



Современные методы оптимизации для обучения нейронных сетей"

Даня Меркулов

Оптимизация для всех! ЦУ

Задача с конечной суммой

Задача с конечной суммой

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

Градиентный спуск действует следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(x) \quad (\text{GD})$$

- Стоимость итерации линейна по n .

Задача с конечной суммой

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

Градиентный спуск действует следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(x) \quad (\text{GD})$$

- Стоимость итерации линейна по n .
- Сходимость с постоянным шагом α или с линейным поиском шага.

Задача с конечной суммой

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

Градиентный спуск действует следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(x) \quad (\text{GD})$$

- Стоимость итерации линейна по n .
- Сходимость с постоянным шагом α или с линейным поиском шага.

Задача с конечной суммой

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

Градиентный спуск действует следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(x) \quad (\text{GD})$$

- Стоимость итерации линейна по n .
- Сходимость с постоянным шагом α или с линейным поиском шага.

Перейдём от полного вычисления градиента к его несмешённой оценке, когда мы на каждой итерации случайным образом равномерно выбираем i_k индекс точки:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) \quad (\text{SGD})$$

При $p(i_k = i) = \frac{1}{n}$ стохастический градиент является несмешённой оценкой градиента и задаётся так:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \sum_{i=1}^n p(i_k = i) \nabla f_i(x) = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x) = \nabla f(x)$$

Это показывает, что ожидаемое значение стохастического градиента равно фактическому градиенту $f(x)$.

Результаты для градиентного спуска

Стохастические итерации в n раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Выпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Невыпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.

Результаты для градиентного спуска

Стохастические итерации в n раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Выпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Невыпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
 - Сублинейная скорость даже в сильно выпуклом случае.

Результаты для градиентного спуска

Стохастические итерации в n раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Выпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Невыпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
 - Сублинейная скорость даже в сильно выпуклом случае.
 - Оценки скорости не могут быть улучшены при стандартных предположениях.

Результаты для градиентного спуска

Стохастические итерации в n раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Выпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Невыпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
 - Сублинейная скорость даже в сильно выпуклом случае.
 - Оценки скорости не могут быть улучшены при стандартных предположениях.
 - Оракул возвращает несмешённую аппроксимацию градиента с ограниченной дисперсией.

Результаты для градиентного спуска

Стохастические итерации в n раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Выпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Невыпуклая	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
 - Сублинейная скорость даже в сильно выпуклом случае.
 - Оценки скорости не могут быть улучшены при стандартных предположениях.
 - Оракул возвращает несмешённую аппроксимацию градиента с ограниченной дисперсией.
- Методы с моментом и квази-Ньютоновские методы не улучшают скорость в стохастическом случае, а только могут улучшить константные множители (бутилочное горлышко — дисперсия, а не число обусловленности).

SGD с постоянным шагом не сходится

Stochastic Gradient Descent. Batch = 2

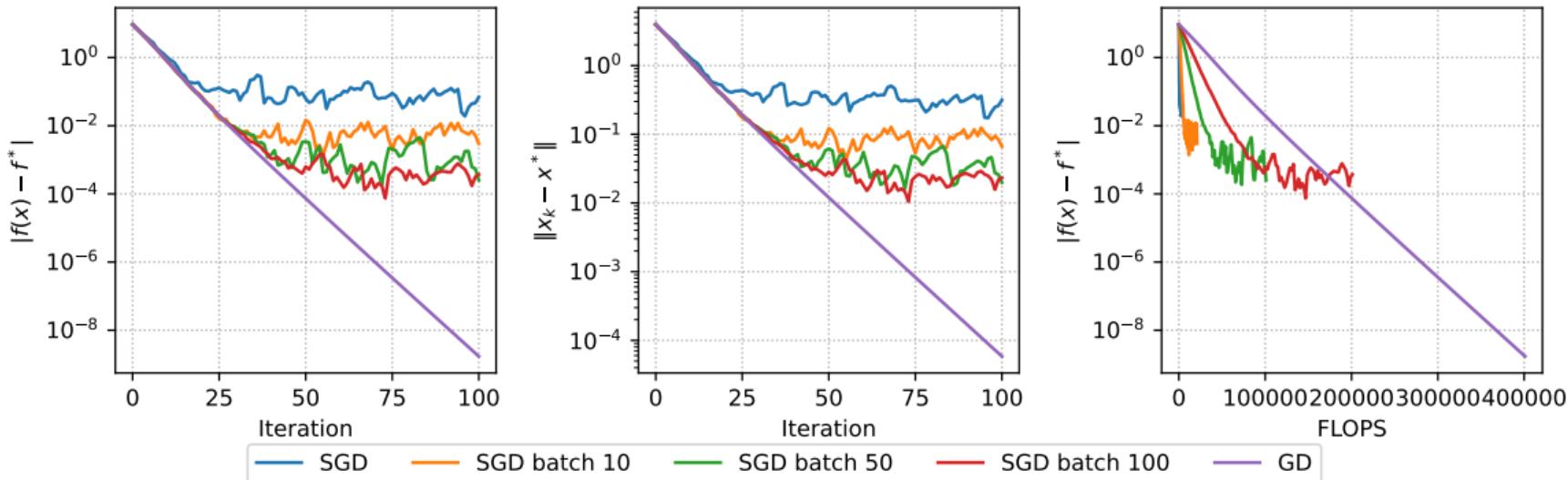


@fminxyz

Основная проблема SGD

$$f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \rightarrow \min_{x \in \mathbb{R}^n}$$

Strongly convex binary logistic regression. m=200, n=10, mu=1.



Адаптивность или масштабирование

Adagrad (Duchi, Hazan, and Singer 2010)

Очень популярный адаптивный метод. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$
$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- AdaGrad не требует настройки шага обучения: $\alpha > 0$ — фиксированная константа, и скорость обучения естественно уменьшается по итерациям.

Adagrad (Duchi, Hazan, and Singer 2010)

Очень популярный адаптивный метод. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$
$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- AdaGrad не требует настройки шага обучения: $\alpha > 0$ — фиксированная константа, и скорость обучения естественно уменьшается по итерациям.
- Шаг обучения для редких информативных признаков убывает медленно.

Adagrad (Duchi, Hazan, and Singer 2010)

Очень популярный адаптивный метод. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$
$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- AdaGrad не требует настройки шага обучения: $\alpha > 0$ — фиксированная константа, и скорость обучения естественно уменьшается по итерациям.
- Шаг обучения для редких информативных признаков убывает медленно.
- Может существенно превосходить SGD на разреженных задачах.

Adagrad (Duchi, Hazan, and Singer 2010)

Очень популярный адаптивный метод. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$
$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- AdaGrad не требует настройки шага обучения: $\alpha > 0$ — фиксированная константа, и скорость обучения естественно уменьшается по итерациям.
- Шаг обучения для редких информативных признаков убывает медленно.
- Может существенно превосходить SGD на разреженных задачах.
- Основной недостаток — монотонное накопление градиентов в знаменателе. AdaDelta, Adam, AMSGrad и др. улучшают это, популярны в обучении глубоких нейронных сетей.

Adagrad (Duchi, Hazan, and Singer 2010)

Очень популярный адаптивный метод. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$
$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- AdaGrad не требует настройки шага обучения: $\alpha > 0$ — фиксированная константа, и скорость обучения естественно уменьшается по итерациям.
- Шаг обучения для редких информативных признаков убывает медленно.
- Может существенно превосходить SGD на разреженных задачах.
- Основной недостаток — монотонное накопление градиентов в знаменателе. AdaDelta, Adam, AMSGrad и др. улучшают это, популярны в обучении глубоких нейронных сетей.
- Константа ϵ обычно устанавливается в 10^{-6} для обеспечения отсутствия деления на ноль или слишком больших шагов.

RMSProp (Tieleman and Hinton, 2012)

Улучшение AdaGrad, которое устраняет его агрессивный, монотонно убывающий шаг обучения. Использует скользящее среднее квадратов градиентов для настройки шага обучения для каждого веса. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$ и правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- RMSProp делит шаг обучения для веса на скользящее среднее величин недавних градиентов для этого веса.

RMSProp (Tieleman and Hinton, 2012)

Улучшение AdaGrad, которое устраняет его агрессивный, монотонно убывающий шаг обучения. Использует скользящее среднее квадратов градиентов для настройки шага обучения для каждого веса. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$ и правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- RMSProp делит шаг обучения для веса на скользящее среднее величин недавних градиентов для этого веса.
- Обеспечивает более тонкую настройку шагов обучения, чем AdaGrad, что делает его подходящим для нестационарных задач.

RMSProp (Tieleman and Hinton, 2012)

Улучшение AdaGrad, которое устраняет его агрессивный, монотонно убывающий шаг обучения. Использует скользящее среднее квадратов градиентов для настройки шага обучения для каждого веса. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$ и правило обновления для $j = 1, \dots, p$:

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)} + \epsilon}}$$

Заметки:

- RMSProp делит шаг обучения для веса на скользящее среднее величин недавних градиентов для этого веса.
- Обеспечивает более тонкую настройку шагов обучения, чем AdaGrad, что делает его подходящим для нестационарных задач.
- Широко используется при обучении нейронных сетей, особенно рекуррентных.

Adadelta (Zeiler, 2012)

Расширение RMSProp, нацеленное на снижение зависимости от вручную заданного глобального шага обучения. Вместо накопления всех прошлых квадратов градиентов Adadelta ограничивает окно накопленных прошлых градиентов фиксированным размером w . Механизм обновления не требует шага обучения α :

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$\tilde{g}_j^{(k)} = \frac{\sqrt{\Delta x_j^{(k-1)} + \epsilon}}{\sqrt{v_j^{(k)} + \epsilon}} g_j^{(k)}$$

$$x_j^{(k)} = x_j^{(k-1)} - \tilde{g}_j^{(k)}$$

$$\Delta x_j^{(k)} = \rho \Delta x_j^{(k-1)} + (1 - \rho)(\tilde{g}_j^{(k)})^2$$

Заметки:

- Adadelta адаптирует шаги обучения на основе скользящего окна обновлений градиентов, а не накопления всех прошлых градиентов. Таким образом, настроенные шаги обучения более устойчивы к изменениям динамики модели.

Adadelta (Zeiler, 2012)

Расширение RMSProp, нацеленное на снижение зависимости от вручную заданного глобального шага обучения. Вместо накопления всех прошлых квадратов градиентов Adadelta ограничивает окно накопленных прошлых градиентов фиксированным размером w . Механизм обновления не требует шага обучения α :

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$\tilde{g}_j^{(k)} = \frac{\sqrt{\Delta x_j^{(k-1)} + \epsilon}}{\sqrt{v_j^{(k)} + \epsilon}} g_j^{(k)}$$

$$x_j^{(k)} = x_j^{(k-1)} - \tilde{g}_j^{(k)}$$

$$\Delta x_j^{(k)} = \rho \Delta x_j^{(k-1)} + (1 - \rho)(\tilde{g}_j^{(k)})^2$$

Заметки:

- Adadelta адаптирует шаги обучения на основе скользящего окна обновлений градиентов, а не накопления всех прошлых градиентов. Таким образом, настроенные шаги обучения более устойчивы к изменениям динамики модели.
- Метод не требует начального установления шага обучения, что упрощает настройку.

Adadelta (Zeiler, 2012)

Расширение RMSProp, нацеленное на снижение зависимости от вручную заданного глобального шага обучения. Вместо накопления всех прошлых квадратов градиентов Adadelta ограничивает окно накопленных прошлых градиентов фиксированным размером w . Механизм обновления не требует шага обучения α :

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$\tilde{g}_j^{(k)} = \frac{\sqrt{\Delta x_j^{(k-1)} + \epsilon}}{\sqrt{v_j^{(k)} + \epsilon}} g_j^{(k)}$$

$$x_j^{(k)} = x_j^{(k-1)} - \tilde{g}_j^{(k)}$$

$$\Delta x_j^{(k)} = \rho \Delta x_j^{(k-1)} + (1 - \rho)(\tilde{g}_j^{(k)})^2$$

Заметки:

- Adadelta адаптирует шаги обучения на основе скользящего окна обновлений градиентов, а не накопления всех прошлых градиентов. Таким образом, настроенные шаги обучения более устойчивы к изменениям динамики модели.
- Метод не требует начального установления шага обучения, что упрощает настройку.
- Часто используется в глубоком обучении, где масштабы параметров существенно различаются между слоями.

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.
- Одна из самых цитируемых научных работ в мире.

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.
- Одна из самых цитируемых научных работ в мире.
- В 2018-2019 годах вышли статьи, указывающие на ошибку в оригинальной статье

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.
- Одна из самых цитируемых научных работ в мире.
- В 2018-2019 годах вышли статьи, указывающие на ошибку в оригинальной статье
- Не сходится для некоторых простых задач (даже выпуклых)

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.
- Одна из самых цитируемых научных работ в мире.
- В 2018-2019 годах вышли статьи, указывающие на ошибку в оригинальной статье
- Не сходится для некоторых простых задач (даже выпуклых)
- Почему-то очень хорошо работает для некоторых сложных задач

Adam (Kingma and Ba, 2014) ¹ ²

Объединяет элементы из AdaGrad и RMSProp. Учитывает экспоненциально убывающее среднее прошлых градиентов и квадратов градиентов.

EMA:

$$m_j^{(k)} = \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)}$$
$$v_j^{(k)} = \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2$$

Коррекция смещения:

$$\hat{m}_j = \frac{m_j^{(k)}}{1 - \beta_1^k}$$
$$\hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k}$$

Обновление:

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon}$$

Заметки:

- Компенсирует смещение к нулю в начальных моментах, наблюдаемое в других методах (например, RMSProp), что делает оценки более точными.
- Одна из самых цитируемых научных работ в мире.
- В 2018-2019 годах вышли статьи, указывающие на ошибку в оригинальной статье
- Не сходится для некоторых простых задач (даже выпуклых)
- Почему-то очень хорошо работает для некоторых сложных задач
- Гораздо лучше работает для языковых моделей, чем для задач компьютерного зрения - почему?

¹Adam: A Method for Stochastic Optimization

²On the Convergence of Adam and Beyond

AdamW (Loshchilov & Hutter, 2017)

Устраняет распространенную проблему с ℓ_2 -регуляризацией в адаптивных оптимизаторах, таких как Adam. Стандартная ℓ_2 -регуляризация добавляет $\lambda \|x\|^2$ к функции потерь, что приводит к градиентному слагаемому λx . В Adam это слагаемое масштабируется адаптивным шагом обучения $(\sqrt{\hat{v}_j} + \epsilon)$, связывая затухание весов (weight decay) с величинами градиента. AdamW разделяет затухание весов от шага адаптации градиентов.

Правило обновления:

$$\begin{aligned} m_j^{(k)} &= \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)} \\ v_j^{(k)} &= \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2 \\ \hat{m}_j &= \frac{m_j^{(k)}}{1 - \beta_1^k}, \quad \hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k} \\ x_j^{(k)} &= x_j^{(k-1)} - \alpha \left(\frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon} + \lambda x_j^{(k-1)} \right) \end{aligned}$$

Заметки:

- Слагаемое затухания весов $\lambda x_j^{(k-1)}$ добавляется после адаптивного шага по градиенту.

