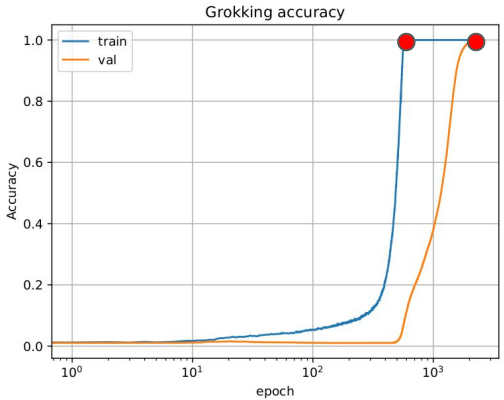
Ландшафт функции потерь в задаче гроккинга: связность оптимумов

Сергей Вячеславович Грозный

Московский государственный университет им. Ломоносова

Науч. рук: к.ф.-м.н. Д.П. Ветров
Консультант: Т. А. Южаков



★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

График точности характерный для эффекта “гроккинга” (левый рисунок)
Синтетические данные для задачи “гроккинга” (правый рисунок)

Отрезок

$$\phi(t) = tw_1 + (1-t)w_2,\ t \in [0,1]$$

Ломаная (колено)

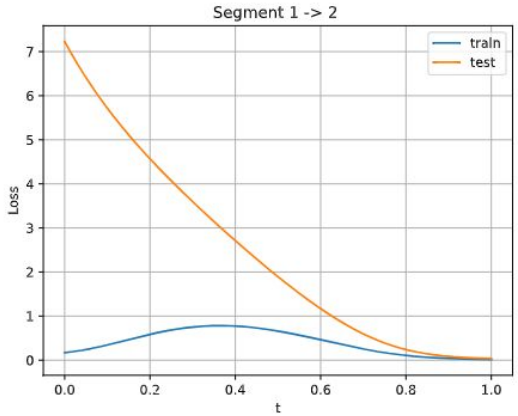
$$\phi_\theta(t) = \begin{cases} 2\left(t\theta + (0.5-t)w_1\right), & t \in [0, 0.5] \\ 2\left((t-0.5)w_2 + (1-t)\theta\right), & t \in [0.5, 1] \end{cases}$$

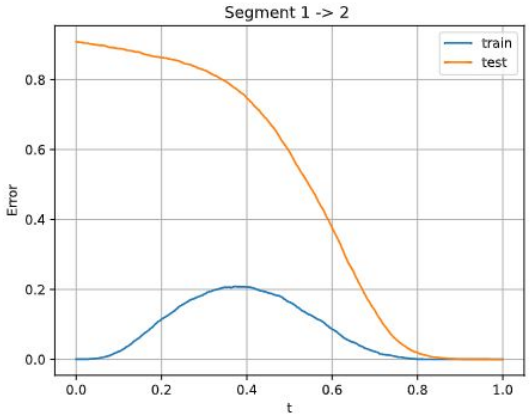
Примеры параметризации кривой для соединения минимумов

Будем оптимизировать лосс следующего вида:

$$L(\theta) = \int_0^1 Loss(\phi_\theta(t))\, dt = E_{t\sim U[0,1]} Loss(t)$$

Соединение отрезком





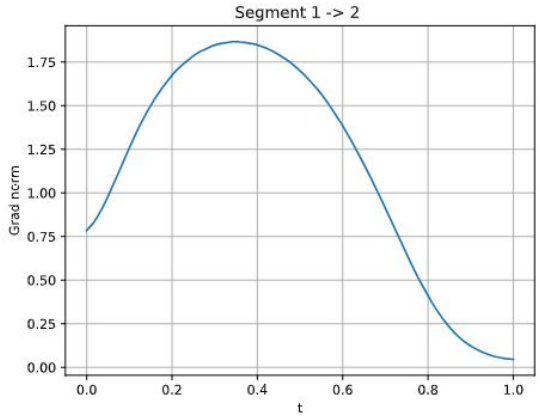


Рис. 2: Кросс-энтропия с l_2 регуляризацией (слева), ошибка (центр) и норма стох. града (справа) как функции от точек на кривой $\phi(t)$ - отрезок.

