параметров градиентного спуска с использованием нижней вариационной оценки интеграла правдоподобия.

### **Цели работы.**

1. Исследовать методы построения моделей глубокого обучения.
2. Предложить критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложить метод построения модели субоптимальной сложности.
4. Разработать алгоритм построения модели и провести вычислительный эксперимент для сравнения различных подходов к решению задачи автоматического построения моделей глубокого обучения.

**Методы исследования.** Для достижения поставленных целей используются методы вариационного байесовского вывода [23, 2, 36]. Рассматриваются суперпозиции порождающей и разделяющей моделей [13, 14, 15]. Для получения оценок вариационных оценок правдоподобия модели используется метод, основанный на градиентном спуске [37, 36]. В качестве метода получения модели субоптимальной сложности используется метод Automatic Relevance Determination [23, 48] с использованием градиентных методов оптимизации гиперпараметров [44, 45, 47, 46].

### **Основные положения, выносимые на защиту.**

1. Предложен метод критерий и субоптимальной сложности модели глубокого обучения.
2. Разработан алгоритм построения модели глубокого обучения субоптимальной сложности.
3. Предложены методы оптимизации параметров и гиперпараметров модели.
4. Разработан программный комплекс для построения моделей глубокого обучения для задач классификации и регрессии.

**Научная новизна.** Разработан новый подход к построению моделей глубокого обучения. Предложены критерии субоптимальной и оптимальной сложности модели, а также исследована их связь. Предложен метод построения модели глубокого обучения субоптимальной сложности. Предложен метод оптимизации гиперпараметров модели, а также методов оптимизации модели. TODO

**Теоретическая значимость.** В данной диссертационной работе предлагаются критерии субоптимальной и оптимальной сложности, основанные на принципе минимальной длины описания. Исследуется взаимосвязь критериев оптимальной и субоптимальной сложности. Предлагаются градиентные методы для получения оценок сложности модели. Доказывается теорема об оценке энтропии эмпирического распределения параметров модели, полученных под действием оператора оптимизации.

**Практическая значимость.** Предложенные в работе методы предназначены для построения моделей глубокого обучения в задачах регрессии и классификации; оптимизации гиперпараметров полученной модели; выборе модели из конечного множества заданных моделей; получения оценок переобучения модели.

**Степень достоверности и апробация работы.** Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах иерархической классификации коллекций тезисов конференции и коллекций сайтов индустриального сектора; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. Всероссийская конференция “Интеллектуализация обработки информации” ММРО-17, 2016 [?].
2. TODO

Работа поддержана грантами Российского фонда фундаментальных исследований и Министерства образования и науки РФ.

1. 16-37-00488, Российский фонд фундаментальных исследований в рамках гранта “Разработка алгоритмов построения сетей глубокого обучения как суперпозиций универсальных моделей”.

**Публикации по теме диссертации.** Основные результаты по теме диссертации изложены в 0 печатных изданиях, 0 из которых изданы в журналах, рекомендованных ВАК.

1. TODO

**Личный вклад.** Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

**Структура и объем работы.** Диссертация состоит из оглавления, введения, четырех разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из  $X$  наименований. Основной текст занимает  $Y$  страниц.

**Краткое содержание работы по главам.** В первой главе вводятся основные понятия и определения, формулируются задачи построения моделей глубокого обучения. Рассматриваются основные критерии выбора моделей. Рассматриваются существующие алгоритмы построения моделей глубокого обучения.

Во второй главе предлагается алгоритм построения модели глубокого обучения.

В третьей главе рассматриваются методы оценки параметров модели. Предлагаются критерии прореживания параметров.

В пятой главе на базе предложенных методов описывается разработанный программный комплекс, позволяющий автоматически построить модель глубокого обучения субпотимальной сложности для заданной выборки для задачи классификации и регрессии. Работа данного комплекса анализируется на  $N$  выборках. Результаты, полученные с помощью предложенных методов, сравниваются с результатами известных алгоритмов.