AdamW (Loshchilov & Hutter, 2017)

Устраняет распространенную проблему с ℓ_2 -регуляризацией в адаптивных оптимизаторах, таких как Adam. Стандартная ℓ_2 -регуляризация добавляет $\lambda \|x\|^2$ к функции потерь, что приводит к градиентному слагаемому λx . В Adam это слагаемое масштабируется адаптивным шагом обучения $(\sqrt{\hat{v}_j} + \epsilon)$, связывая затухание весов (weight decay) с величинами градиента. AdamW разделяет затухание весов от шага адаптации градиентов.

Правило обновления:

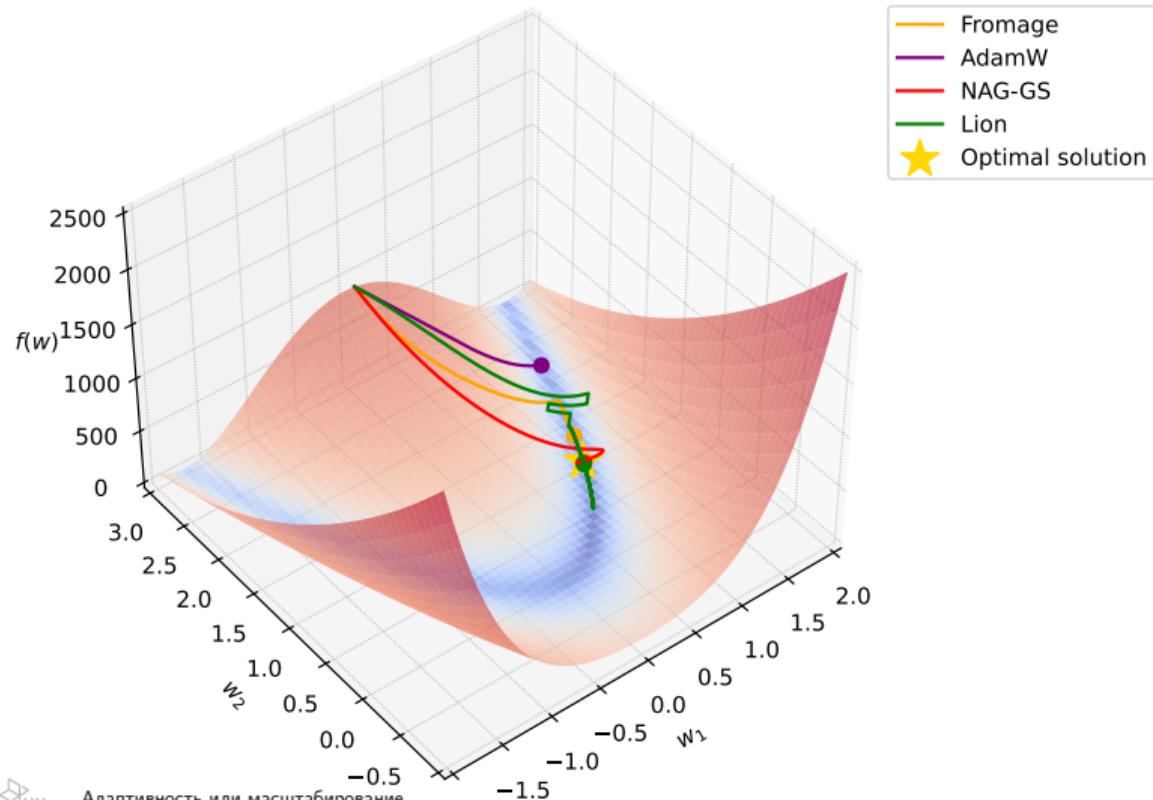
$$\begin{aligned} m_j^{(k)} &= \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)} \\ v_j^{(k)} &= \beta_2 v_j^{(k-1)} + (1 - \beta_2)(g_j^{(k)})^2 \\ \hat{m}_j &= \frac{m_j^{(k)}}{1 - \beta_1^k}, \quad \hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k} \\ x_j^{(k)} &= x_j^{(k-1)} - \alpha \left(\frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon} + \lambda x_j^{(k-1)} \right) \end{aligned}$$

Заметки:

- Слагаемое затухания весов $\lambda x_j^{(k-1)}$ добавляется после адаптивного шага по градиенту.
- Широко используется в обучении трансформаторов и других крупных моделей. Вариант по умолчанию для Hugging Face Trainer.

Много методов

Rosenbrock Function.
Adaptive stochastic gradient algorithms.
Learning rate 0.003

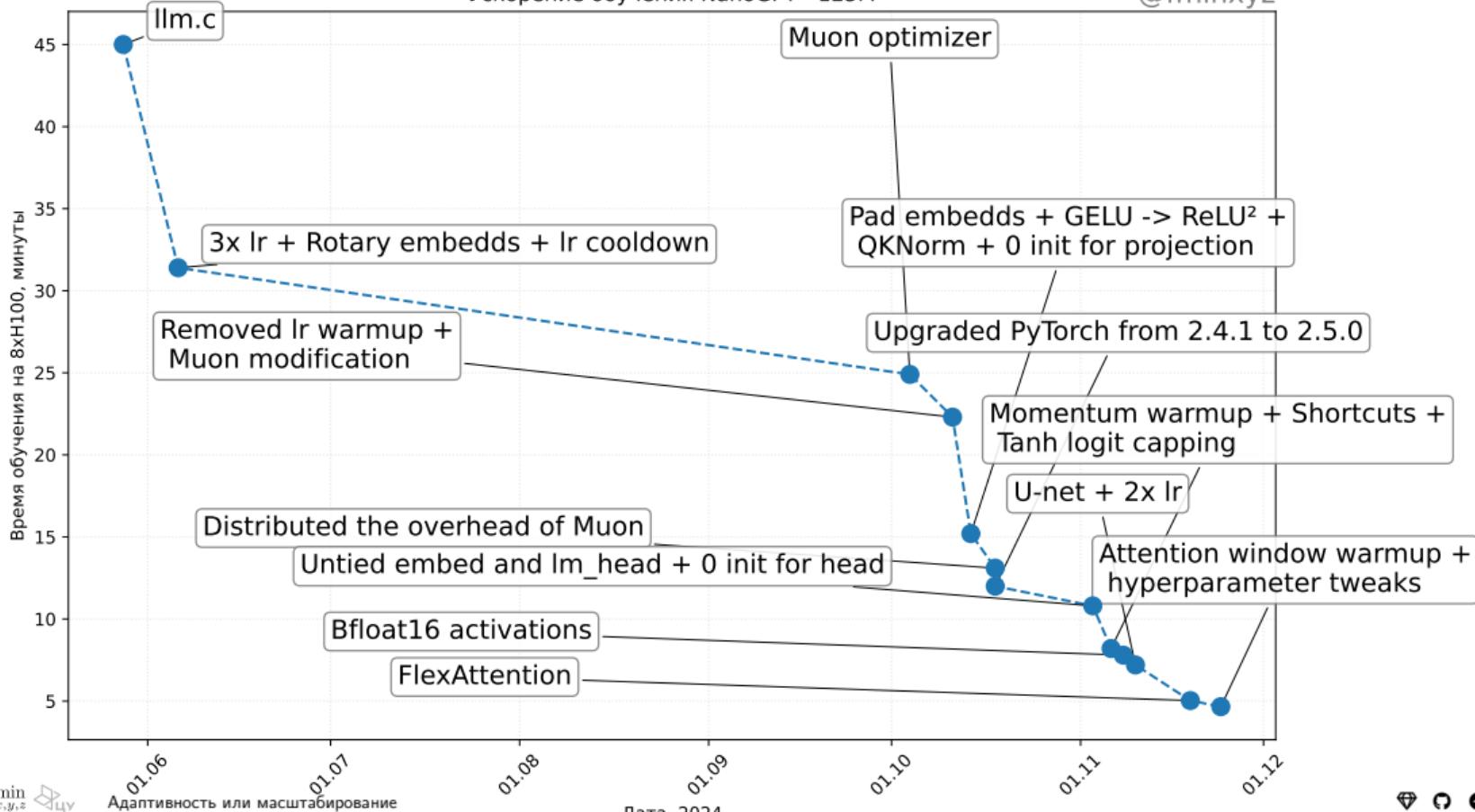


Как их сравнить? AlgoPerf benchmark

NanoGPT speedrun

Ускорение обучения NanoGPT - 125M

@fminxyz



Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Расшифровывается как **Stochastic Hessian-Approximation Matrix Preconditioning for Optimization Of deep networks**: стохастическое предобуславливание матрицей, основанной на аппроксимации гессиана, для оптимизации глубоких сетей. Это метод, вдохновлённый оптимизацией второго порядка и рассчитанный на крупномасштабное глубокое обучение.

Основная идея: аппроксимировать полноматричный предобуславливатель AdaGrad с помощью эффективных матричных структур, в частности произведений Кронекера.

Для матрицы весов $W \in \mathbb{R}^{m \times n}$, обновление включает предобуславливание с использованием приближений статистических матриц $L \approx \sum_k G_k G_k^T$ и $R \approx \sum_k G_k^T G_k$, где G_k — градиенты.

Упрощённая концепция:

1. Вычислить градиент G_k .

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Расшифровывается как **Stochastic Hessian-Approximation Matrix Preconditioning for Optimization Of deep networks**: стохастическое предобуславливание матрицей, основанной на аппроксимации гессиана, для оптимизации глубоких сетей. Это метод, вдохновлённый оптимизацией второго порядка и рассчитанный на крупномасштабное глубокое обучение.

Основная идея: аппроксимировать полноматричный предобуславливатель AdaGrad с помощью эффективных матричных структур, в частности произведений Кронекера.

Для матрицы весов $W \in \mathbb{R}^{m \times n}$, обновление включает предобуславливание с использованием приближений статистических матриц $L \approx \sum_k G_k G_k^T$ и $R \approx \sum_k G_k^T G_k$, где G_k — градиенты.

Упрощённая концепция:

1. Вычислить градиент G_k .
2. Обновить статистику $L_k = \beta L_{k-1} + (1 - \beta) G_k G_k^T$ и $R_k = \beta R_{k-1} + (1 - \beta) G_k^T G_k$.

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Расшифровывается как **Stochastic Hessian-Approximation Matrix Preconditioning for Optimization Of deep networks**: стохастическое предобуславливание матрицей, основанной на аппроксимации гессиана, для оптимизации глубоких сетей. Это метод, вдохновлённый оптимизацией второго порядка и рассчитанный на крупномасштабное глубокое обучение.

Основная идея: аппроксимировать полноматричный предобуславливатель AdaGrad с помощью эффективных матричных структур, в частности произведений Кронекера.

Для матрицы весов $W \in \mathbb{R}^{m \times n}$, обновление включает предобуславливание с использованием приближений статистических матриц $L \approx \sum_k G_k G_k^T$ и $R \approx \sum_k G_k^T G_k$, где G_k — градиенты.

Упрощённая концепция:

1. Вычислить градиент G_k .
2. Обновить статистику $L_k = \beta L_{k-1} + (1 - \beta) G_k G_k^T$ и $R_k = \beta R_{k-1} + (1 - \beta) G_k^T G_k$.
3. Вычислить предобуславливатели $P_L = L_k^{-1/4}$ и $P_R = R_k^{-1/4}$. (Обратный корень матрицы)

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Расшифровывается как **Stochastic Hessian-Approximation Matrix Preconditioning for Optimization Of deep networks**: стохастическое предобуславливание матрицей, основанной на аппроксимации гессиана, для оптимизации глубоких сетей. Это метод, вдохновлённый оптимизацией второго порядка и рассчитанный на крупномасштабное глубокое обучение.

Основная идея: аппроксимировать полноматричный предобуславливатель AdaGrad с помощью эффективных матричных структур, в частности произведений Кронекера.

Для матрицы весов $W \in \mathbb{R}^{m \times n}$, обновление включает предобуславливание с использованием приближений статистических матриц $L \approx \sum_k G_k G_k^T$ и $R \approx \sum_k G_k^T G_k$, где G_k — градиенты.

Упрощённая концепция:

1. Вычислить градиент G_k .
2. Обновить статистику $L_k = \beta L_{k-1} + (1 - \beta) G_k G_k^T$ и $R_k = \beta R_{k-1} + (1 - \beta) G_k^T G_k$.
3. Вычислить предобуславливатели $P_L = L_k^{-1/4}$ и $P_R = R_k^{-1/4}$. (Обратный корень матрицы)
4. Update: $W_{k+1} = W_k - \alpha P_L G_k P_R$.

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Заметки:

- Цель — эффективнее учитывать информацию о кривизне, чем методы первого порядка.

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Заметки:

- Цель — эффективнее учитывать информацию о кривизне, чем методы первого порядка.
- Вычислительно дороже, чем Adam, но может сходиться быстрее или приводить к лучшим решениям по числу шагов.

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Заметки:

- Цель — эффективнее учитывать информацию о кривизне, чем методы первого порядка.
- Вычислительно дороже, чем Adam, но может сходиться быстрее или приводить к лучшим решениям по числу шагов.
- Требует аккуратной реализации для эффективности (например, эффективного вычисления корней из обратной матрицы, обработки больших матриц).

Shampoo (Gupta, Anil, et al., 2018; Anil et al., 2020)

Заметки:

- Цель — эффективнее учитывать информацию о кривизне, чем методы первого порядка.
- Вычислительно дороже, чем Adam, но может сходиться быстрее или приводить к лучшим решениям по числу шагов.
- Требует аккуратной реализации для эффективности (например, эффективного вычисления корней из обратной матрицы, обработки больших матриц).
- Существуют варианты для разных форм тензоров (например, для свёрточных слоёв).