Попробуем соединить несколькими отрезками

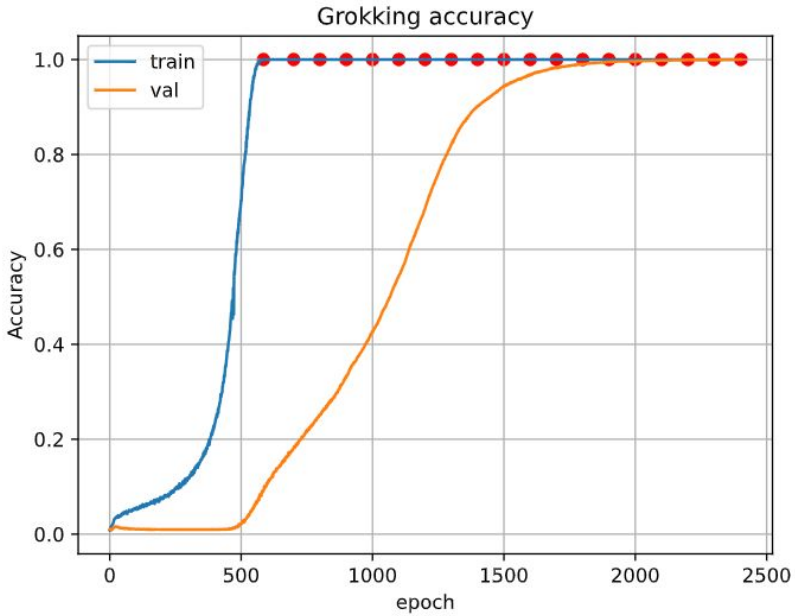
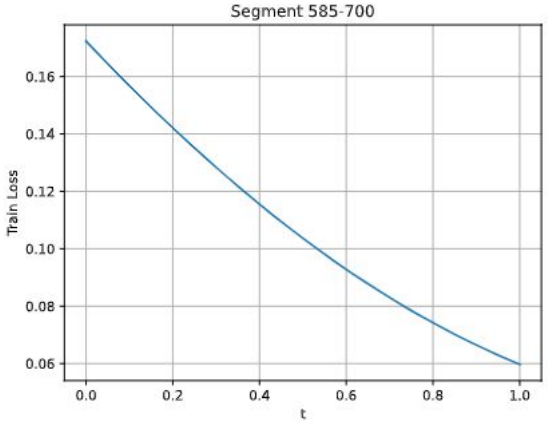
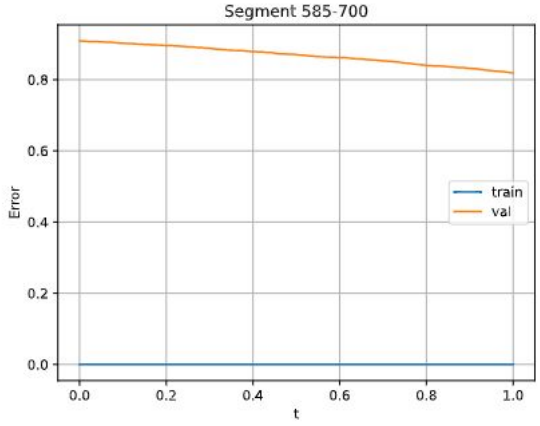
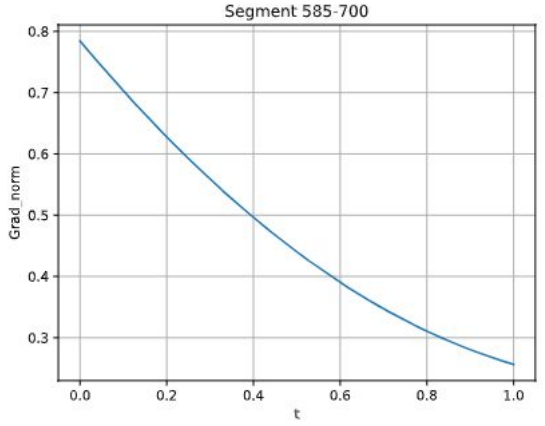