## **Глава 1**

### **Постановка задачи**

## Постановка задачи

### 1.1. Статистическая сложность модели

Одна из проблем построения моделей глубокого обучения — большое количество параметров моделей [3, 4]. Поэтому задача выбора моделей глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. В работе [5] приводятся некоторые теоретические оценки построения нейросетей с использованием жадных стратегий. В работе [6] предлагается жадная стратегия выбора модели нейросети с использованием релевантных априорных распределений. Данный метод был также применен к задаче построения модели метода релевантных векторов [7]. Альтернативой данным алгоритмам построения моделей являются методы, основанные на прореживании сетей глубокого обучения [8, 9]. Популярным методом построения моделей глубокого обучения является жадное послойное построение модели с отдельным критерием оптимизации для каждого слоя [11, 12]. В ряде работ [13, 14, 15] предлагается декомпозиция модели на порождающую и разделяющую, оптимизируемых последовательно.

В качестве порождающих моделей в сетях глубокого обучения могут выступать ограниченные машины Больцмана [3] и автокодировщики [17]. В работе [18] рассматриваются некоторые типы регуляризации автокодировщиков, позволяющие формально рассматривать данные модели как порождающие модели с использованием байесового вывода. В работе [19] также рассматриваются регуляризованные автокодировщики и свойства оценок их правдоподобия. В работе [20] предлагается обобщение автокодировщика с использованием вариационного байесовского вывода [2]. В работе [21] рассматриваются модификации вариационного автокодировщика и ступенчатых сетей (англ. ladder network) [22] для случая построения многослойных порождающих моделей.

В качестве критерия выбора модели в ряде работ [23, 24, 25, 26, 27] выступает правдоподобие модели. В работах [24, 25, 26, 27] рассматривается проблема выбора модели и оценки гиперпараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [1], являющаяся показателем статистической сложности модели и заданной выборки. В работе [28] рассматривается перечень критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [29] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [30] анализируется показатель избыточности параметров сети. Утверждается, что по небольшому набору параметров в глубокой сети с большим количеством избыточных параметров можно спрогнозировать значения остальных. В работе [31] рассматривается показатель робастности моделей, а также его взаимосвязь с топологией выборки и классами функций, в частности рассматривается влияние функции ошибки и ее липшицевой константы на робастность моделей. Схожие идеи были рассмот-



рены в работе [32], в которой исследуется устойчивость классификации модели под действием шума.

Одним из методов получения приближенного значения интеграла правдоподобия является вариационный метод получения нижней оценки интеграла [2]. В работе [33] рассматривается стохастическая версия вариационного метода. В работе [34] рассматривается алгоритм получения вариационной нижней оценки правдоподобия для оптимизации гиперпараметров моделей глубокого обучения. В работе [35] рассматривается взаимосвязь градиентных методов получения вариационной нижней оценки интеграла с методом Монте-Карло. В работе [36] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. В данной работе отмечается, что стохастический градиентный спуск не оптимизирует вариационную оценку интеграла правдоподобия и приближает ее только до некоторого количества итераций оптимизации. Схожий подход рассматривается в работе [37], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [38] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [39, 24]. Проблемой такого подхода является возможная высокая вычислительная сложность [40, 41]. В работах [42, 43] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании  $k$ -fold метода скользящего контроля, при котором выборка делится на  $k$ -частей с обучением на  $k - 1$  части и валидацией результата на оставшейся части выборки.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров [23, 2]. В работе [24] рассматривается оптимизация гиперпараметров с использованием метода скользящего контроля и методов оптимизации интеграла правдоподобия моделей, отмечается низкая скорость сходимости гиперпараметров при использовании метода скользящего контроля. В ряде работ [44, 45] рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе [44] предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемого функционала рассматривается ошибка на валидационной части выборки. В работе [37] рассматривается задача оптимизации параметров градиентного спуска с использованием нижней вариационной оценки интеграла правдоподобия. В работе [36] отмечается возможность использовать градиентный метод для оптимизации гиперпараметров с использованием вариационной нижней оценки интеграла правдоподобия в качестве оптимизируемого функционала.