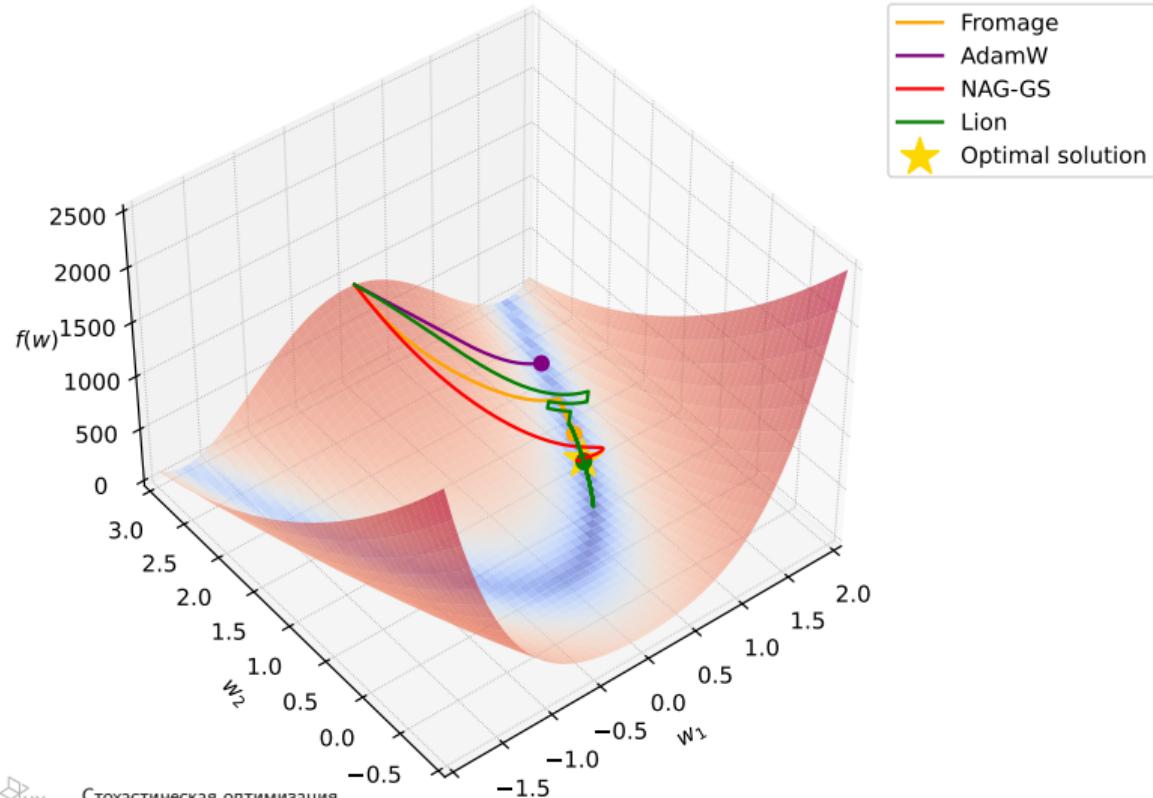
$$\begin{aligned}W_{t+1} &= W_t - \eta(G_t G_t^\top)^{-1/4} G_t (G_t^\top G_t)^{-1/4} \\&= W_t - \eta(US^2U^\top)^{-1/4}(USV^\top)(VS^2V^\top)^{-1/4} \\&= W_t - \eta(US^{-1/2}U^\top)(USV^\top)(VS^{-1/2}V^\top) \\&= W_t - \eta US^{-1/2} S S^{-1/2} V^\top \\&= W_t - \eta U V^\top\end{aligned}$$

³Deriving Muon

Стохастическая оптимизация

Много методов

Rosenbrock Function.
Adaptive stochastic gradient algorithms.
Learning rate 0.003



SGD расходится с любым шагом обучения для LLS

Оптимизация для глубокого обучения с практической точки зрения

Как их сравнить? AlgoPerf benchmark^{4 5}

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:

Как их сравнить? AlgoPerf benchmark^{4 5}

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
 - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.

Как их сравнить? AlgoPerf benchmark^{4 5}

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
 - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
 - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет $\times 3$). Оценка — медианное время выполнения по 5 наборам задач.

Как их сравнить? AlgoPerf benchmark

4 5

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
 - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
 - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет $\times 3$). Оценка — медианное время выполнения по 5 наборам задач.
- **Оценка:** результаты агрегируются с помощью профилей производительности. Профили показывают долю задач, решённых за время, не превышающее фактор τ относительно самой быстрой посылки. Итоговый балл — нормированная площадь под кривой профиля ($1.0 =$ самая быстрая на всех задачах).

Как их сравнить? AlgoPerf benchmark^{4 5}

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
 - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
 - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет $\times 3$). Оценка — медианное время выполнения по 5 наборам задач.
- **Оценка:** результаты агрегируются с помощью профилей производительности. Профили показывают долю задач, решённых за время, не превышающее фактор τ относительно самой быстрой посылки. Итоговый балл — нормированная площадь под кривой профиля ($1.0 =$ самая быстрая на всех задачах).
- **Вычислительная стоимость:** оценка требует $\sim 49,240$ суммарных часов на 8x NVIDIA V100 GPUs (в среднем ~ 3469 ч/внешняя настройка, ~ 1847 ч/самонастройка).

⁴Benchmarking Neural Network Training Algorithms

⁵Accelerating neural network training: An analysis of the AlgoPerf competition

AlgoPerf benchmark

Summary фиксированных базовых задач в AlgoPerf benchmark. Функции потерь включают кросс-энтропию (CE), среднюю абсолютную ошибку (L1) и CTC-потерю (Connectionist Temporal Classification, CTC).

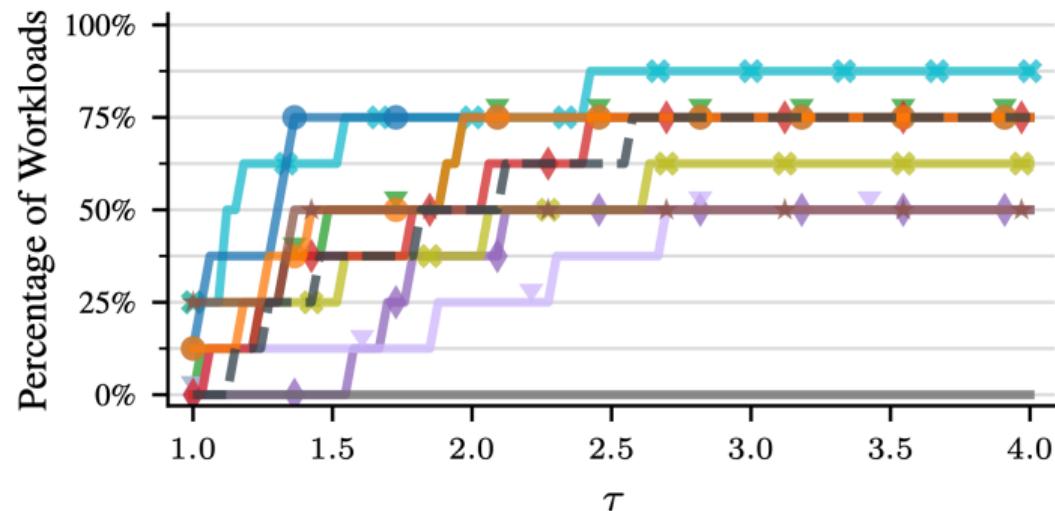
Дополнительные метрики оценки: индекс структурного сходства (SSIM), коэффициент ошибок (ER) и доля ошибок по словам (WER), средняя усреднённая точность (mAP) и метрика BLEU (bilingual evaluation understudy). Бюджет времени выполнения соответствует набору правил внешней настройки; набор правил самонастройки допускает обучение, в 3 раза более длительное.

Задача	Датасет	Модель	Потери	Метрика	Целевое значение на валидации	Бюджет времени
Clickthrough rate prediction	CRITEO 1TB	DLRMSMALL	CE	CE	0.123735	7703
MRI reconstruction	FASTMRI	U-NET	L1	SSIM	0.7344	8859
Image classification	IMAGENET	ResNet-50	CE	ER	0.22569	63,008
		ViT	CE	ER	0.22691	77,520
Speech recognition	LIBRISPEECH	Conformer	CTC	WER	0.085884	61,068
		DeepSpeech	CTC	WER	0.119936	55,506
Molecular property prediction	OGBG	GNN	CE	mAP	0.28098	18,477
Translation	WMT	Transformer	CE	BLEU	30.8491	48,151

AlgoPerf benchmark

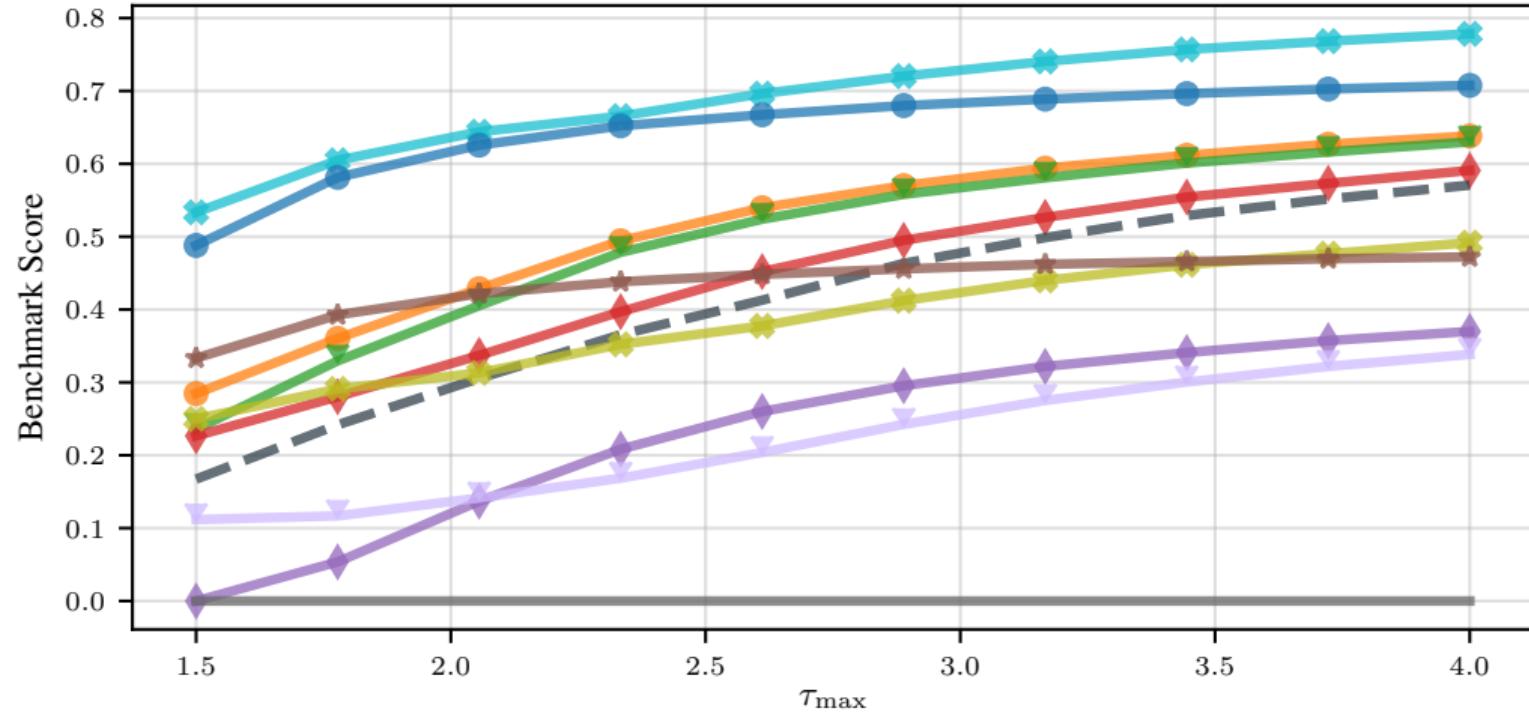
Submission	Line	Score
PYTORCH DISTRIBUTED SHAMPOO		0.7784
SCHEDULE FREE ADAMW		0.7077
GENERALIZED ADAM		0.6383
CYCLIC LR		0.6301
NADAMP		0.5909
BASELINE		0.5707
AMOS		0.4918
CASPR ADAPTIVE		0.4722
LAWA QUEUE		0.3699
LAWA EMA		0.3384
SCHEDULE FREE PRODIGY		0

(a) External tuning leaderboard



(b) External tuning performance profiles

AlgoPerf benchmark

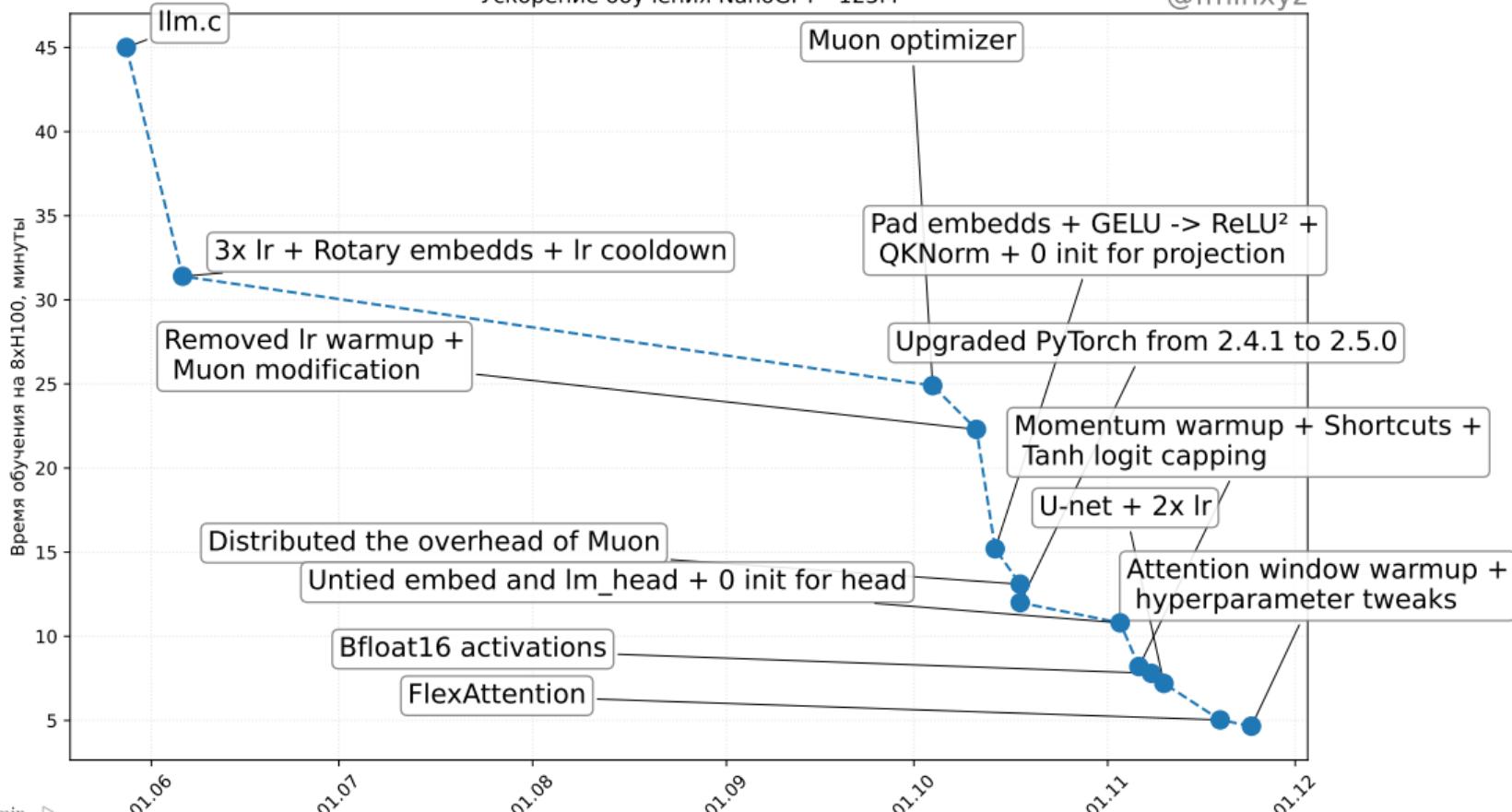


- PyTorch Distr. Shampoo
- Schedule Free AdamW
- Generalized Adam
- Cyclic LR
- NadamP
- Baseline
- Amos
- CASPR Adaptive
- Lawa Queue
- Lawa EMA
- Schedule Free Prodigy

NanoGPT speedrun

Ускорение обучения NanoGPT - 125M

@fminxyz



Работают ли трюки, если увеличить размер модели?

