

Теория обучения как физика: как объяснить гроккинг

С.В. Козырев

Гроккинг (отложенное обобщение в теории обучения) объясняется организованным (овражным) ландшафтом. Переход к обобщению происходит в силу второго закона термодинамики.

Задержка при гроккинге связана с застреванием случайного блуждания в овраге.

S. V. Kozyrev, How to explain grokking, arXiv:2412.18624

Метод хищник–жертва — использование пары взаимодействующих агентов в организованном (овражном) ландшафте ускоряет гроккинг в 100 раз

I. A. Lopatin, S. V. Kozyrev, A. N. Pechen, Predator–Prey Model: Driven Hunt for Accelerated Grokking, arXiv:2509.10562

Овражные (организованные) ландшафты и метод оврагов (ускоренный градиент)

И. М. Гельфанд, М. Л. Цетлин, О некоторых способах управления сложными системами, Успехи математических наук, 17 (1), 3–25 (1962).

Оптимизация для перепараметризованных систем — многообразие минимумов ("режим интерполяции")

C. Liu, L. Zhu, and M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks, Appl. Comput. Harmon. Anal. 59, 85–116 (2022).

Гроккинг (отложенное обобщение)

A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra, Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, arXiv:2201.02177 (2022)

Гроккинг будем объяснять случайным блужданием по оврагам многообразия нулевого риска для перепараметризованной системы.

Градиентный спуск — дифференциальное уравнение

$$\frac{dx}{dt} = -\nabla f(x).$$

Разностная схема для вектора x

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Стохастический градиентный спуск

Динамика Ланжевена стохастического градиента, SGLD

$$x_{k+1} = x_k + w_k - \alpha_k \nabla f(x_k),$$

w_k — независимые гауссовские векторы с нулевым средним.

Стохастическое дифференциальное уравнение (SDE)

$$d\xi^i(t) = \sqrt{2\theta} dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i} dt, \quad (1)$$

$dw^i(t)$ стохастический дифференциал винеровского процесса

$$dw^i(t)dw^j(t) = \delta_{ij}dt.$$

Уравнение Фоккера–Планка — диффузия в потенциале

$$\frac{\partial u}{\partial t} = \theta \Delta u + \nabla u \cdot \nabla f + u \Delta f, \quad (2)$$

где $x \in \mathbb{R}^d$, $u = u(x, t)$ распределение, $f = f(x)$ потенциал, $f \in C^2(\mathbb{R}^d)$, $\theta > 0$ температура, эквивалентно

$$\frac{\partial u}{\partial t} = \theta \operatorname{div} \left[e^{-\beta f} \operatorname{grad} \left[u e^{\beta f} \right] \right].$$

Распределение Гиббса $e^{-\beta f}$, $\beta = 1/\theta$ — стационарное решение.
Сходимость к распределению Гиббса (термализация).

Свободная энергия состояния равна $F = E - \theta S$, где E энергия, а S энтропия состояния, в общем случае (для области U)

$$e^{-\beta F(U)} = \int_U e^{-\beta E(x)} dx.$$

Второй закон термодинамики — минимизация свободной энергии (при равной энергии, максимизация энтропии).

Задача обучения — минимизация эмпирического риска (суммы по выборке $\{z\}$) по гипотезе (параметру) x

$$f(\{z\}, x) = \frac{1}{L} \sum_{l=1}^L \mathcal{L}(z_l, x) \rightarrow \min_x.$$

Если $f(\{z\}, x)$ обозвать энергией, то градиентный спуск есть минимизация энергии, а стохастический градиентный спуск есть минимизация свободной энергии, то есть для SGLD в пространстве гипотез и диффузии в потенциале (2)

$$d\xi^i(t) = \sqrt{2\theta} dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i} dt,$$

$$\frac{\partial u}{\partial t} = \theta \Delta u + \nabla u \cdot \nabla f + u \Delta f,$$

$u(x)$ сходится к распределению Гиббса, сосредоточенному в ямах с низкой свободной энергией.

— Отбор широких ям, то есть улучшение обобщения решения задачи обучения в подходе широких минимумов.