- 1.2. Эвристические и прикладные методы
- 1.3. Метаоптимизация
- 1.4. Адаптивное обучение
- 1.5. Байесовские методы порождения и выбора моделей
- 1.6. Способы прогнозирования графовых структур

**Глава 2**  
**Методы порождения и выбора модели**

**Глава 3**  
**Оценка параметров модели**

**Глава 4**  
**Анализ прикладных задач порождения и выбора моделей глубокого обучения**

**Заключение**

## Список иллюстраций

## Список таблиц

### СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Grünwald Peter*. A Tutorial Introduction to the Minimum Description Length Principle // *Advances in Minimum Description Length: Theory and Applications*. — MIT Press, 2005.
2. *Bishop Christopher M*. Pattern Recognition and Machine Learning (Information Science and Statistics). — Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
3. *Salakhutdinov Ruslan, Hinton Geoffrey E*. Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure // *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)* / Ed. by Marina Meila, Xiaotong Shen. — Vol. 2. — *Journal of Machine Learning Research - Proceedings Track*, 2007. — Pp. 412–419. <http://jmlr.csail.mit.edu/proceedings/papers/v2/salakhutdinov07a/salakhutdinov07a.pdf>.
4. On the importance of initialization and momentum in deep learning / Ilya Sutskever, James Martens, George E. Dahl, Geoffrey E. Hinton // *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — *JMLR Workshop and Conference Proceedings*, 2013. — Май. — Pp. 1139–1147. <http://jmlr.org/proceedings/papers/v28/sutskever13.pdf>.
5. Approximation and learning by greedy algorithms / Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, Ronald A. DeVore // *Ann. Statist.* — 2008. — 02. — Vol. 36, no. 1. — Pp. 64–94. <http://dx.doi.org/10.1214/009053607000000631>.
6. *Tzikas Dimitris, Likas Aristidis*. An Incremental Bayesian Approach for Training Multilayer Perceptrons // *Artificial Neural Networks – ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part I* / Ed. by Konstantinos Diamantaras, Wlodek Duch, Lazaros S. Iliadis. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. — Pp. 87–96. [http://dx.doi.org/10.1007/978-3-642-15819-3\\_12](http://dx.doi.org/10.1007/978-3-642-15819-3_12).
7. *Tipping Michael E*. Sparse Bayesian Learning and the Relevance Vector Machine // *J. Mach. Learn. Res.* — 2001. — Сентябрь. — Vol. 1. — Pp. 211–244. <http://dx.doi.org/10.1162/15324430152748236>.
8. *Cun Yann Le, Denker John S., Solla Sara A*. Optimal Brain Damage // *Advances in Neural Information Processing Systems*. — Morgan Kaufmann, 1990. — Pp. 598–605.
9. *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // *Информатика и ее применения*. — 2015. — Т. 9(1). — С. 79–89. <http://strijov.com/papers/Popova2014OptimalModelSelection.pdf>.
10. Learning both Weights and Connections for Efficient Neural Network / Song Han, Jeff Pool, John Tran, William Dally // *Advances in Neural Information Processing Systems 28* / Ed. by C. Cortes, N. D. Lawrence,

- D. D. Lee et al. — Curran Associates, Inc., 2015. — Pp. 1135–1143. <http://papers.nips.cc/paper/5784-learning-both-weights-and-connections-for-efficient-neural-network.pdf>.
11. Greedy Layer-Wise Training of Deep Networks / Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle // *Advances in Neural Information Processing Systems 19* / Ed. by B. Schölkopf, J. C. Platt, T. Hoffman. — MIT Press, 2007. — Pp. 153–160. <http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf>.
  12. Hinton Geoffrey E., Osindero Simon, Teh Yee-Whye. A Fast Learning Algorithm for Deep Belief Nets // *Neural Comput.* — 2006. — Июль. — Vol. 18, no. 7. — Pp. 1527–1554. <http://dx.doi.org/10.1162/neco.2006.18.7.1527>.
  13. Semi-supervised Learning with Deep Generative Models / Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling // *Advances in Neural Information Processing Systems 27* / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Curran Associates, Inc., 2014. — Pp. 3581–3589. <http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf>.
  14. Li Yi, Shapiro L. O., Bilmes J. A. A generative/discriminative learning algorithm for image classification // *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1.* — Vol. 2. — 2005. — Oct. — Pp. 1605–1612 Vol. 2.
  15. J. Lasserre. Hybrid of generative and discriminative methods for machine learning: Ph.D. thesis / University of Cambridge. — 2008.
  16. AdaNet: Adaptive Structural Learning of Artificial Neural Networks / Corinna Cortes, Xavi Gonzalvo, Vitaly Kuznetsov et al. // *CoRR.* — 2016. — Vol. abs/1607.01097. <http://arxiv.org/abs/1607.01097>.
  17. Cho Kyunghyun. Foundations and Advances in Deep Learning: G5 Artikkeliväitöskirja. — Aalto University; Aalto-yliopisto, 2014. — P. 277. <http://urn.fi/URN:ISBN:978-952-60-5575-6>.
  18. Alain Guillaume, Bengio Yoshua. What regularized auto-encoders learn from the data-generating distribution // *Journal of Machine Learning Research.* — 2014. — Vol. 15, no. 1. — Pp. 3563–3593. <http://dl.acm.org/citation.cfm?id=2750359>.
  19. Kamyshanska Hanna, Memisevic Roland. On autoencoder scoring // *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Май. — Pp. 720–728. <http://jmlr.org/proceedings/papers/v28/kamyshanska13.pdf>.
  20. D. Kingma M. Welling. Auto-Encoding Variational Bayes // *Proceedings of the International Conference on Learning Representations (ICLR).* — 2014.
  21. How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks. / Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe

- et al. // *CoRR*. — 2016. — Vol. abs/1602.02282. <http://dblp.uni-trier.de/db/journals/corr/corr1602.html#SonderbyRMSW16>.
22. Semi-Supervised Learning with Ladder Network. / Antti Rasmus, Harri Valpola, Mikko Honkala et al. // *CoRR*. — 2015. — Vol. abs/1507.02672. <http://dblp.uni-trier.de/db/journals/corr/corr1507.html#RasmusVHBR15>.
23. *MacKay David J. C.* Information Theory, Inference & Learning Algorithms. — New York, NY, USA: Cambridge University Press, 2002.
24. *Токмакова А. А., Стрижов В. В.* Оценивание гиперпараметров линейных и регрессионных моделей при отборе шумовых и коррелирующих признаков // *Информатика и её применения*. — 2012. — Т. 6(4). — С. 66–75. [http://strijov.com/papers/Tokmakova2011HyperParJournal\\_Preprint.pdf](http://strijov.com/papers/Tokmakova2011HyperParJournal_Preprint.pdf).
25. *Зайцев А. А., Стрижов В. В., Токмакова А. А.* Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия // *Информационные технологии*. — 2013. — Vol. 2. — Pp. 11–15. [http://strijov.com/papers/ZaytsevStrijovTokmakova2012Likelihood\\_Preprint.pdf](http://strijov.com/papers/ZaytsevStrijovTokmakova2012Likelihood_Preprint.pdf).
26. *Strijov V., Weber Gerhard-Wilhelm.* NONLINEAR REGRESSION MODEL GENERATION USING HYPERPARAMETERS OPTIMIZATION: Preprint 2009-21. — Middle East Technical University, 06800 Ankara, Turkey: Institute of Applied Mathematics, 2009. — Октябрь. — Preprint No. 149.
27. *Стрижов В. В.* Порождение и выбор моделей в задачах регрессии и классификации: Ph.D. thesis / Вычислительный центр РАН. — 2014. <http://strijov.com/papers/Strijov2015ModelSelectionRu.pdf>.
28. *Перекрыстенко Д.О.* Анализ структурной и статистической сложности суперпозиции нейронных сетей. — 2014. <http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Perekrestenko2014Complexity>
29. *Vladislavleva E.* Other publications TiSEM: : Tilburg University, School of Economics and Management, 2008. <http://EconPapers.repec.org/RePEc:tiu:tiutis:65a72d10-6b09-443f-8cb9-88f3bb3bc31b>.
30. Predicting Parameters in Deep Learning / Misha Denil, Babak Shakibi, Laurent Dinh et al. // *Advances in Neural Information Processing Systems* 26 / Ed. by C.j.c. Burges, L. Bottou, M. Welling et al. — 2013. — Pp. 2148–2156. [http://media.nips.cc/nipsbooks/nipspapers/paper\\_files/nips26/1053.pdf](http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1053.pdf).
31. *Xu Huan, Mannor Shie.* Robustness and generalization // *Machine Learning*. — 2012. — Vol. 86, no. 3. — Pp. 391–423. <http://dx.doi.org/10.1007/s10994-011-5268-1>.
32. Intriguing properties of neural networks. / Christian Szegedy, Wojciech Zaremba, Ilya Sutskever et al. // *CoRR*. — 2013. — Vol. abs/1312.6199. <http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SzegedyZSBEGF13>.
33. Stochastic Variational Inference / Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley // *J. Mach. Learn. Res.* — 2013. — Май. — Vol. 14, no. 1. — Pp. 1303–1347. <http://dl.acm.org/citation.cfm?id=2502581.2502622>.