Scaling up the NanoGPT (124M) speedrun

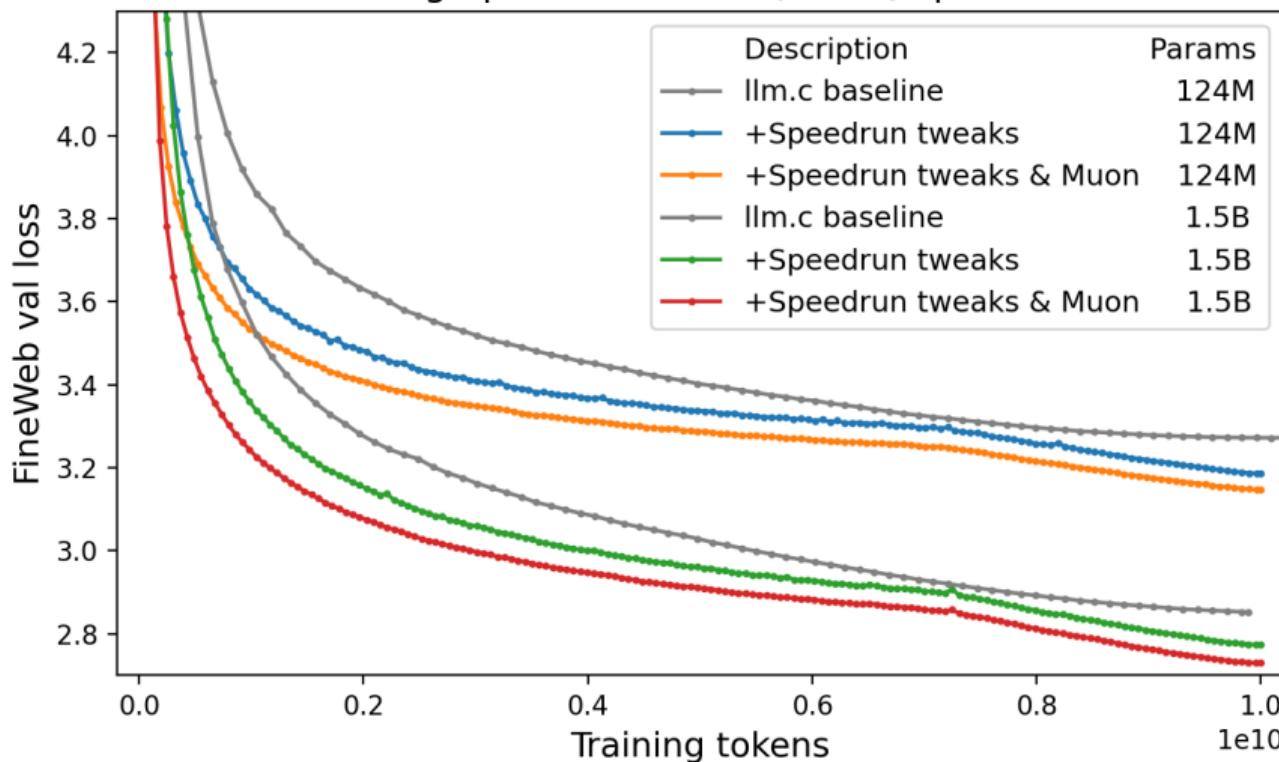


Рис. 4: Источник

Работают ли трюки, если увеличить размер модели?

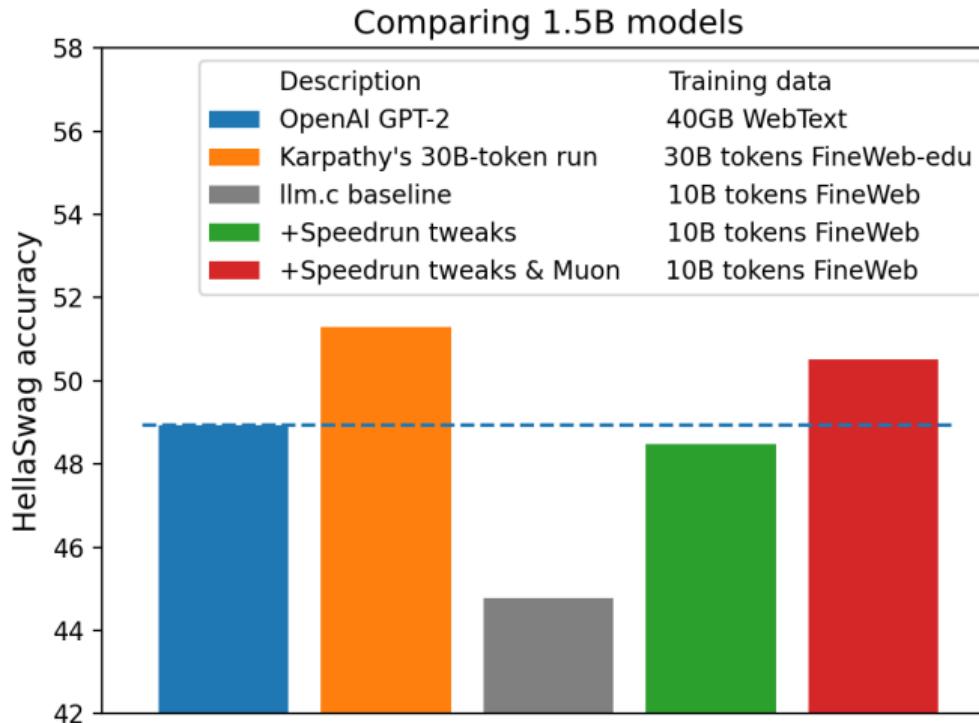


Рис. 5: Источник

Неожиданные истории

Adam работает хуже для CV, чем для LLM? ⁶

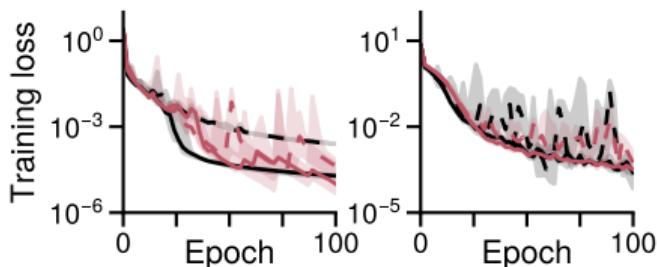


Рис. 6: CNNs on MNIST and CIFAR10

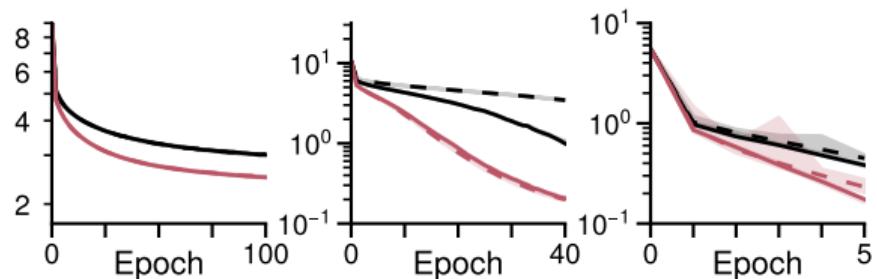


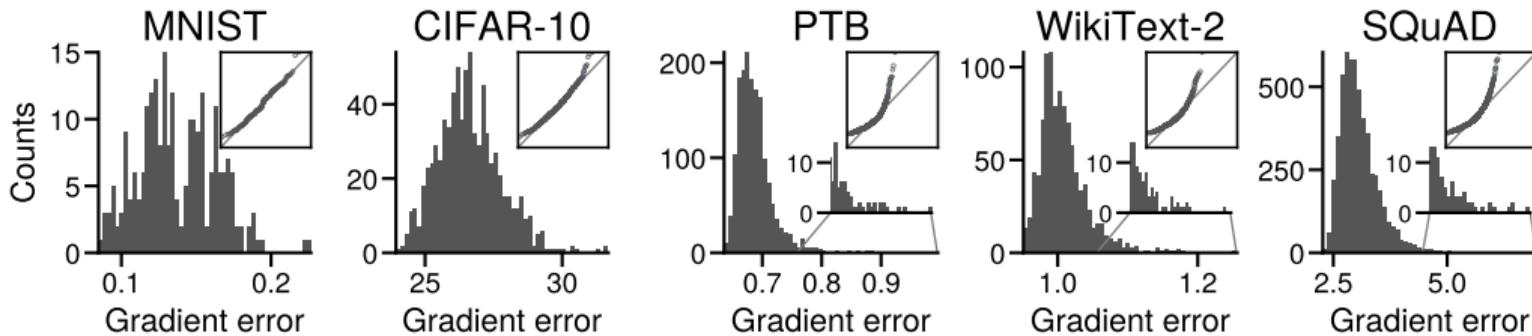
Рис. 7: Transformers on PTB, WikiText2, and SQuAD

Черные линии - SGD; красные линии - Adam.

⁶Linear attention is (maybe) all you need (to understand transformer optimization)

Почему Adam работает хуже для CV, чем для LLM? ⁷

Потому что шум градиентов в языковых моделях имеет тяжелые хвосты?



⁷Linear attention is (maybe) all you need (to understand transformer optimization)

Почему Adam работает хуже для CV, чем для LLM? ⁸

Нет! Метки имеют тяжелые хвосты!

В компьютерном зрении датасеты часто сбалансированы: 1000 котиков, 1000 песелей и т.д.
В языковых датасетах почти всегда не так: слово *the* встречается часто, слово *tie* на порядки реже

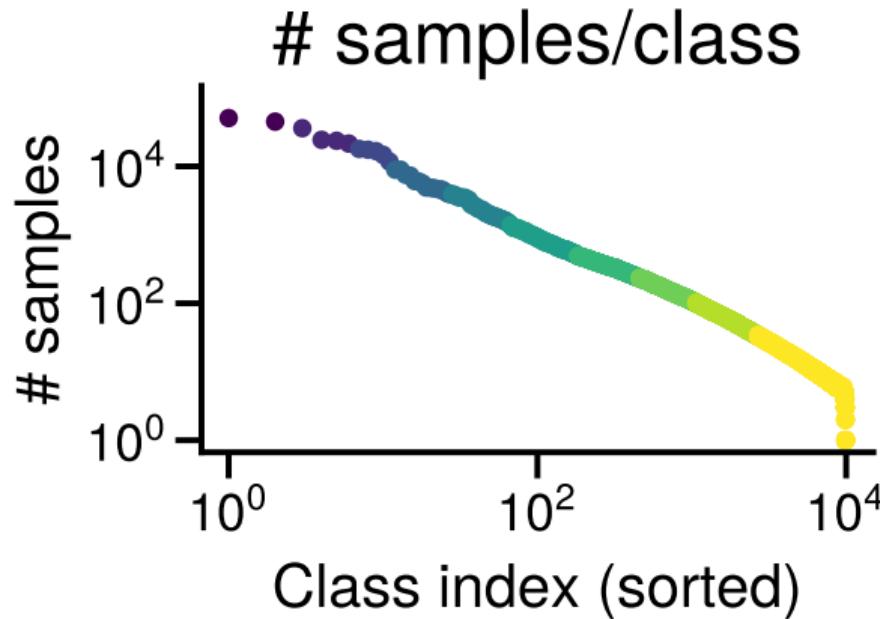
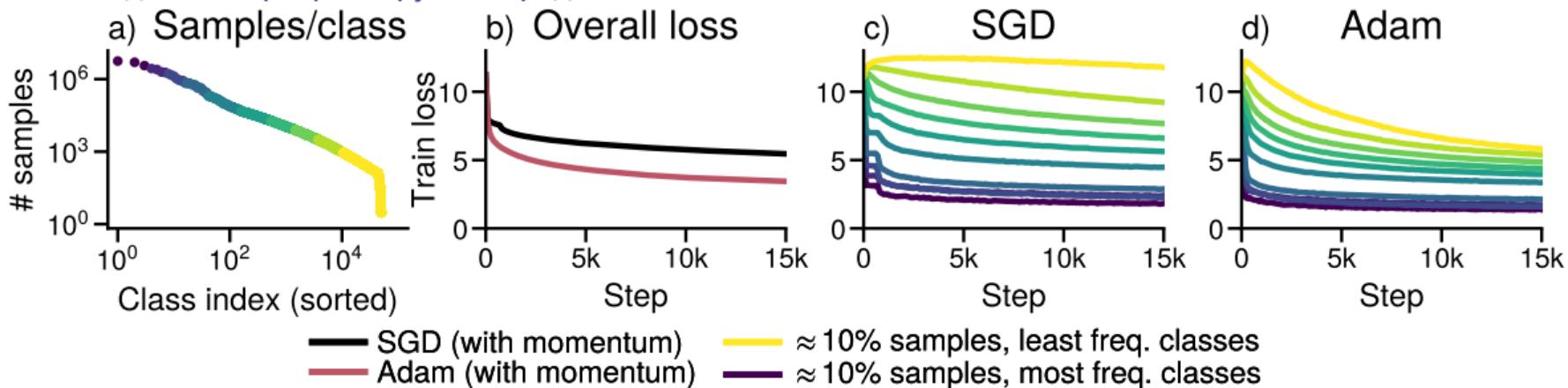


Рис. 8: Распределение частоты токенов в PTB

⁸Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models

Почему Adam работает хуже для CV, чем для LLM? ⁹

SGD медленно прогрессирует на редких классах



SGD не добивается прогресса на низкочастотных классах, в то время как Adam добивается. Обучение GPT-2 S на WikiText-103. (a) Распределение классов, отсортированных по частоте встречаемости, разбитых на группы, соответствующие $\approx 10\%$ данных. (b) Значение функции потерь при обучении. (c, d) Значение функции потерь при обучении для каждой группы при использовании SGD и Adam.

⁹Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models

- 💡 Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.
- Не инициализируйте все веса одинаково — почему?



Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.

- Не инициализируйте все веса одинаково — почему?
- Случайная инициализация: задавайте случайно, например, из гауссовского распределения $N(0, \sigma^2)$, где стандартное отклонение σ зависит от числа нейронов в слое. Это обеспечивает нарушение симметрии. *Symmetry breaking*.



Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.

