Пространство параметров нейронных сетей для различных априорных распределений

Аминов Тимур

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель А. А. Зайцев

Москва,
2020 г.

Пространство параметров

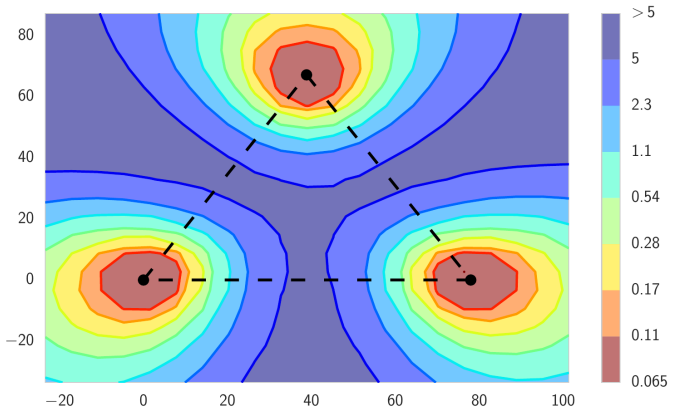


Рис.: изолированные оптимумы

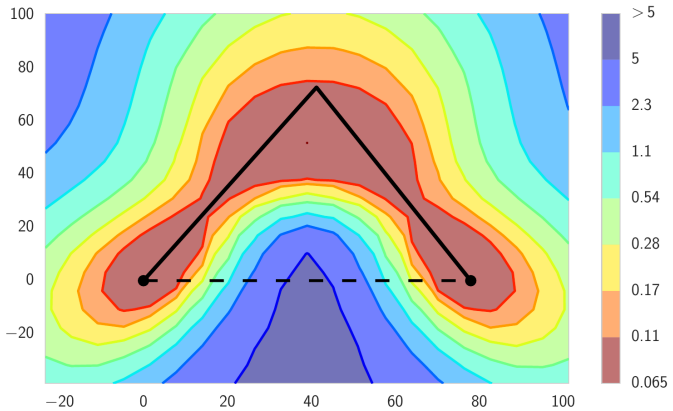


Рис.: не изолированные оптимумы

[1] https://arxiv.org/pdf/1802.10026.pdf

Цель работы

Задача

Изучить пространство параметров широких и глубоких нейронных сетей:

- для различных априорных распределений
- для различных архитектур сетей

Причины изучения

- повысить понимание нейронных сетей, изучив пространство их параметров
- использовать полученное понимание для эффективного построения ансамблей нейронных сетей
- новые методы оптимизации на основе понимания нейронных сетей

Существующие работы

Методы построения кривых предложены в следующих работах

1. T.Garipov, P.Izmailov, D.Podoprikhin, D.Vetrov, A.Gordon Wilson. *Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs.2018. (arxiv.org/abs/1802.10026)*
2. I.Anokhin, D.Yarotsky *Low-loss connection of weight vectors: distribution-based approaches (accepted to ICML 2020)*

Новизна

Таким образом, сейчас не существует решения для двух задач:

1. при каком количестве нейронов изолированные локальные минимумы лосс функции объединяются в единую структуру?
2. как на сложность пространства параметров влияет наличие априорного распределения для параметров?

В данной работе мы

1. будем оценивать сложность пространства параметров по тому, насколько соединены в этом пространстве локальные минимумы
2. построим зависимость сложности пространства параметров от числа нейронов в скрытом слое полносвязной нейронной сети
3. построим зависимость сложности от информативности априорного распределения параметров

Обозначения модели

- $\mathbf{x} \in \mathbb{R}^{d_0}$ — входные данные (признаковое описание одного объекта)
- $\hat{\mathbf{y}}_n(\mathbf{x}; \mathbf{W}_1, \mathbf{W}_2) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$ — модель прогнозирования, многослойный перцептрон (MLP)
- $\mathbf{\Theta} = \{\mathbf{W}_1, \mathbf{W}_2\}$ — множество параметров модели
- $\mathbf{\Theta}^A, \mathbf{\Theta}^B$ — два локальных оптимума функции ошибки нашей модели
- $\mathbf{W}_1, \mathbf{W}_2$ — матрицы весов, имеющие размерность $(d_1 \times d_0, d_2 \times d_1)$ соответсвенно
- σ — функция активации (ReLU)
- $\boldsymbol{\psi} : [0, 1] \to \mathbb{R}^D$, такая траектория что $\psi(0) = \mathbf{\Theta}^A, \psi(1) = \mathbf{\Theta}^B, t \in [0, 1]$

Кривая, вдоль которой функция ошибки мала

Дано

- два вектора параметров $\mathbf{\Theta}^A, \mathbf{\Theta}^B$ для нейронной сети, соответствующих локальным минимумам
- loss function $\mathrm{L}(\mathbf{\Theta})$ для них

Требуется построить

непрерывную кривую $\psi(t)$, такую что

- $\psi(0) = \mathbf{\Theta}^A, \psi(1) = \mathbf{\Theta}^B$
- $\int_0^1 L(\psi(t))dt \to \min_\psi$

Базовые методы построения кривой

Линейная аппроксимация

$$\psi(t) = (1-t)\mathbf{\Theta}^A + t\mathbf{\Theta}^B$$

Arc-аппроксимация

$$\psi(t) = \mu + \cos\left(\frac{\pi}{2}t\right)(\mathbf{\Theta}^A - \mu) + \sin\left(\frac{\pi}{2}t\right)(\mathbf{\Theta}^B - \mu),$$

где $\mu = \mathbb{E}\mathbf{\Theta}^A = \mathbb{E}\mathbf{\Theta}^B$

Улучшения

- Optimal Transportation — решение задачи с помощью задачи оптимального транспорта (OT)
- Weight Adjustment — послойное решение задачи (WA)

Цель эксперимента и данные

Цель

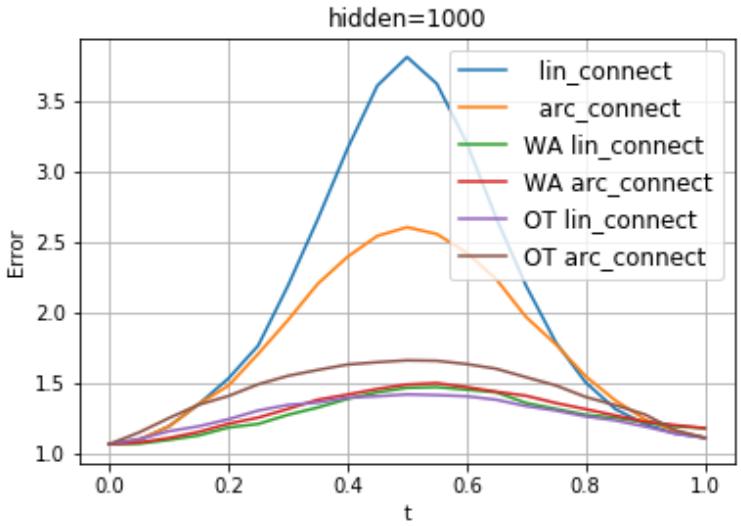
- построение кривой различными методами
- изучение зависимости пространства параметров модели (изолированность локальных оптимумов) от
  - метода построения кривой
  - сложности модели
  - априорного распредения параметров

Данные

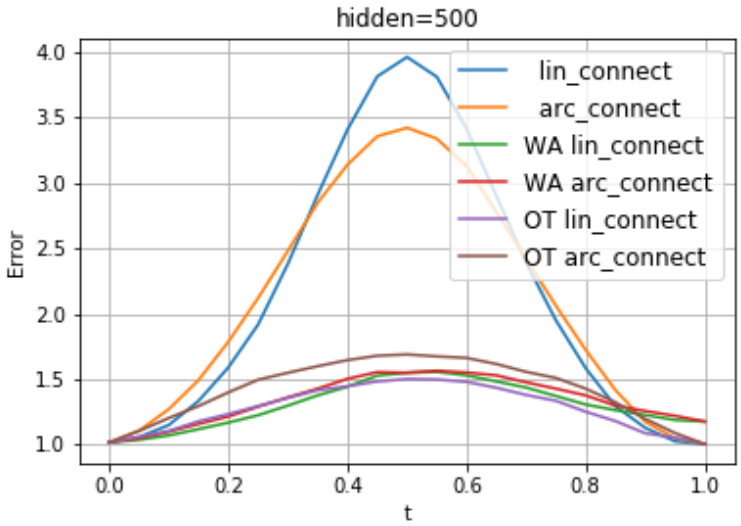
MNIST

Результаты: Сравнение методов построения кривых

Использование продвинутых методов (WA, OT) позволяет строить кривые с меньшей средней ошибкой



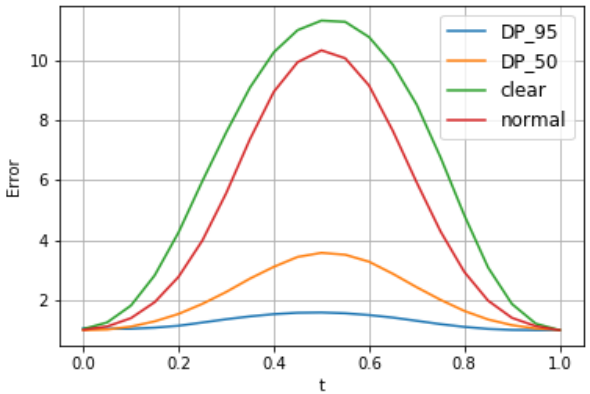
(a) Без априорного распределения



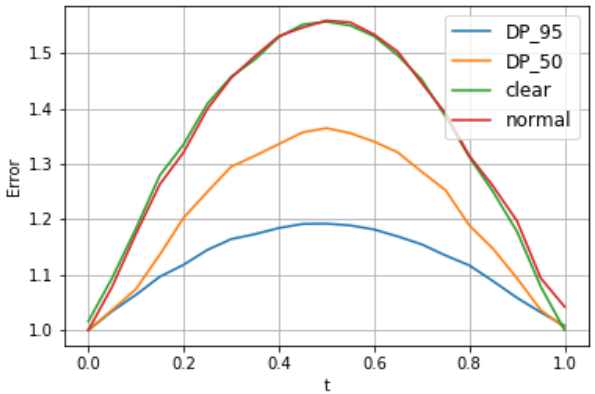
(b) Нормальное априорное распределение

Результаты: Сравнение априорных распределений

Наложение априорного распределения делаем локальные минимумы пространства параметров более связанными. Приведены нормированные ошибки.

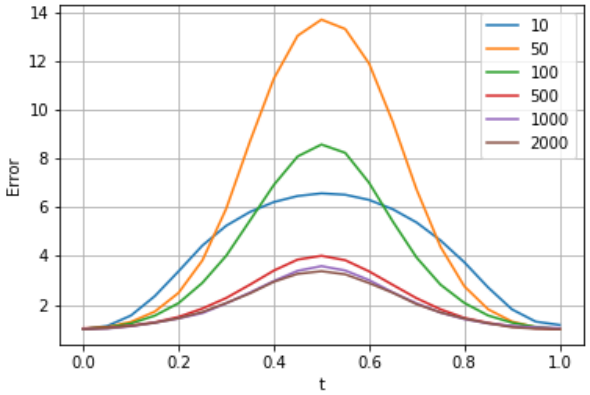


(c) Arc (hid = 50)

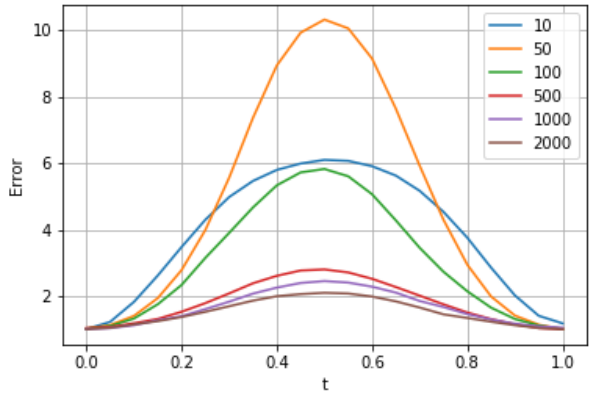


(d) OT+arc (hid = 1000).

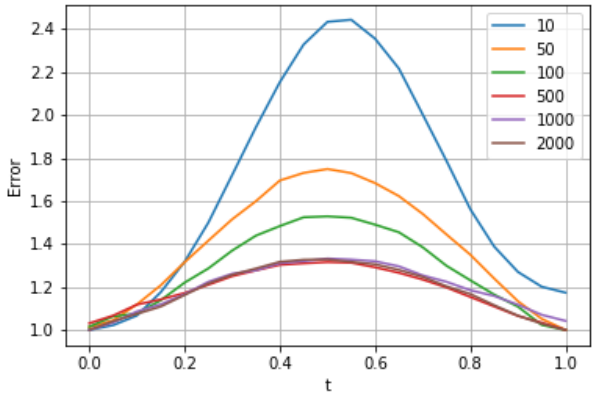
Результаты: Минимумы становятся более связанными при увеличении числа нейронов



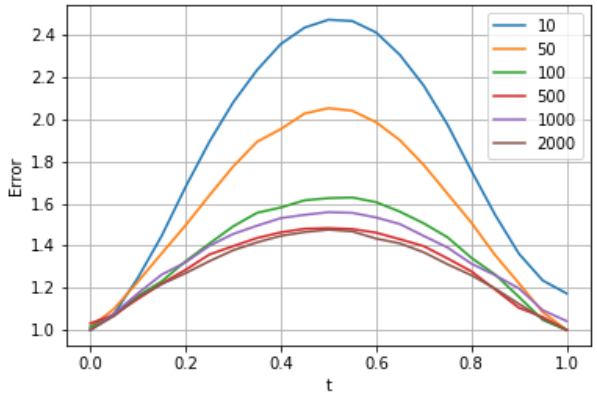
(e) Lin.



(f) Arc.



(g) OT+lin.



(h) OT+arc.

Выводы

- сложные методы построения кривой, позволяют находить кривые с меньшей средней ошибкой.
- с увеличением количества нейронов локальные минимумы функции потерь перестают быть изолированными.
- использование априорного распределения делает пространство более гладким и улучшает связанность локальных оптимумов
- более информативное априорное распределение задает более гладкое пространство параметров модели.

Дальнейшие исследования

- выявление критического числа нейронов, при котором локальные оптимумы становятся сильно изолированными
- создание новых методов эффективного построения ансамблей
- проведение аналогичных исследований для более сложных моделей
- применение других методов для построения кривой

Результаты

Max accuracy loss without prior						
model	10	50	100	500	1000	2000
Linear	45.74	24.35	10.83	4.84	3.83	4.17
Arc	35.94	21.58	9.53	4.21	3.39	2.98
Linear + Weight Adjustment	17.3	5.91	3.79	2.46	2.24	2.11
Arc + Weight Adjustment	17.11	5.77	3.71	2.43	2.23	2.11
Linear + OT	12.92	4.02	2.86	2.38	2.11	2.24
Arc + OT	13.77	4.81	3.5	2.59	2.36	2.37

Max accuracy loss with DropOut 0.5						
model	10	50	100	500	1000	2000
Linear	41.21	14.17	9.68	3.57	3.1	2.75
Arc	28.75	12.64	8.29	3.7	3.36	2.6
Linear + Weight Adjustment	20.31	7.49	4.57	2.57	2.23	2.06
Arc + Weight Adjustment	19.65	7.43	4.43	2.57	2.23	2.09
Linear + OT	15.03	4.61	3.09	2.17	2.05	2.09
Arc + OT	15.34	4.66	3.43	2.28	2.29	2.25