Grokking (отложенное обобщение)

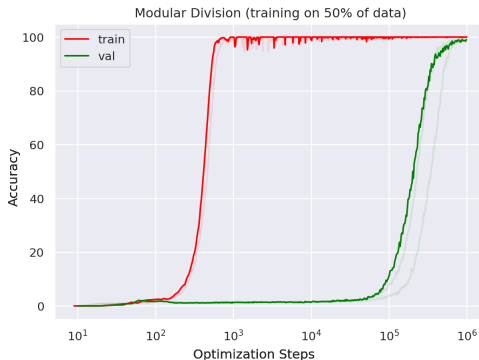


Рис.: 1. Гроккинг: Пример обобщения, возникающего после переобучения на алгоритмическом наборе данных. Красные кривые показывают точность на обучающей выборке, а зелёные на контрольной выборке. Точность на обучающей выборке становится близкой к идеальной при 10^3 шагах оптимизации (запоминание), но требуется 10^6 шагов, чтобы точность на контроле достигла этого уровня (обобщение).

Гроккинг объясняется овражным ландшафтом.

Запоминание выборки — достижение области нулевого риска $f(\{z\}, x) = 0$ путём стохастического градиентного спуска с ненулевым градиентом (сваливание в овраг, снос $\sim t$).

Гроккинг — случайное блуждание в области нулевого риска $f(\{z\}, x) = 0$ (броуновское движение $\sim \sqrt{t}$).

Переход к обобщению есть случайное блуждание из оврага в область расширения оврага (долину) и происходит в силу второго закона термодинамики.

Примечательно, что количество шагов стохастического градиента, необходимых для гроккинга, пропорционально квадрату времени запоминания (10^3 и 10^6). В процессе запоминания и гроккинга в пространстве гипотез преодолеваются близкие расстояния, но в разных режимах (снос $\sim t$, броуновское движение $\sim \sqrt{t}$).

Наблюдаемые свойства гроккинга (пороговое поведение, экспоненциальная зависимость времени гроккинга от размера выборки).

Многообразие нулевого риска — область с обобщением, область без обобщения.

Область с обобщением не зависит от выборки, область без обобщения уменьшается с ростом выборки. Предположение — линейное уменьшение энтропии с ростом выборки.

Гроккинг становится возможен, когда с ростом выборки энтропия области без обобщения становится меньше энтропии области с обобщением.

Формула Эйринга $e^{-\beta(F_1 - F_0)}$ скорости реакции (перехода между областями). $F_1 - F_0$ разность свободных энергий обобщающей области (долины) и начальной области (овраг, нет обобщения), $F = E - \theta S$. Даёт экспоненциальное падение времени гроккинга с ростом выборки.

Алгоритм хищник–жертва в теории обучения

I. A. Lopatin, S. V. Kozyrev, A. N. Pechen, Predator–Prey Model: Driven Hunt for Accelerated Grokking, arXiv:2509.10562

S. V. Kozyrev, I. A. Lopatin, A. N. Pechen, Control of Overfitting with Physics, Entropy, 26, 1090 (2024). arXiv: 2412.10716

S. V. Kozyrev, I. A. Lopatin, A. N. Pechen, Generalization in learning: Eyring formula and predator-prey model, Lobachevskii Journal of Mathematics, 46 (6), 2591–2600 (2025).

Овражный ландшафт: ускорение продвижения вдоль оврага за счёт взаимодействия между двумя агентами.

Агент Заяц и агент Волк, помещаются в овраг,
Волк гонит Зайца по оврагу.

Модель загонной охоты.

Модель хищник–жертва

Координаты x_k, x'_k в $\mathbb{R}^d$ жертвы; y_k, y'_k хищника;
 f потенциал, ∇f есть градиент потенциала.

Шаги вниз по потенциалу для жертвы и хищника

$$x'_k = x_k - \alpha \nabla f(x_k);$$

$$y'_k = y_k - \alpha \nabla f(y_k);$$

овражные шаги, $\alpha > 0$ регулирует длины шагов

$$x_{k+1} = x'_k + \alpha U(|x'_k - y'_k|) \frac{x'_k - y'_k}{|x'_k - y'_k|};$$

$$y_{k+1} = y'_k + \alpha V(|x'_k - y'_k|) \frac{x'_k - y'_k}{|x'_k - y'_k|}.$$

Взаимодействие хищник–жертва

$$U(r) = U_0 e^{-r/\sigma}, \quad V(r) = V_0 = \text{const}, \quad U_0 > V_0 > 0, \quad \sigma > 0.$$

Система уравнений модели хищник–жертва,
жертва вектор x и хищник вектор y .

$$\frac{dx}{dt} = -\nabla f(x) + \frac{x - y}{|x - y|} U(|x - y|), \quad (3)$$

$$\frac{dy}{dt} = -\nabla f(y) + \frac{x - y}{|x - y|} V(|x - y|). \quad (4)$$

Функцию $V(r) = V_0$ выбирают равной константе, функцию $U(r)$ для жертвы можно выбрать экспоненциально убывающей $U(r) = U_0 e^{-r/\sigma}$, с константами $U_0 > V_0 > 0$, $\sigma > 0$.