- Не инициализируйте все веса одинаково — почему?
- Случайная инициализация: задавайте случайно, например, из гауссовского распределения $N(0, \sigma^2)$, где стандартное отклонение σ зависит от числа нейронов в слое. Это обеспечивает нарушение симметрии. *Symmetry breaking*.
- Можно найти более полезные советы здесь

¹⁰On the importance of initialization and momentum in deep learning Ilya Sutskever, James Martens, George Dahl, Geoffrey Hinton

Влияние инициализации весов нейронной сети на сходимость методов¹¹

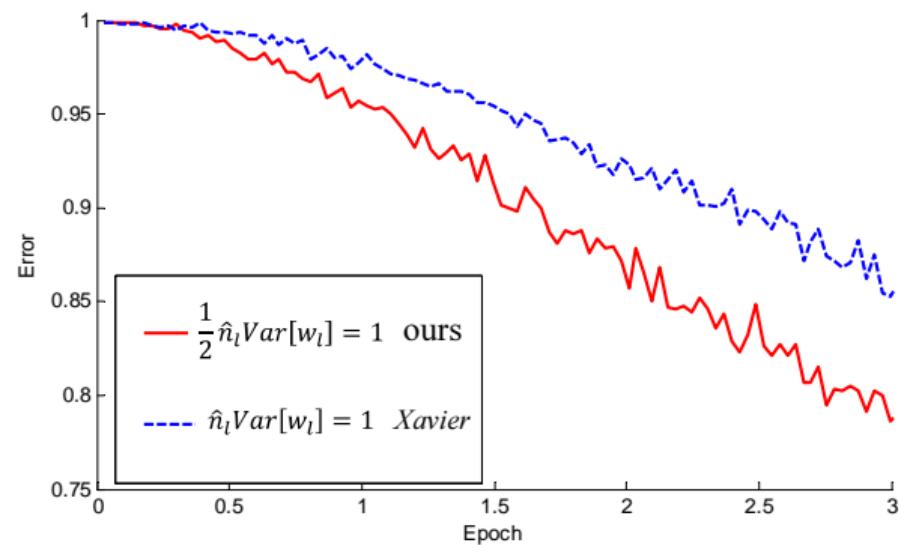


Рис. 9: 22-layer ReLU net: good init converges faster

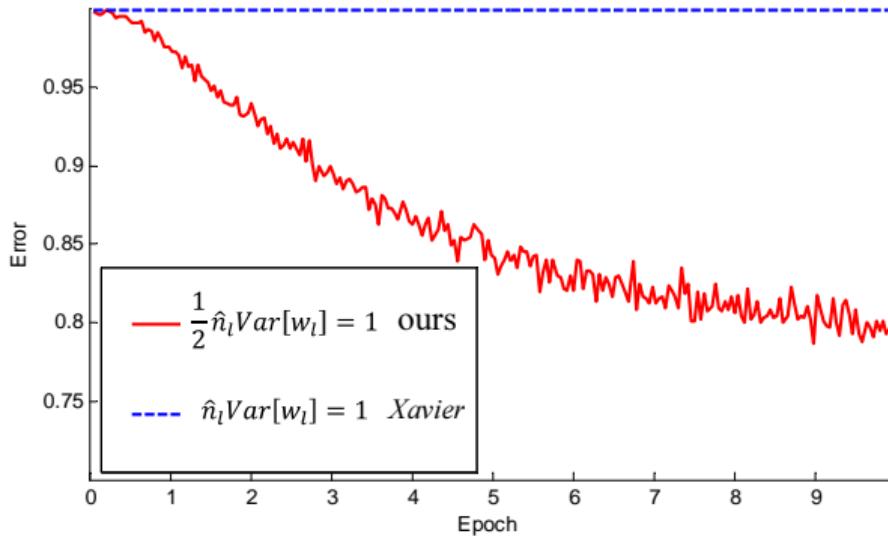
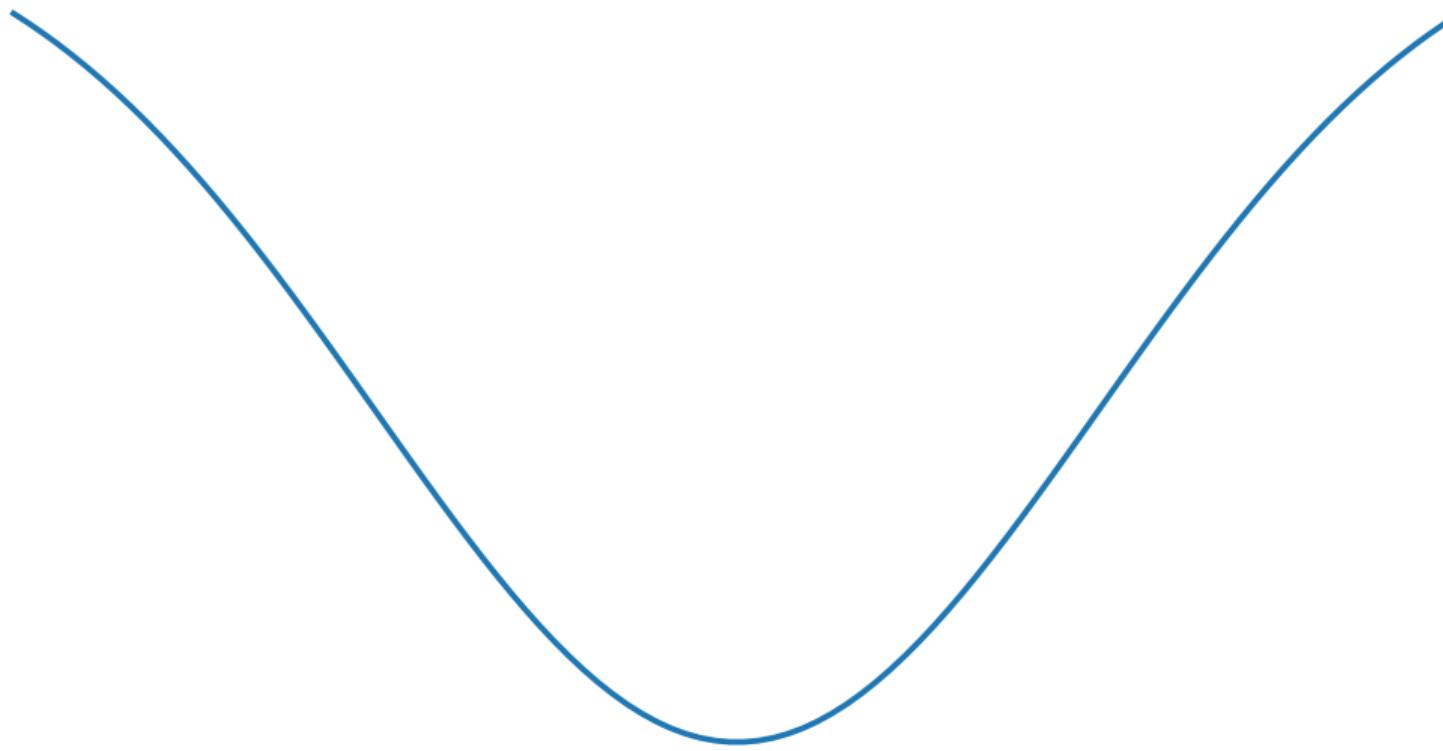


Рис. 10: 30-layer ReLU net: good init is able to converge

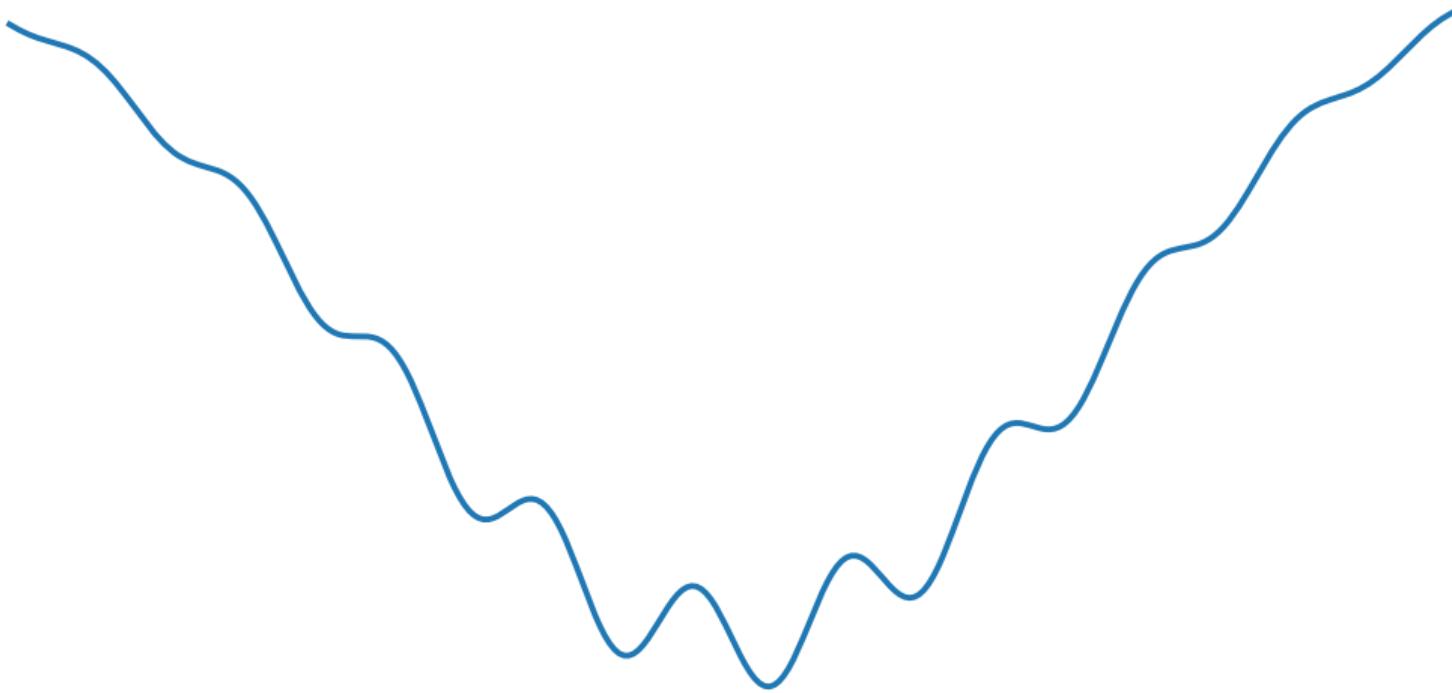
¹¹Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

Весёлые истории

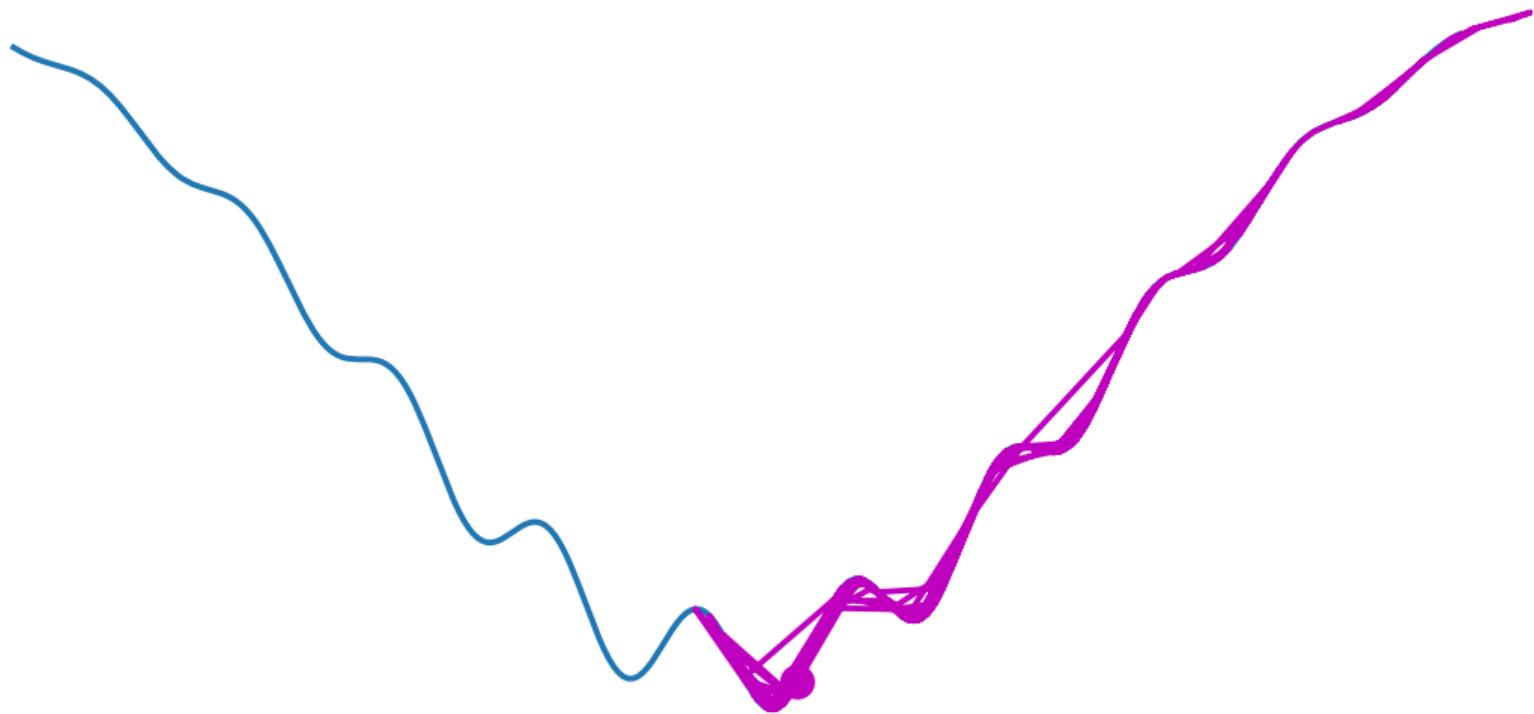
Градиентный спуск сходится к локальному минимуму



Градиентный спуск сходится к локальному минимуму



Стохастический градиентный спуск
выпрыгивает из локальных минимумов



Визуализация с помощью проекции на прямую

- Обозначим начальную точку как w_0 , представляющую собой веса нейронной сети при инициализации. Веса, полученные после обучения, обозначим как \hat{w} .