34. *Graves Alex.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett et al. — Curran Associates, Inc., 2011. — Pp. 2348–2356. <http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf>.
35. *Salimans Tim, Kingma Diederik P., Welling Max.* Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. // ICML / Ed. by Francis R. Bach, David M. Blei. — Vol. 37 of *JMLR Proceedings*. — JMLR.org, 2015. — Pp. 1218–1226. <http://dblp.uni-trier.de/db/conf/icml/icml2015.html#SalimansKW15>.
36. *Maclaurin Dougal, Duvenaud David K., Adams Ryan P.* Early Stopping is Nonparametric Variational Inference // *CoRR*. — 2015. — Vol. abs/1504.01344. <http://arxiv.org/abs/1504.01344>.
37. *Mandt Stephan, Hoffman Matthew D, Blei David M.* Continuous-Time Limit of Stochastic Gradient Descent Revisited.
38. *Welling Max, Teh Yee Whye.* Bayesian Learning via Stochastic Gradient Langevin Dynamics // Proceedings of the 28th International Conference on Machine Learning (ICML-11) / Ed. by Lise Getoor, Tobias Scheffer. — ICML '11. — New York, NY, USA: ACM, 2011. — June. — Pp. 681–688.
39. *Arlot Sylvain, Celisse Alain.* A survey of cross-validation procedures for model selection // *Statist. Surv.* — 2010. — Vol. 4. — Pp. 40–79. <http://dx.doi.org/10.1214/09-SS054>.
40. Fast and Accurate Support Vector Machines on Large Scale Systems / Abhinav Vishnu, Jeyanthi Narasimhan, Lawrence Holder et al. // 2015 IEEE International Conference on Cluster Computing, CLUSTER 2015, Chicago, IL, USA, September 8-11, 2015. — 2015. — Pp. 110–119. <http://dx.doi.org/10.1109/CLUSTER.2015.26>.
41. Cross-validation pitfalls when selecting and assessing regression and classification models / Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, Simon Thomas // *Journal of Cheminformatics*. — 2014. — Vol. 6, no. 1. — Pp. 1–15. <http://dx.doi.org/10.1186/1758-2946-6-10>.
42. *Hornung Roman, Bernau Christoph, Truntzer Caroline et al.* Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation. — 2014. <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-20682-6>.
43. *Bengio Yoshua, Grandvalet Yves.* No Unbiased Estimator of the Variance of K-Fold Cross-Validation // *J. Mach. Learn. Res.* — 2004. — Декабрь. — Vol. 5. — Pp. 1089–1105. <http://dl.acm.org/citation.cfm?id=1005332.1044695>.
44. *Maclaurin Dougal, Duvenaud David, Adams Ryan.* Gradient-based Hyperparameter Optimization through Reversible Learning // Proceedings of the 32nd International Conference on Machine Learning (ICML-15) / Ed. by David Blei, Francis Bach. — JMLR Workshop and Conference Proceedings, 2015. — Pp. 2113–2122. <http://jmlr.org/proceedings/papers/v37/maclaurin15.pdf>.



45. *Domke Justin*. Generic Methods for Optimization-Based Modeling. // AISTATS / Ed. by Neil D. Lawrence, Mark A. Girolami. — Vol. 22 of *JMLR Proceedings*. — JMLR.org, 2012. — Pp. 318–326. <http://dblp.uni-trier.de/db/journals/jmlr/jmlrp22.html#Domke12>.
46. *Pedregosa Fabian*. Hyperparameter optimization with approximate gradient // Proceedings of the 33rd International Conference on Machine Learning (ICML). — 2016. <http://jmlr.org/proceedings/papers/v48/pedregosa16.html>.
47. Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters / Jelena Luketina, Tapani Raiko, Mathias Berglund, Klaus Greff // Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 / Ed. by Maria-Florina Balcan, Kilian Q. Weinberger. — Vol. 48 of *JMLR Workshop and Conference Proceedings*. — JMLR.org, 2016. — Pp. 2952–2960.
48. *Karaletsos Theofanis, Rätsch Gunnar*. Automatic Relevance Determination For Deep Generative Models // *arXiv preprint arXiv:1505.07765*. — 2015.