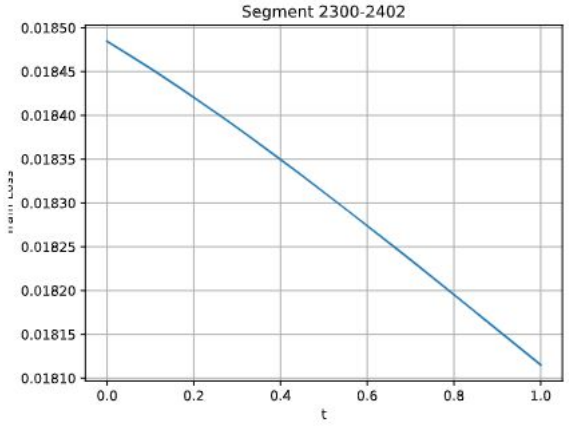
Рис. 3: Однослойный decoder-only трансформер, оптимизатор SGD с параметрами lr=0.1, weight decay=0.001, логирование весов модели каждые сто эпох (красные круги).

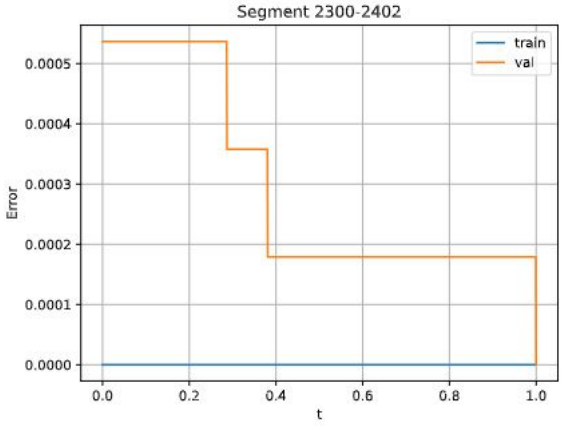
Первый и последний отрезок

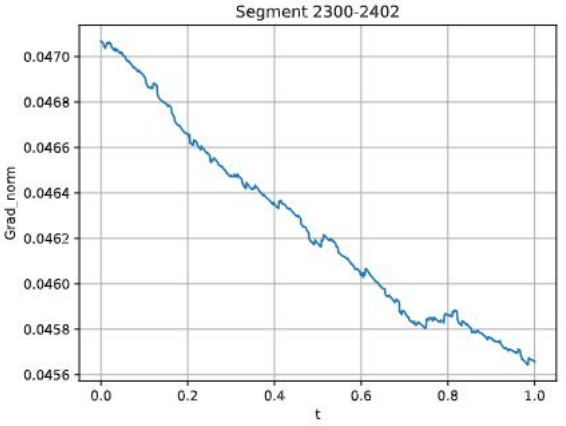




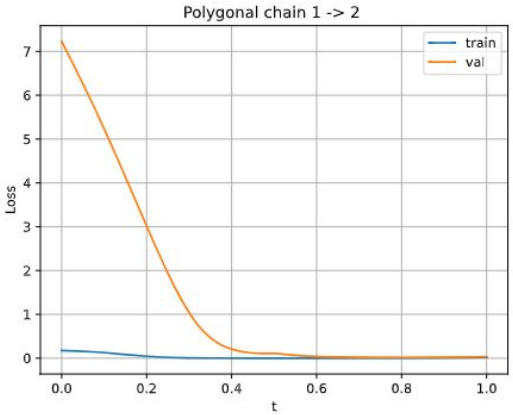


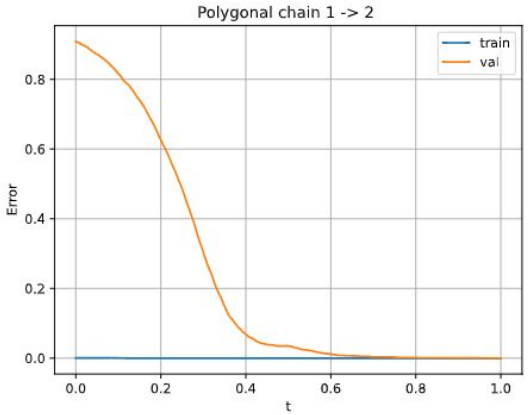


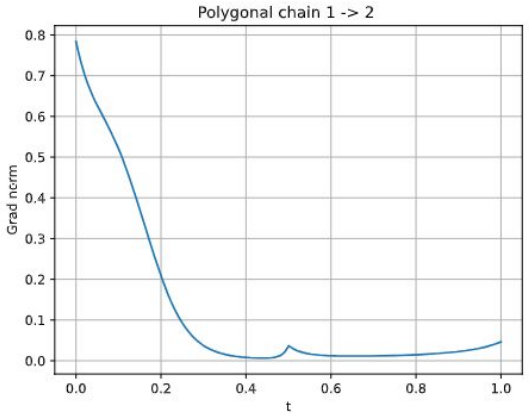