$$L(\alpha) = L(w_0 + \alpha w_1), \text{ where } \alpha \in [-b, b].$$

Визуализация с помощью проекции на прямую

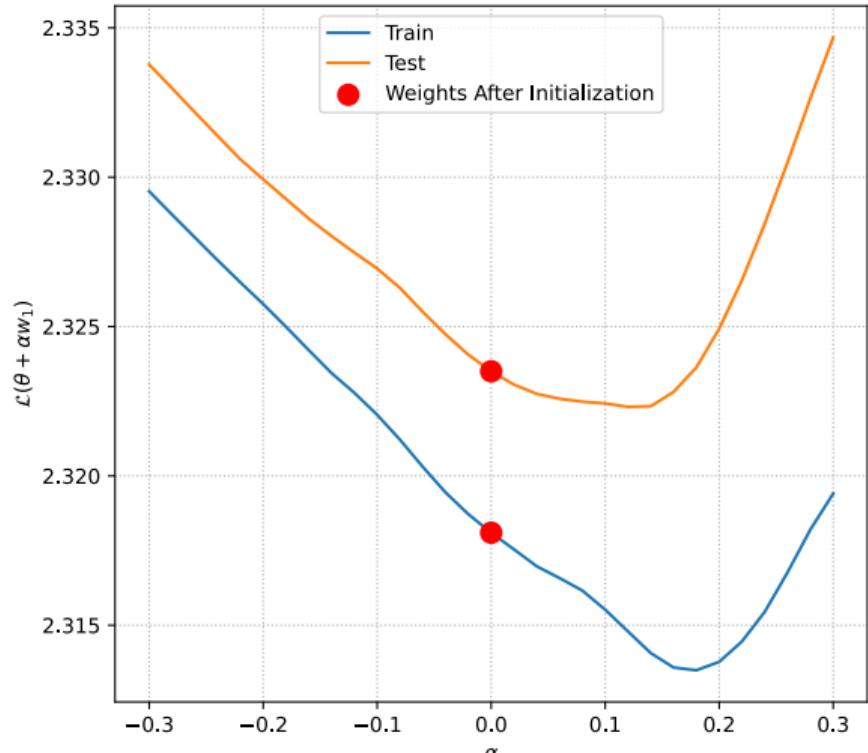
- Обозначим начальную точку как w_0 , представляющую собой веса нейронной сети при инициализации. Веса, полученные после обучения, обозначим как \hat{w} .
- Генерируем случайный вектор такой же размерности и нормы $w_1 \in \mathbb{R}^p$, затем вычисляем значение функции потерь вдоль этого вектора:

$$L(\alpha) = L(w_0 + \alpha w_1), \text{ where } \alpha \in [-b, b].$$

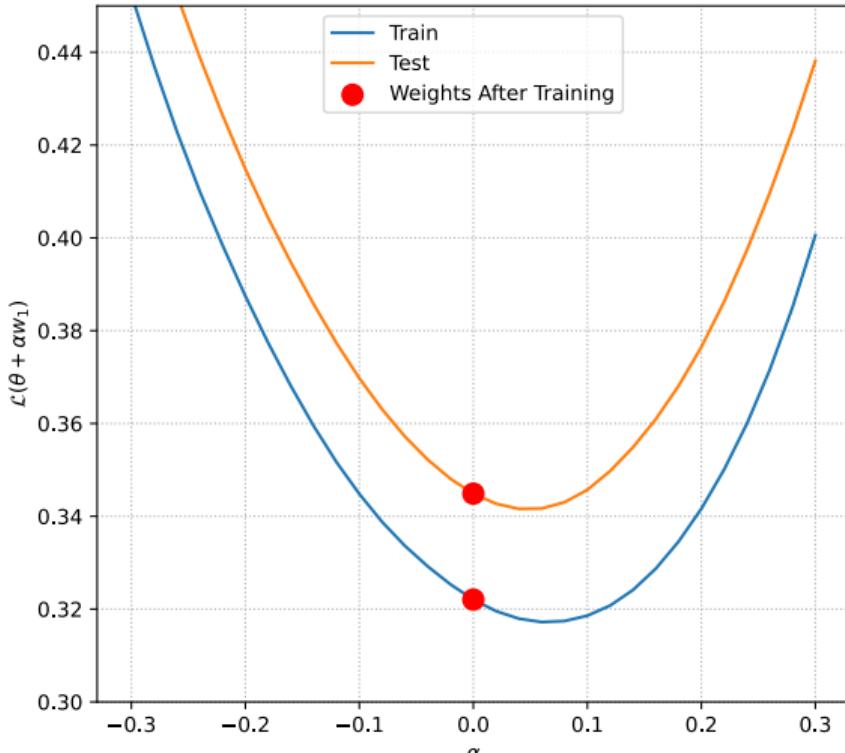
Проекция функции потерь нейронной сети на прямую

No Dropout

Loss surface, Line projection around the starting point



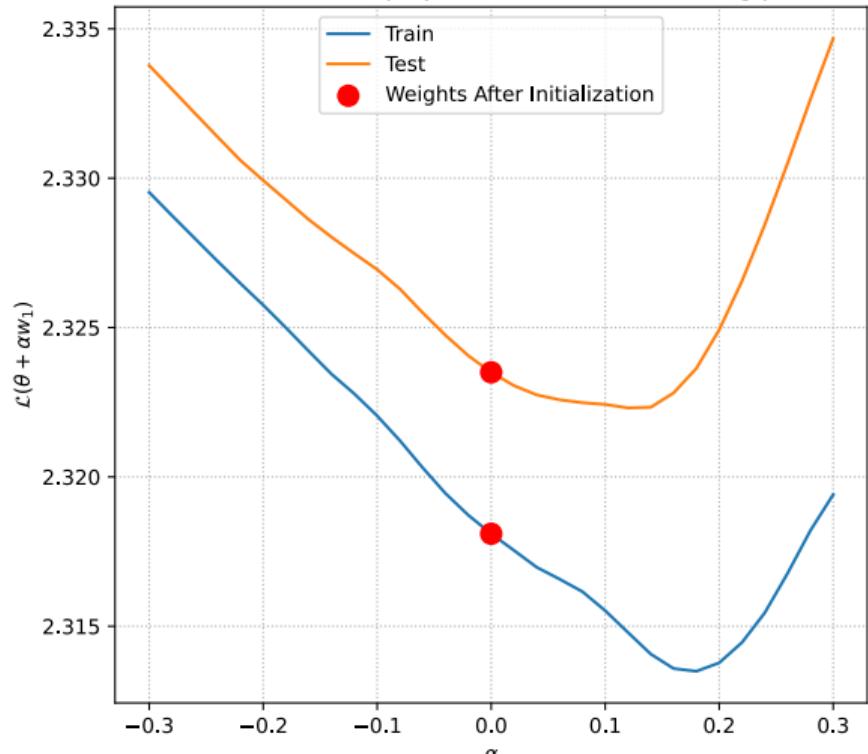
Loss surface, Line projection around the final point



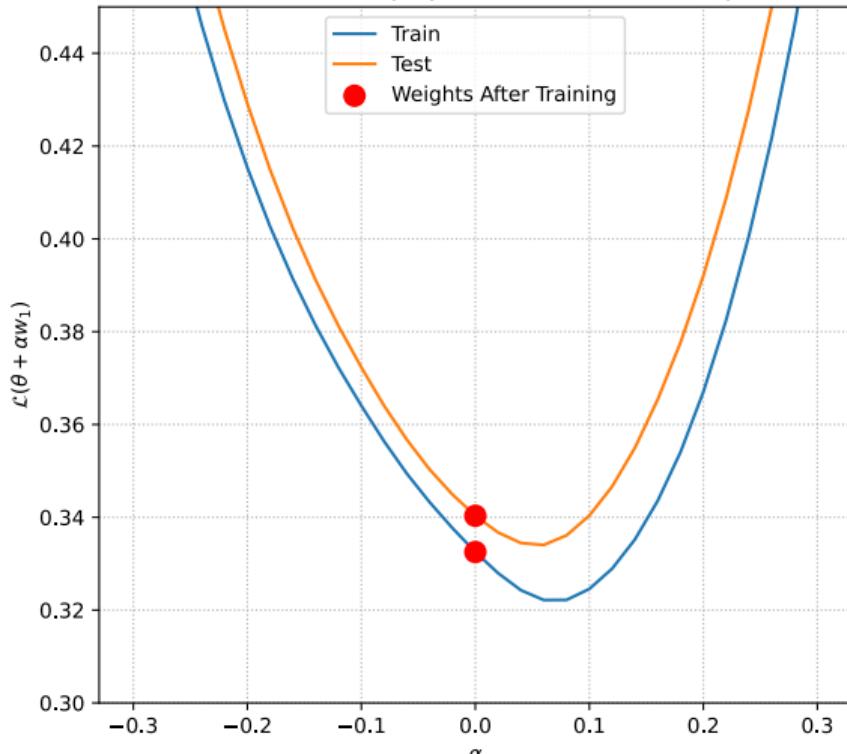
Проекция функции потерь нейронной сети на прямую

Dropout 0.2

Loss surface, Line projection around the starting point



Loss surface, Line projection around the final point

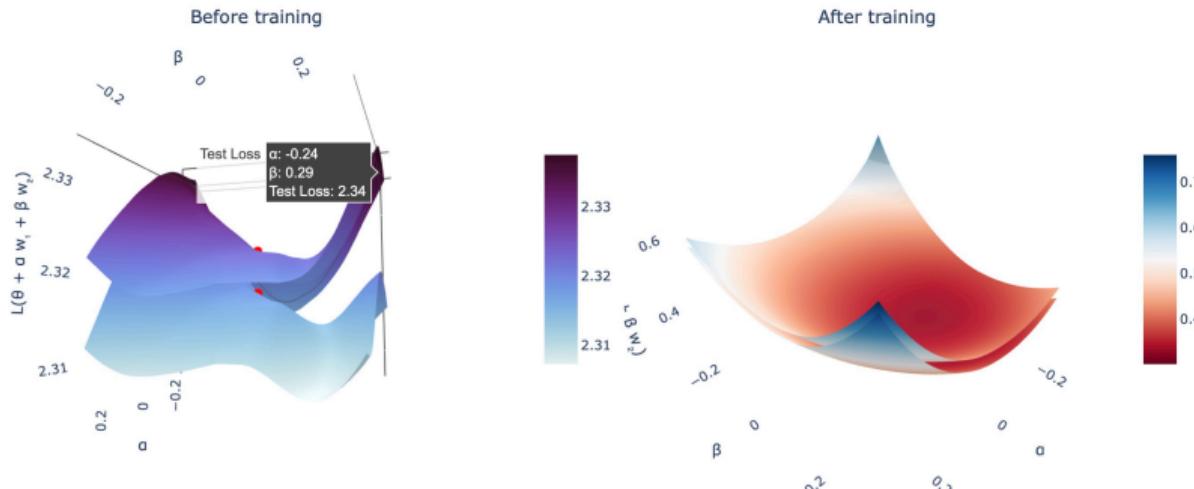


Проекция функции потерь нейронной сети на плоскость

- Мы можем расширить эту идею и построить проекцию поверхности потерь на плоскость, которая задается 2 случайными векторами.

$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2), \text{ where } \alpha, \beta \in [-b, b]^2.$$

No Dropout. Plane projection of loss surface.

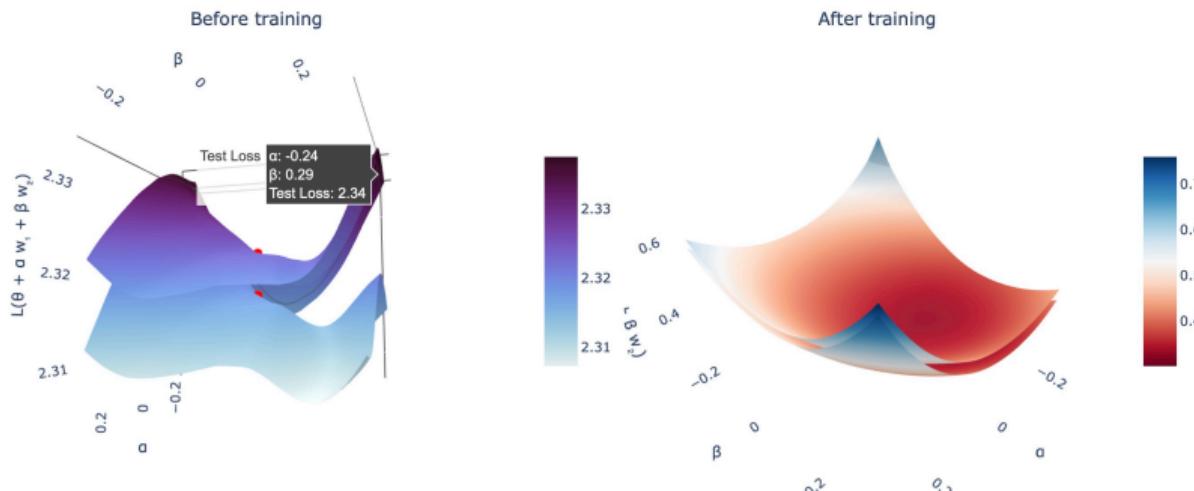


Проекция функции потерь нейронной сети на плоскость

- Мы можем расширить эту идею и построить проекцию поверхности потерь на плоскость, которая задается 2 случайными векторами.
- Два случайных гауссовых вектора в пространстве большой размерности с высокой вероятностью ортогональны.

$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2), \text{ where } \alpha, \beta \in [-b, b]^2.$$

No Dropout. Plane projection of loss surface.