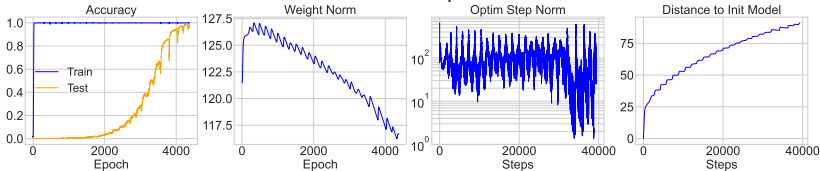
Если из первого уравнения вычесть второе, то

$$\frac{d(x-y)}{dt} = -(\nabla f(x) - \nabla f(y)) + \frac{x-y}{|x-y|} \left(U_0 e^{-|x-y|/\sigma} - V_0 \right).$$

Член с градиентами для близких x, y в овраге можно рассматривать как шум, второй член будет описывать выход пары хищник–жертва на постоянное овражное расстояние (скобка обращается в нуль).

Пара агентов на постоянном расстоянии $\sigma \ln \frac{U_0}{V_0}$ друг от друга бегут с постоянной скоростью V_0 вдоль оврага.

Baseline ModuloOperation

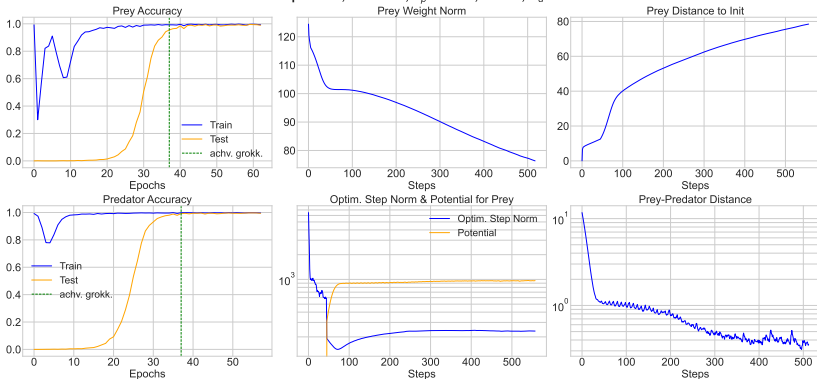


Гроккинг для модулярной арифметики, алгоритм AdamW.

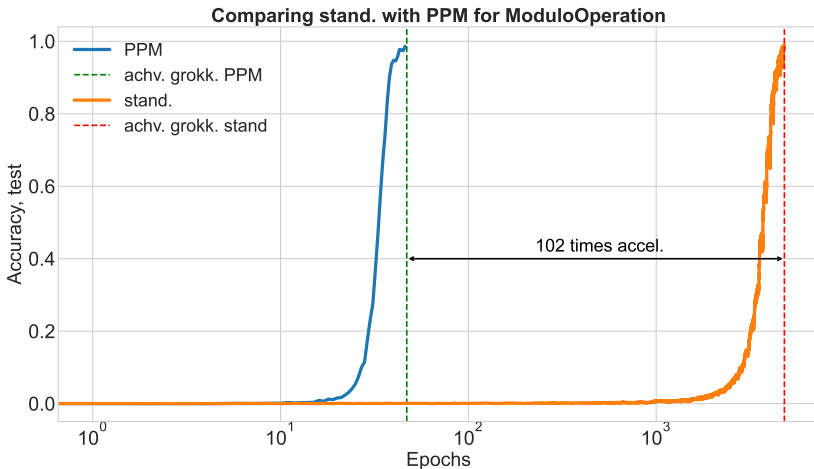
Гроккинг (переход к обобщению) требует 4000 эпох.

Хищник-жертва, гроккинг занимает около 40 эпох.

PP-conn. ModuloOperation, $A = 1000$, $\alpha_p = 1000$, $\sigma = 10$, $N_d = 5$ no use EMA



Ускорение в 100 раз!



Можно предположить, что метод позволяет достичь гроккинга за короткое время для всех моделей, где гроккинг в принципе есть (то есть ландшафт организован как овражный).