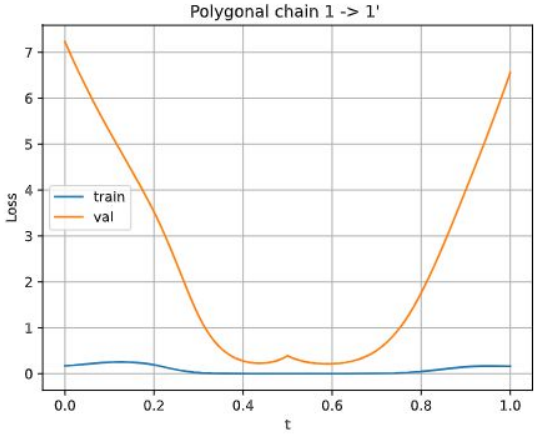
Ломаная с одним изгибом

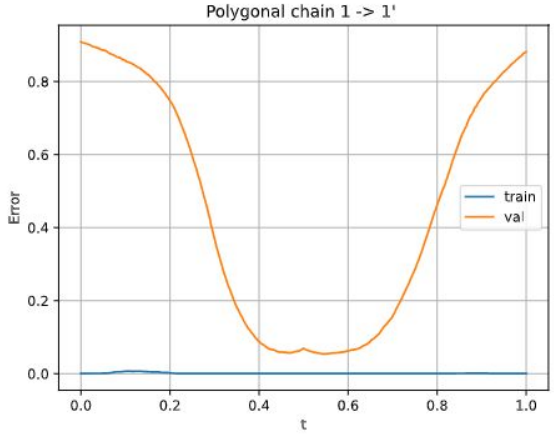


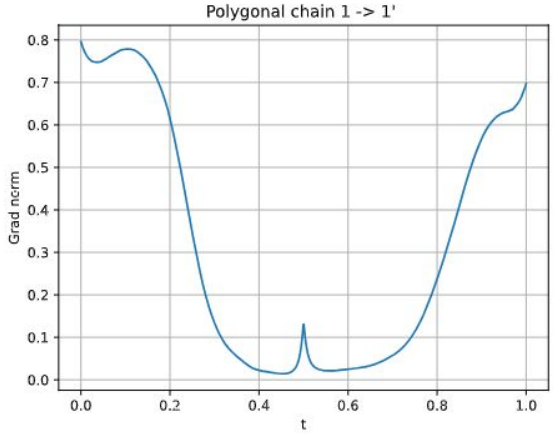




Соединяем другие точки







Выводы:

- ландшафт функции потерь имеет сложную структуру
- гроккинг возникает в силу блуждания по многообразию функции потерь с нулевой ошибкой.
- с помощью построения ломаной с одним изгибом удалось достичь довольно низкую ошибку на отложенной выборке, стартуя из точки с низкой валидацией