Может ли быть полезно изучение таких проекций? ¹²

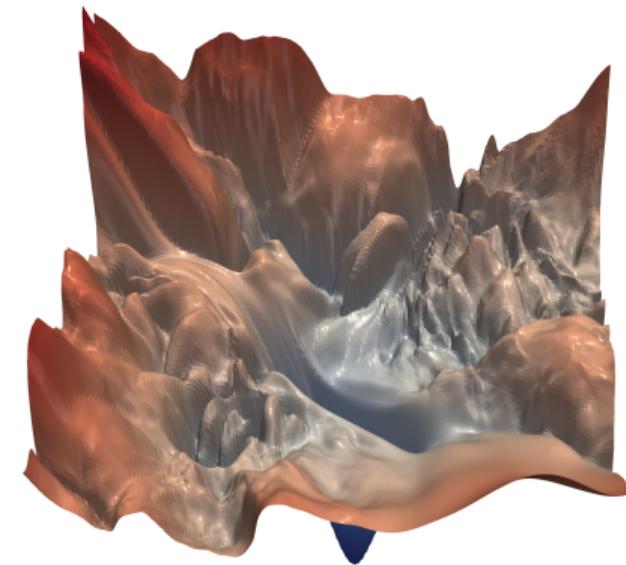


Рис. 14: The loss surface of ResNet-56
without skip connections

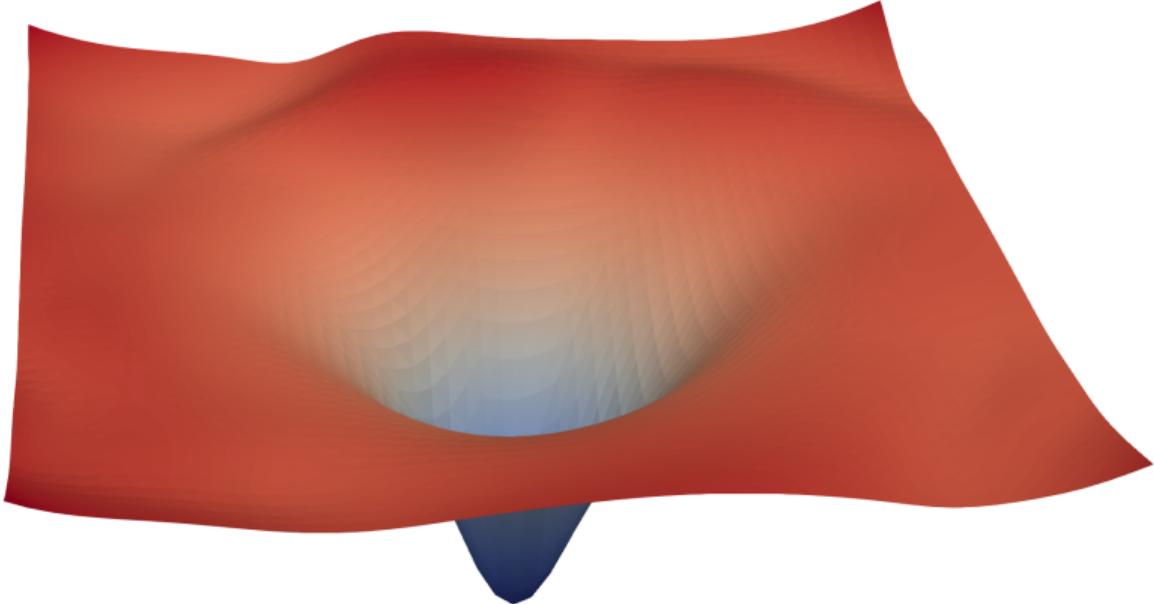


Рис. 15: The loss surface of ResNet-56 with skip connections

¹²Visualizing the Loss Landscape of Neural Nets, Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein

Может ли быть полезно изучение таких проекций, если серьезно? ¹³

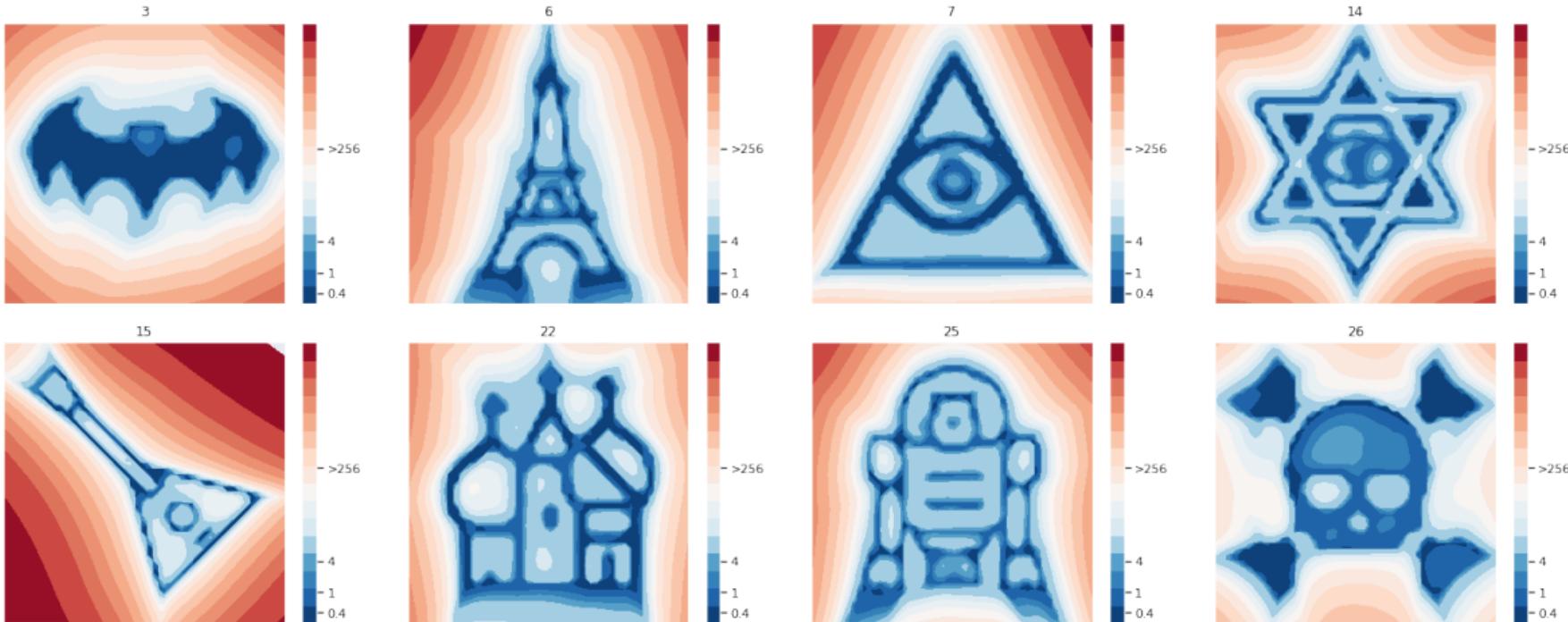
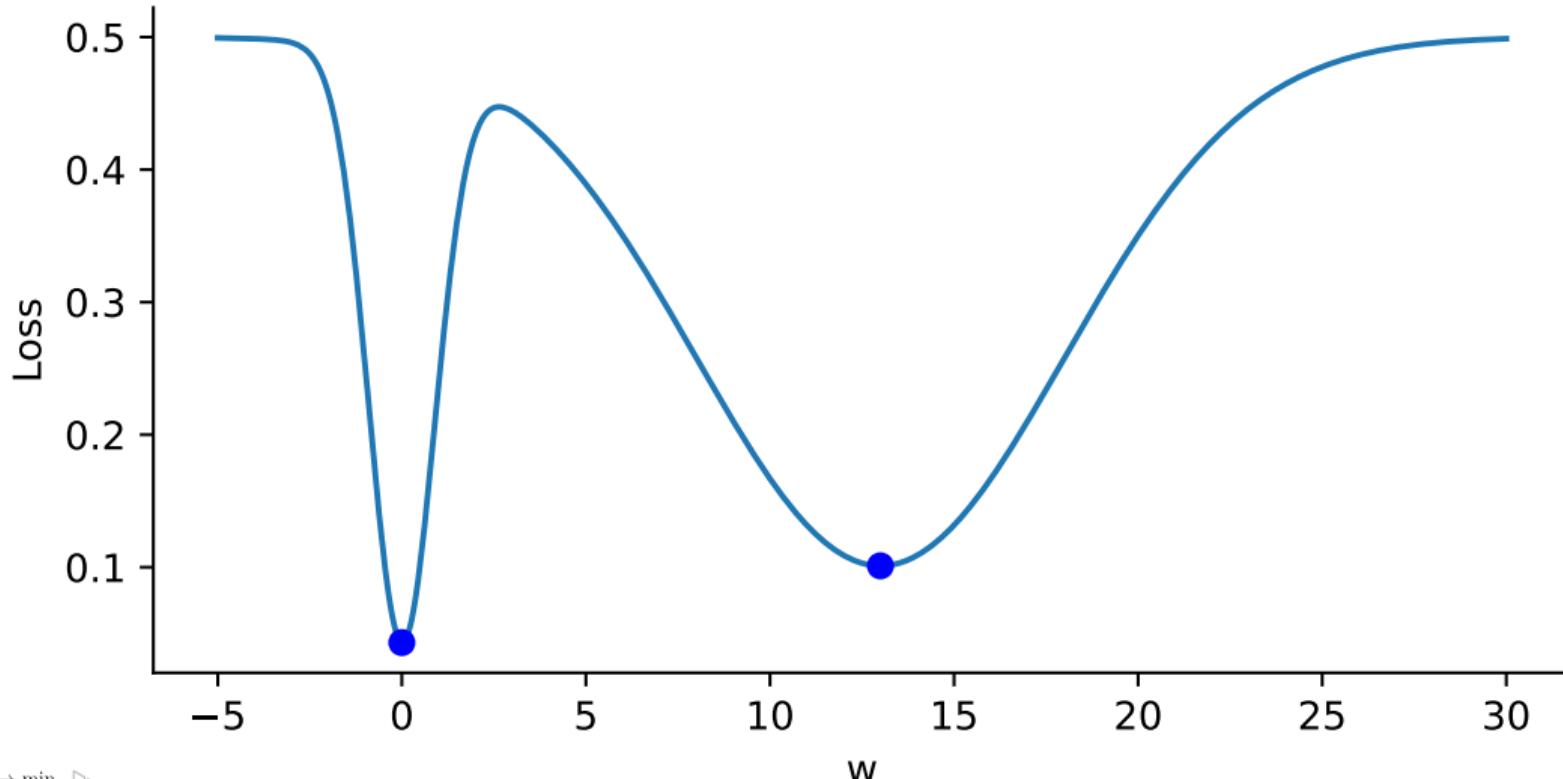


Рис. 16: Examples of a loss landscape of a typical CNN model on FashionMNIST and CIFAR10 datasets found with MPO. Loss values are color-coded according to a logarithmic scale

¹³Loss Landscape Sightseeing with Multi-Point Optimization, Ivan Skorokhodov, Mikhail Burtsev
 $f \rightarrow \min_{x,y,z}$ Весёлые истории

