# Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs [1]

Никита Свербягин

НИУ Высшая Школа Экономики

17 октября 2019 г.

# План

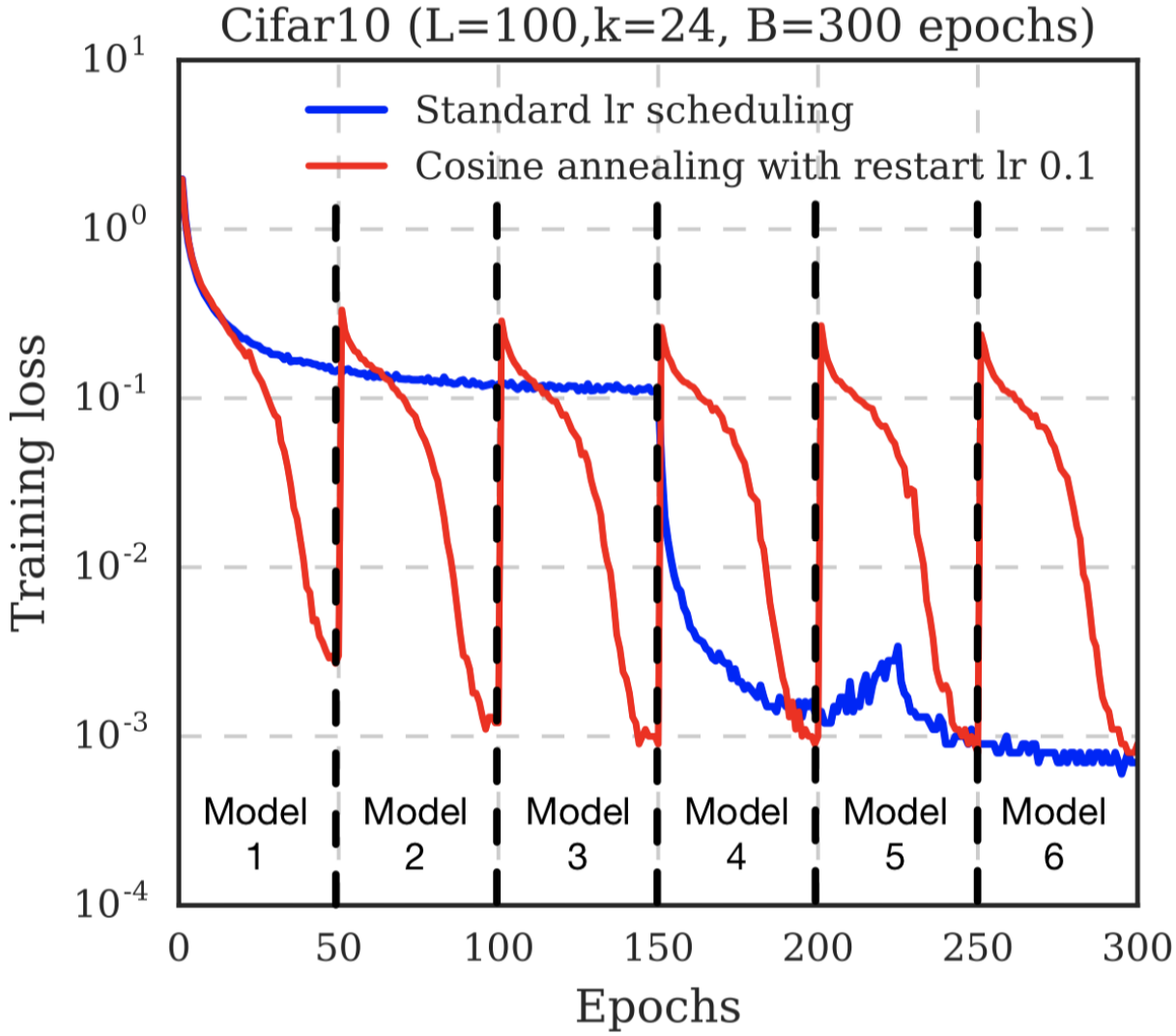
# Введение

В машинном обучении всё основано на оптимизации некой функции потерь. Проблемы функции потерь:

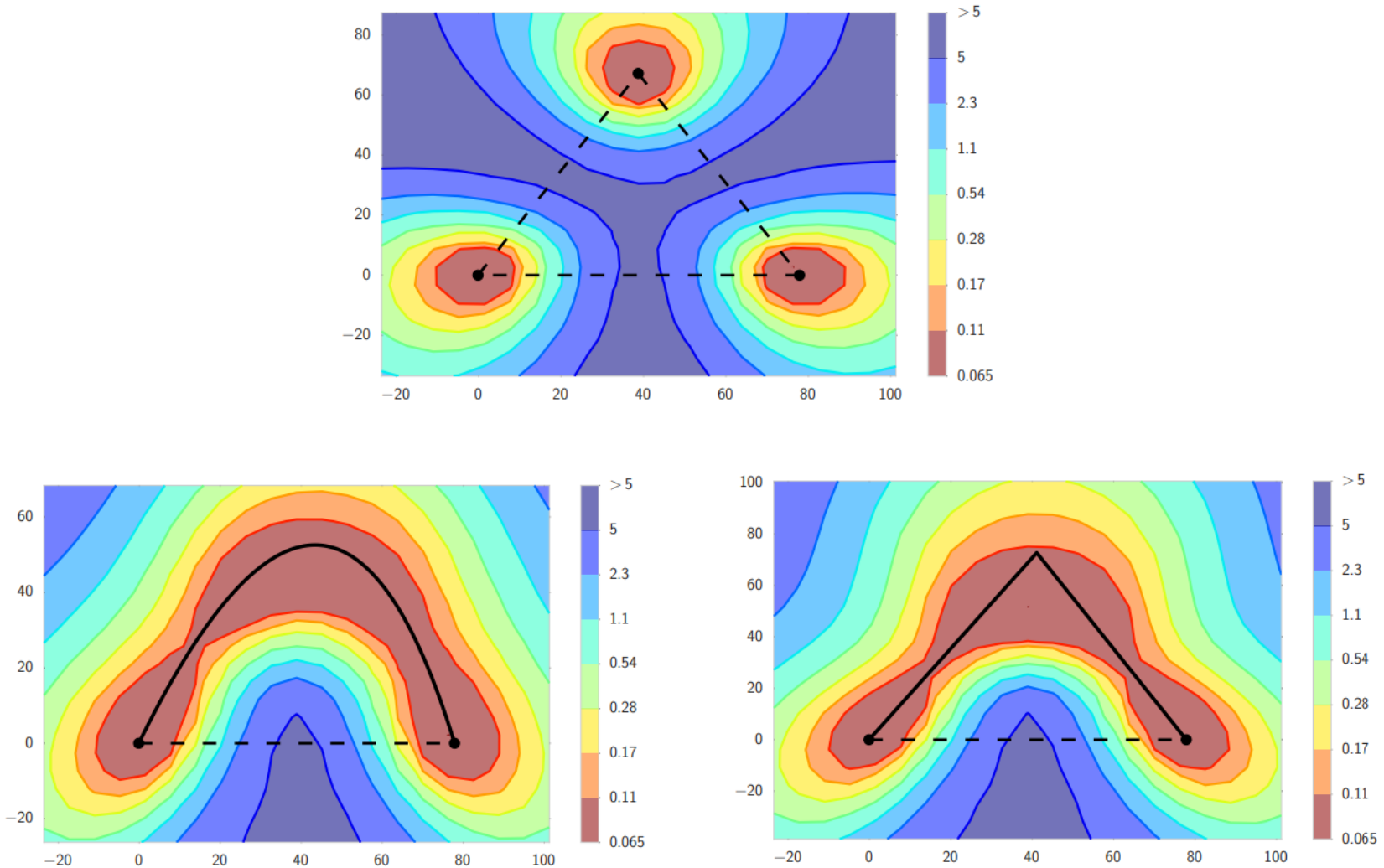
- зависит от миллионов параметров.
- Геометрические свойства практически не изучены
- Количество локальных экстремумов и седловых точек очень велико и может расти экспоненциально от количества параметров
- Вдоль отрезков, соединяющих локальные оптимумы, значения функции потерь велики.

# Общая идея

Авторы статьи предлагают технику построения ансамбля, очень похожую на технику под названием Snapshot Ensembling [2].



# Обнаруженное свойство функций потерь




функция потерь ResNet-164 на CIFAR-100 с L2-регуляризацией.

# Обнаруженное свойство функций потерь

- Существуют пути, вдоль которых сохраняются низкие значения функции потерь
- Более того, существуют очень простые пути, например кривая из двух отрезков одинаковой длины.

# Поиск кривых с низкими значениями функции потерь между двумя оптимумами

Введем нужные обозначения:

- $w_1, w_2 \in \mathbb{R}^{|net|}$ - два локальных оптимума функции потерь $\mathcal{L}(w)$
- $\phi_\theta : [0, 1] \to \mathbb{R}^{|net|}$ - кривая, соединяющая $w_1$ и $w_2$, такая, что $\phi_\theta(0) = w_1, ; \phi_\theta(1) = w_2$
- $\theta$ - параметры кривой

# Формулировка задачи поиска кривой

Будем искать такой набор параметров $\theta$, который минимизирует математическое ожидание функции потерь при равномерном распределении вдоль кривой.

$$
\begin{aligned}
l(\theta) &= \frac{\int \mathcal{L}(\phi_\theta) d\phi_\theta}{\int d\phi_\theta} \\
&= \frac{\int_0^1 \mathcal{L}(\phi_\theta(t)) \|\phi'_\theta(t)\| dt}{\int_0^1 \|\phi'_\theta(t)\| dt} \\
&= \int_0^1 \mathcal{L}(\phi_\theta(t)) q_\theta(t) dt \\
&= \mathbb{E}_{t \sim q_\theta(t)} \mathcal{L}(\phi_\theta(t))
\end{aligned}
$$

# Формулировка задачи поиска кривой

$q_\theta(t) = \|\phi'_\theta(t)\| \left( \int\limits_0^1 \|\phi'_\theta(t)\| dt \right)^{-1}$ распределение $t$, соответствующее равномерному распределению вдоль кривой.

Для некоторых кривых

$$\mathbb{E}_{t \sim q_\theta(t)} \mathcal{L}(\phi_\theta(t)) = \mathbb{E}_{t \sim \mathcal{U}(0,1)} \mathcal{L}(\phi_\theta(t))$$

Например, для кривой из двух отрезков одинаковой длины.

# Процесс подбора параметра кривой $\theta$

1. Сэмплируем $\hat{t} \sim \mathcal{U}(0, 1)$
2. Обновляем веса в соответствии с $\nabla_\theta \mathcal{L}(\phi_\theta(\hat{t}))$
3. повторяем, пока не сойдется

С помощью $\nabla_\theta \mathcal{L}(\phi_\theta(\hat{t}))$ получаем несмещенную оценку $\nabla_\theta l(\theta)$, т.к.

$$\nabla_\theta \mathcal{L}(\phi_\theta(\hat{t})) \approx \mathbb{E}_{t \sim \mathcal{U}(0,1)} \nabla_\theta \mathcal{L}(\phi_\theta(t)) = \nabla_\theta \mathbb{E}_{t \sim \mathcal{U}(0,1)} \mathcal{L}(\phi_\theta(t)) = \nabla_\theta l(\theta)$$

# Параметризация кривой

Параметризация кривой, состоящей из двух отрезков:

$$\phi_\theta(t) = \begin{cases} 2(t\theta + (0.5 - t)w_1), & 0 \leqslant t \leqslant 0.5 \\ 2((t - 0.5)w_2 + (1 - t)\theta), & 0.5 < t \leqslant 1 \end{cases}$$

# Описание экспериментов

Все эксперименты, показанные дальше, проводились с сетью ResNet-164 на датасете CIFAR-100.
В статье есть эксперименты с другими моделями, но все эксперименты дали похожий результат.

# Эксперименты с кривыми

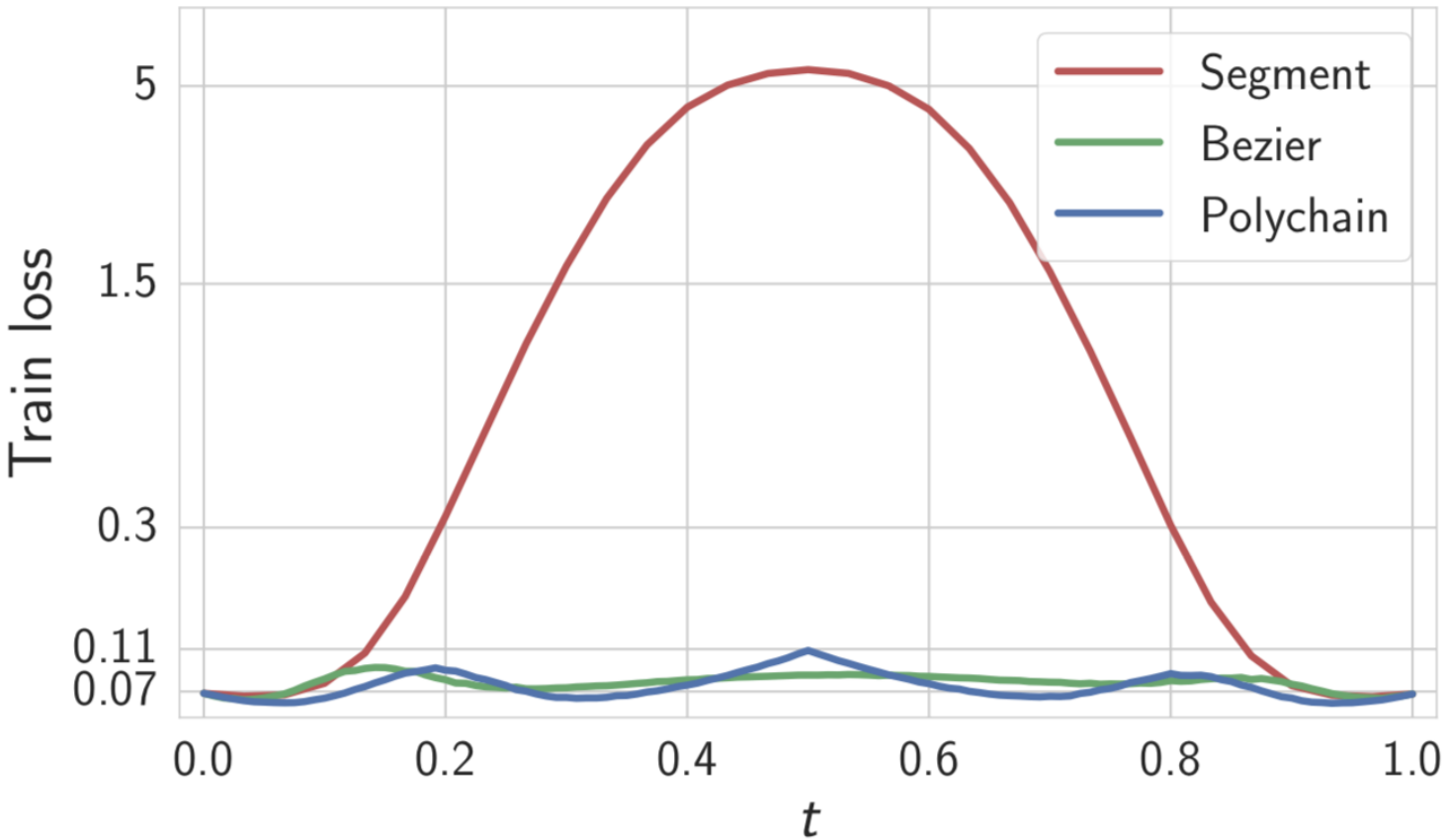



Рис.: Значения функции потерь на обучающей выборке вдоль кривых разного вида

# Эксперименты с кривыми

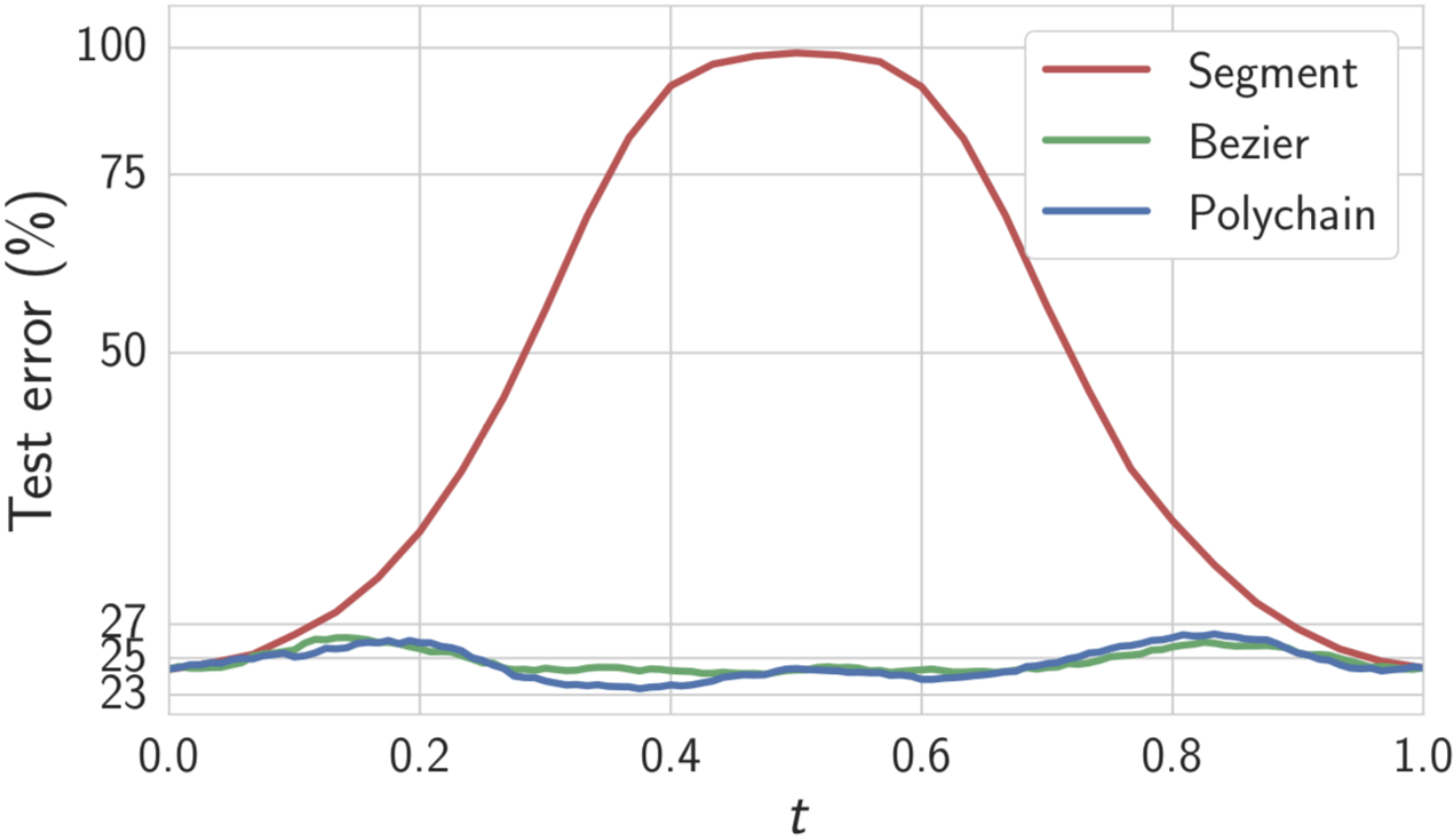



Рис.: Значения ошибки на тестовой выборке вдоль кривых разного вида

# Эксперименты с кривыми

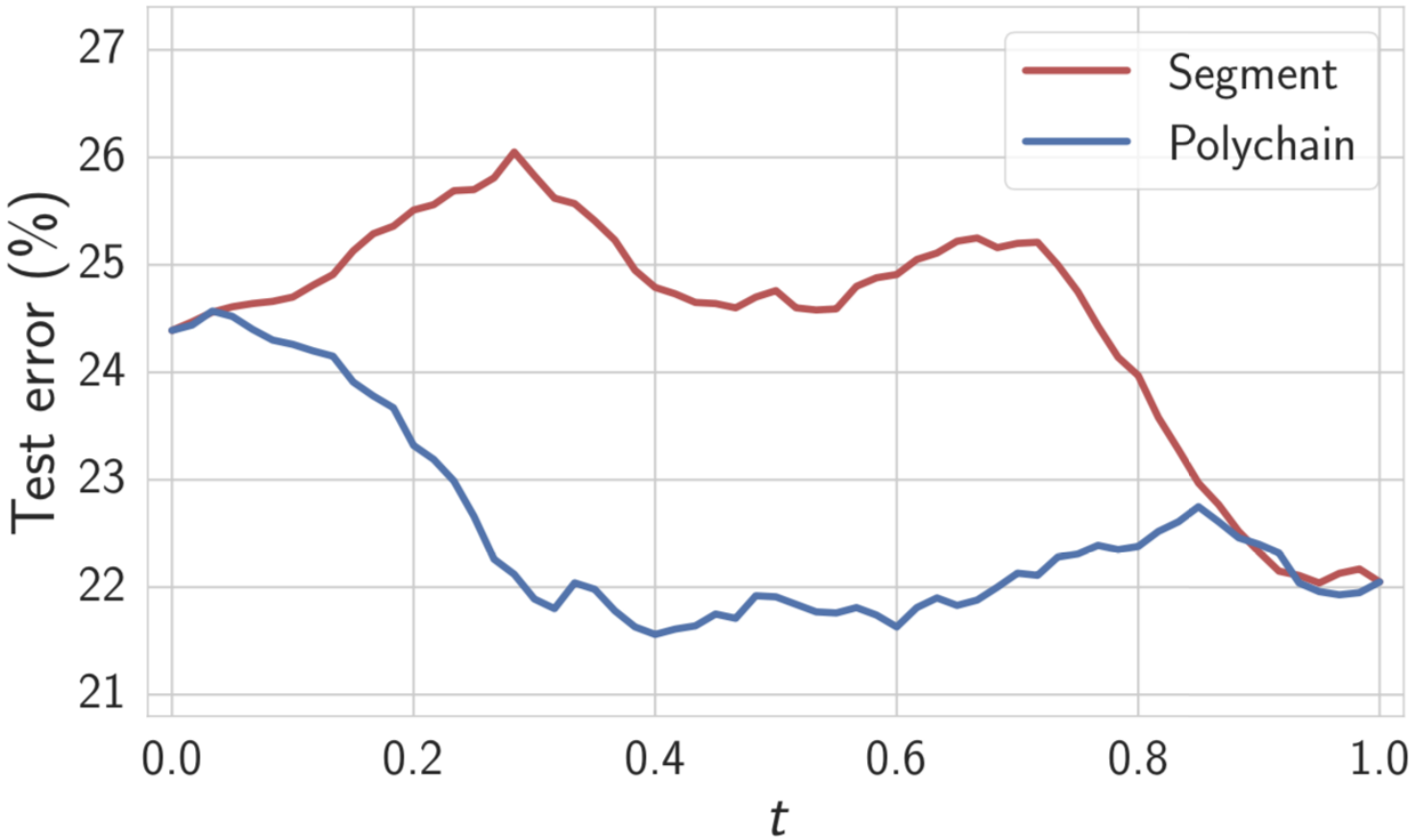



Рис.: Ошибка ансамбля из двух сетей, одна из которых - $\phi_\theta(0)$, а вторая - $\phi_\theta(t)$

# Выводы из экспериментов

1. Т.к. эксперименты проводились на разных моделях с разными сидами, и во всех случаях находились хорошие кривые, можно предположить, что они существуют в большинстве случаев.
2. По предыдущему слайду видно, что минимум функции потерь достигается при $t = 0.4$. Это означает, что значительные изменения в модели проявляются при достаточно небольшом смещении от начальной модели.
3. Можно генерировать ансамбль, идя вдоль кривой.

# Fast Geometric Ensembling

Проблема построения ансамбля по кривым в том, что нужно независимо обученных модели.

В FGE кривые не строятся явно; используется предположение, что достаточно недалеко отойти от текущего оптимума, чтобы получить новую модель с отличиями от начальной.

Идея алгоритма: заставлять сошедшуюся модель выходить из минимума, увеличивая learning rate, затем уменьшать его и при его уменьшении до минимума сохранять очередную версию модели. Так повторяется несколько раз.

# Fast Geometric Ensembling. Схема изменения learning rate

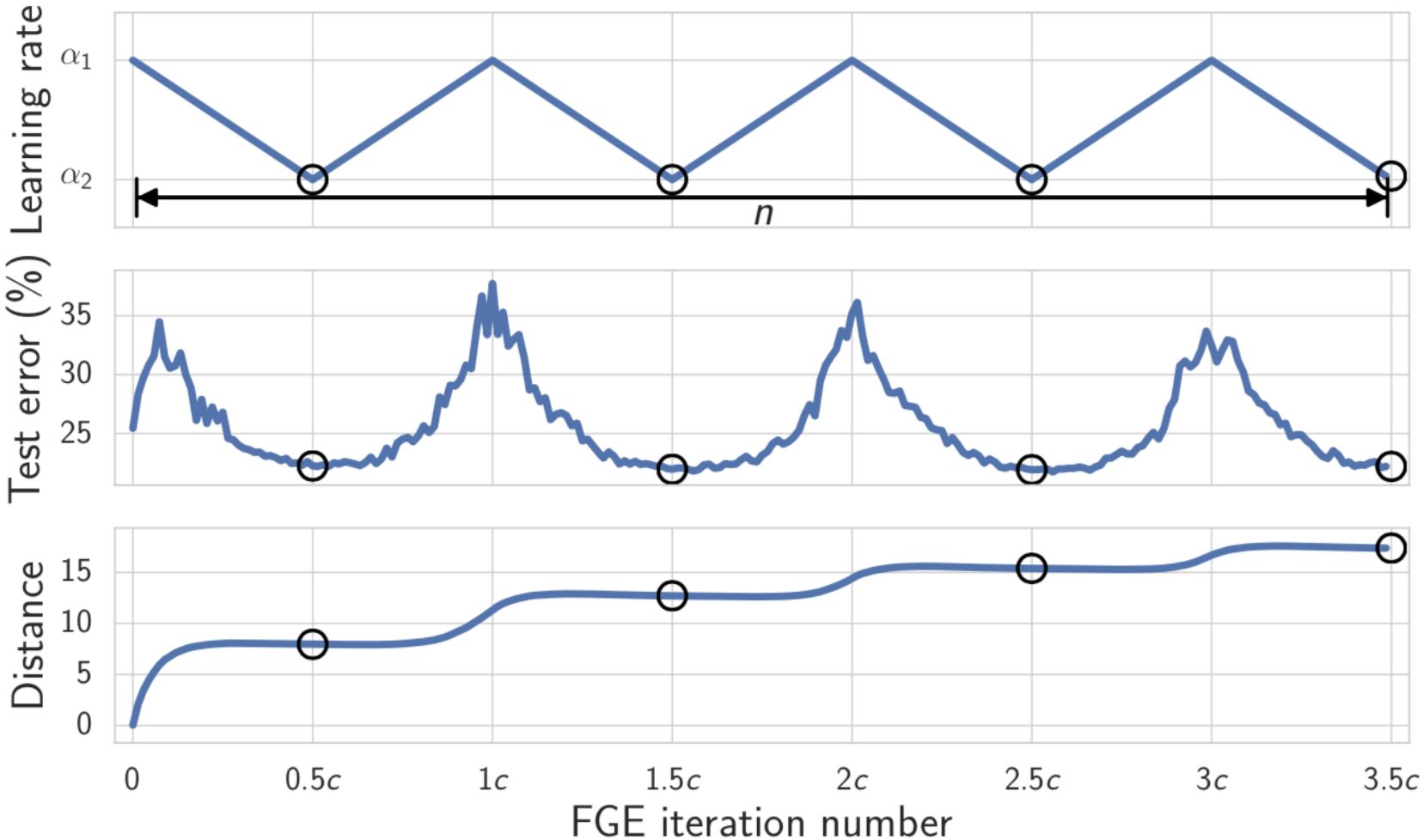



Рис.: Схема изменения LR и выбора моделей. $c$ - количество итераций (батчей) в эпохе

# Fast Geometric Ensembling. Отличие от Snapshot Ensembling

В SSE циклический LR имеет период 20-40 эпох и применяется на протяжении всего обучения. В FGE сначала находится локальный оптимум, и затем используется циклический LR с периодом 1 эпоха.

# Сравнение Fast Geometric Ensembling и Snapshot Ensembling

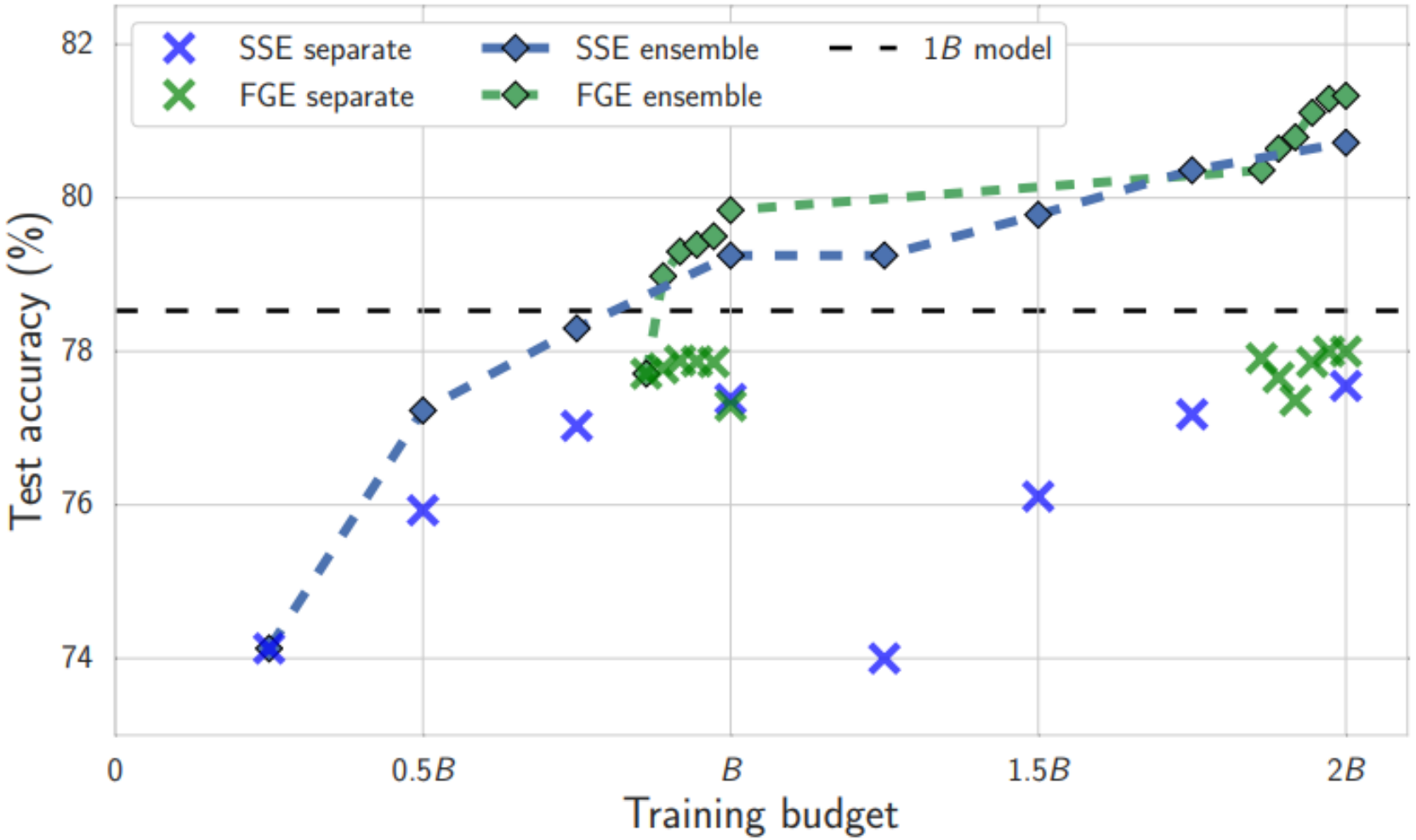



Рис.: Крестики - отдельные модели из ансамбля, ромбы - ансамбль из всех накопленных к данному моменту моделей. $B = 150$ эпох

# Сравнение разных моделей

| DNN (Budget) | method | CIFAR-100 | | | CIFAR-10 | | |
|---|---|---|---|---|---|---|---|
| | | $1B$ | $2B$ | $3B$ | $1B$ | $2B$ | $3B$ |
| VGG-16 (200) | Ind | $27.4 \pm 0.1$ | 25.28 | 24.45 | $6.75 \pm 0.16$ | 5.89 | 5.9 |
| | SSE | $26.4 \pm 0.1$ | 25.16 | 24.69 | $6.57 \pm 0.12$ | 6.19 | 5.95 |
| | FGE | $\mathbf{25.7 \pm 0.1}$ | **24.11** | **23.54** | $\mathbf{6.48 \pm 0.09}$ | **5.82** | **5.66** |
| ResNet-164 (150) | Ind | $21.5 \pm 0.4$ | 19.04 | 18.59 | $4.72 \pm 0.1$ | **4.1** | **3.77** |
| | SSE | $20.9 \pm 0.2$ | 19.28 | 18.91 | $4.66 \pm 0.02$ | 4.37 | 4.3 |
| | FGE | $\mathbf{20.2 \pm 0.1}$ | **18.67** | **18.21** | $\mathbf{4.54 \pm 0.05}$ | 4.21 | 3.98 |
| WRN-28-10 (200) | Ind | $19.2 \pm 0.2$ | 17.48 | 17.01 | $3.82 \pm 0.1$ | 3.4 | **3.31** |
| | SSE | $17.9 \pm 0.2$ | 17.3 | 16.97 | $3.73 \pm 0.04$ | 3.54 | 3.55 |
| | FGE | $\mathbf{17.7 \pm 0.2}$ | **16.95** | **16.88** | $\mathbf{3.65 \pm 0.1}$ | **3.38** | 3.52 |

Рис.: Сравнение ошибки разных моделей с различными техниками ансемблирования; при бюджете $kB$ модель независимо запускается $k$ раз

# ImageNet

Основной фокус авторов был на датасете CIFAR, но был проведен эксперимент с предобученной ResNet-50. Они запустили FGE на 5 эпох и это позволило уменьшить топ-1 ошибку на тесте на 0.56%

# Выводы

- Между локальными минимумами функции потерь существуют простые пути, вдоль которых функция потерь остается низкой, но при этом модели имеют различия в разных точках кривой.
- На этом факте основан алгоритм Fast Geometric Ensembling.
- Данный алгоритм хорош, когда вычислительные мощности ограничены - он позволяет получить ансамбль, обучая одну модель.

# Вопросы

- Алгоритм поиска кривой: формула функции потерь, что представляет собой итерация поиска кривой?
- Поиск среди кривых простого вида (два соединенных отрезка равной длины): как это позволяет упростить функцию потерь?
- Описание алгоритма Fast Geometric Ensembling.
- Как Fast Geometric Ensembling связан с поиском кривых?

# Источники


- Timur Garipov и др. *Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs*. 2018. arXiv: 1802.10026 [stat.ML].
- Gao Huang и др. “Snapshot Ensembles: Train 1, get M for free”. в: *CoRR* abs/1704.00109 (2017). arXiv: 1704.00109. URL: http://arxiv.org/abs/1704.00109.