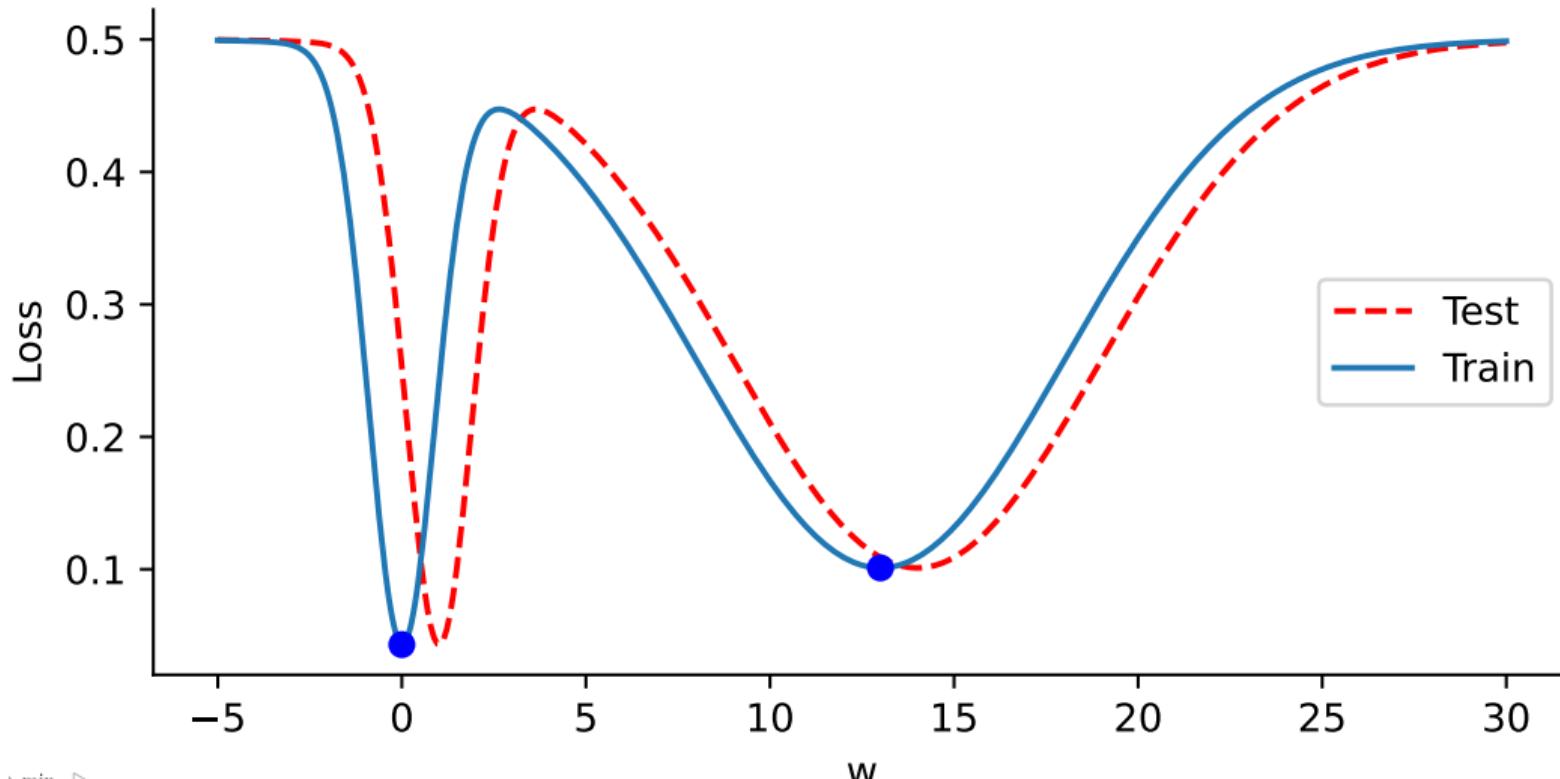
Ширина локальных минимумов

Узкие и широкие локальные минимумы



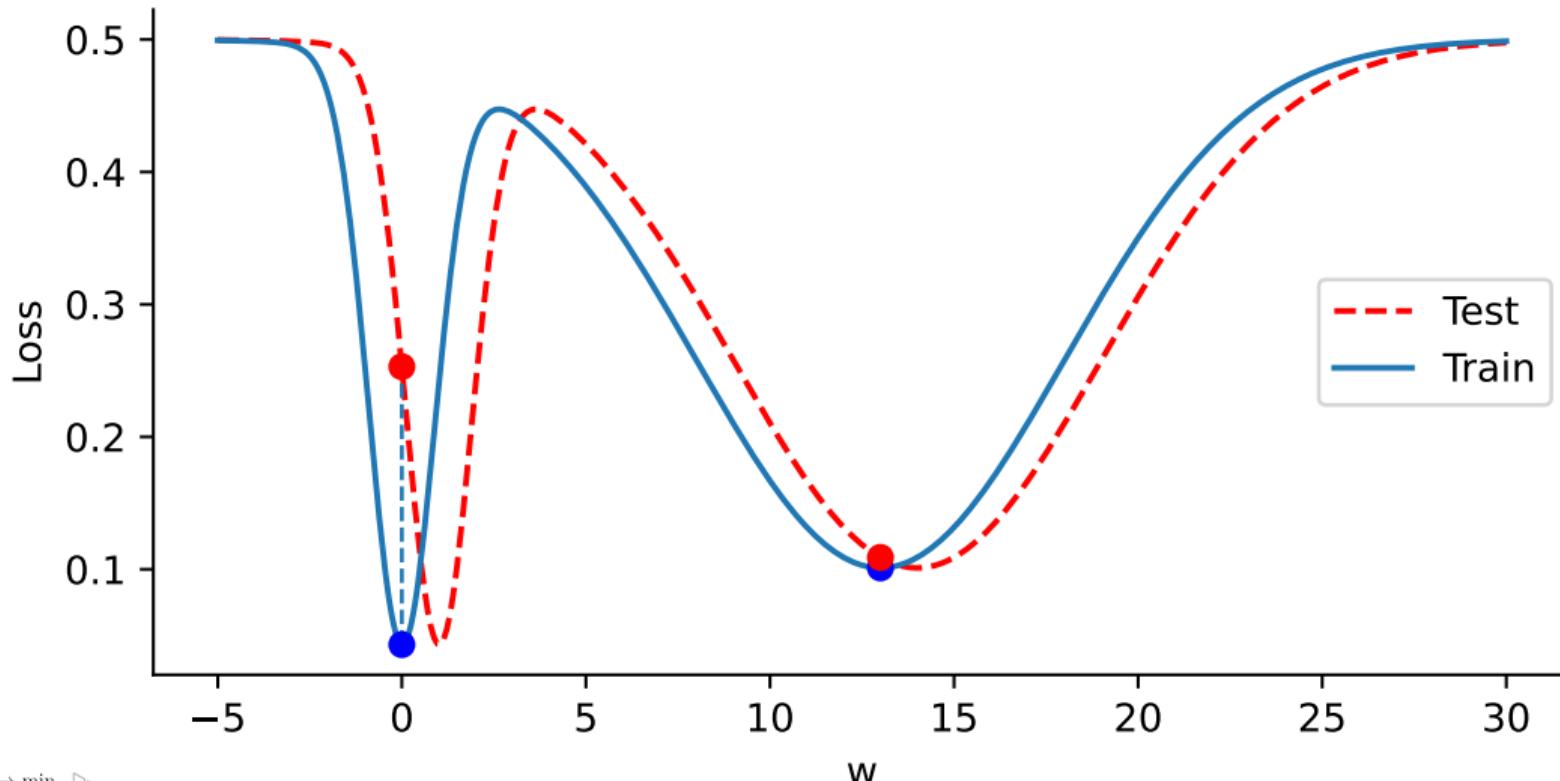
Ширина локальных минимумов

Узкие и широкие локальные минимумы



Ширина локальных минимумов

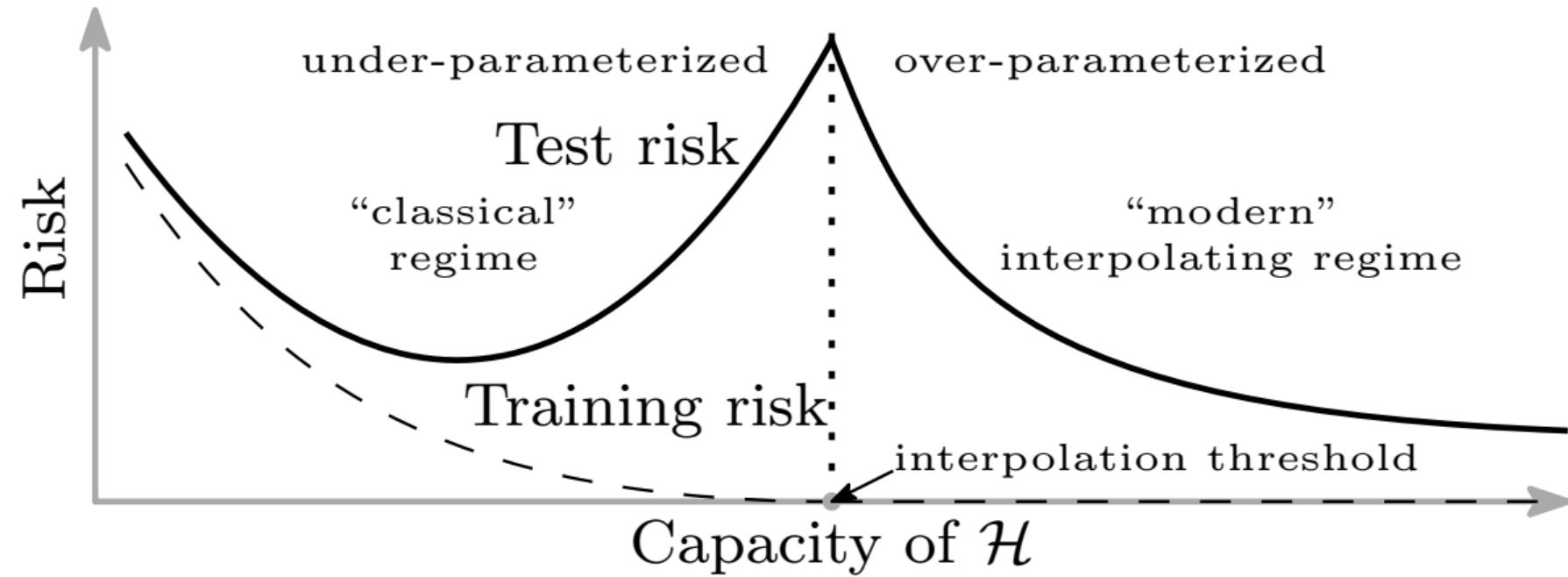
Узкие и широкие локальные минимумы



Экспоненциальный шаг обучения

- Exponential Learning Rate Schedules for Deep Learning

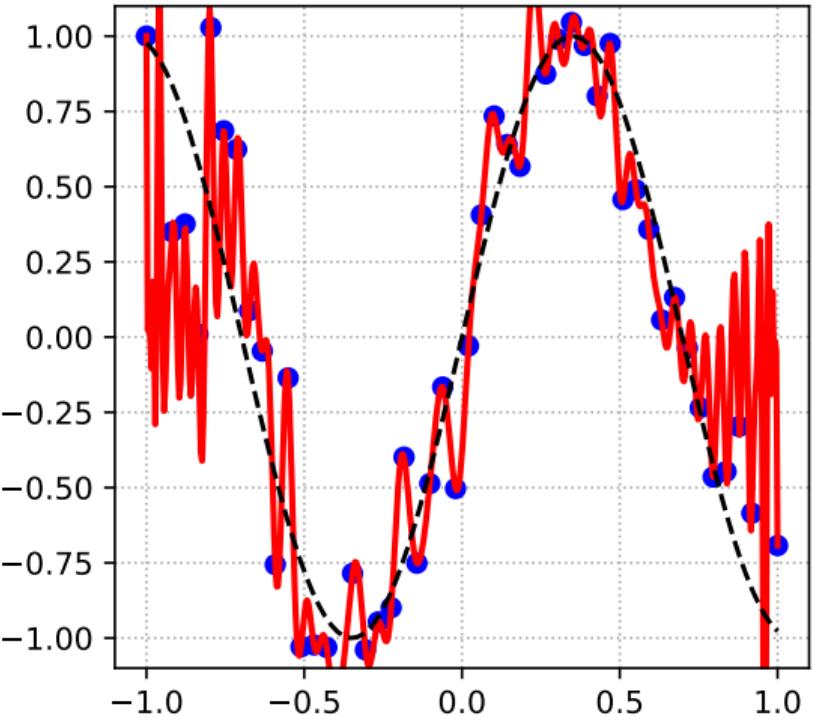
Double Descent¹⁴



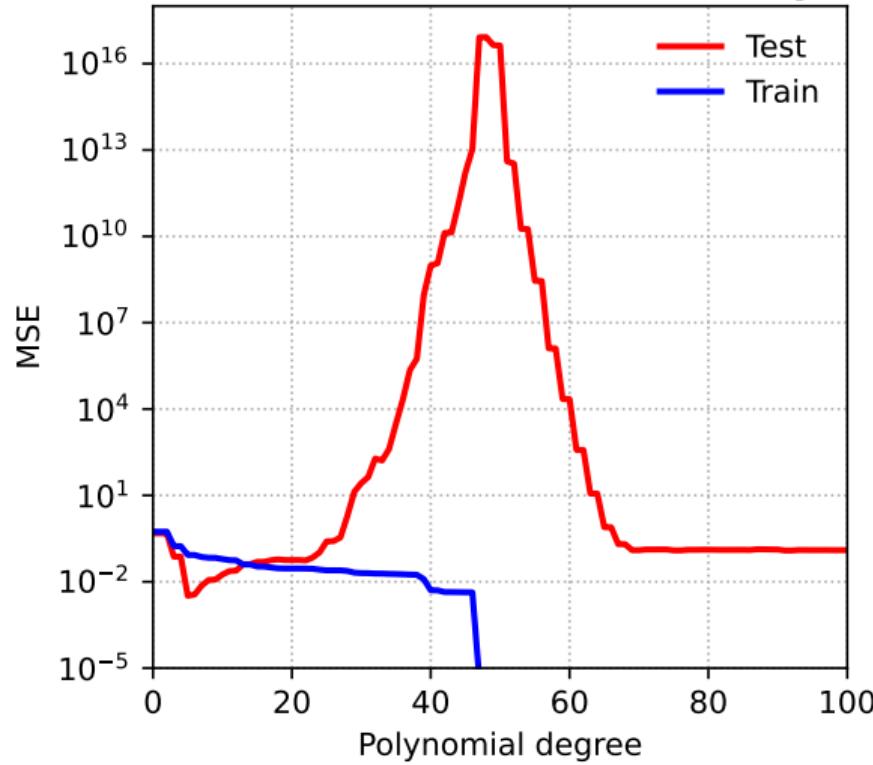
¹⁴Reconciling modern machine learning practice and the bias-variance trade-off, Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal

Double Descent

Polynomial Fitting



@fminxyz



Modular Division (training on 50% of data)

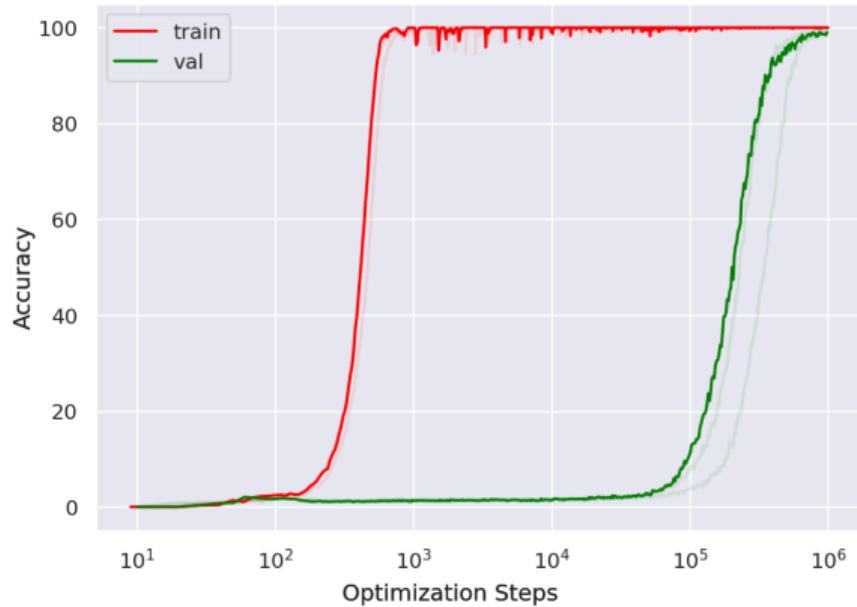


Рис. 17: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about $4 \cdot 10^5$ non-embedding parameters. Reproduction of experiments (~ half an hour) is available [here](#)

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация

Modular Division (training on 50% of data)

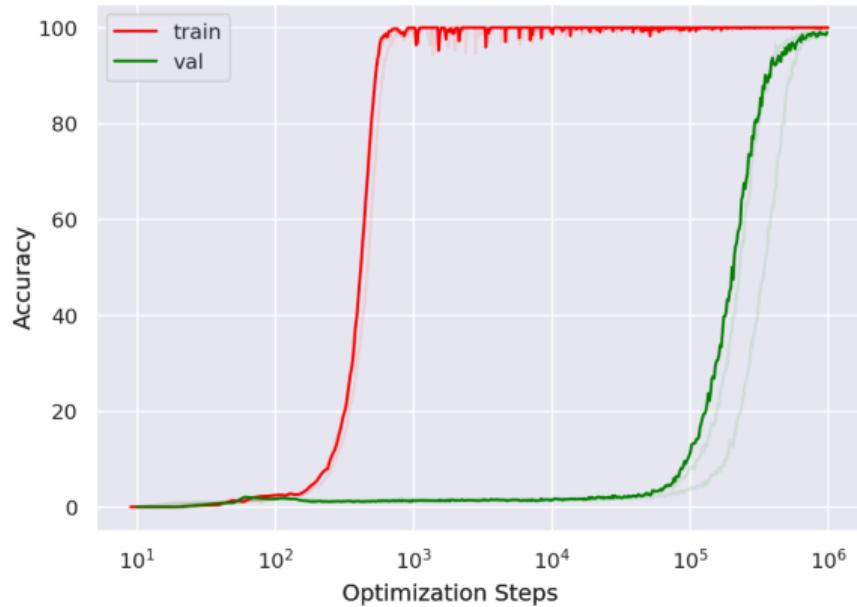


Рис. 17: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about $4 \cdot 10^5$ non-embedding parameters. Reproduction of experiments (\sim half an hour) is available here

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация
- Автор канала Свидетели Градиента собирает интересные наблюдения и эксперименты про гроккинг.

Modular Division (training on 50% of data)

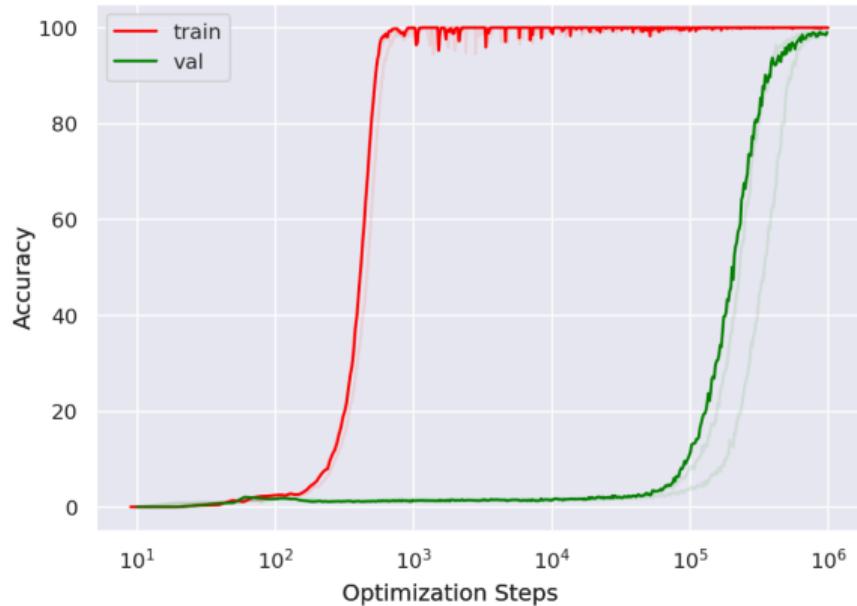


Рис. 17: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about $4 \cdot 10^5$ non-embedding parameters. Reproduction of experiments (\sim half an hour) is available here

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация
- Автор канала Свидетели Градиента собирает интересные наблюдения и эксперименты про гроккинг.
- Также есть видео с его докладом **Чем не является гроккинг**.

¹⁵Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra