

# Neural Tangent Kernel

Федоров Игорь, БПМИ171

07.10.2020

# Мотивация

- Изучаем процесс обучения
- Изучаем обобщающие свойства

# Спойлер

- Не работает, но все равно интересно (С)
- Когда-нибудь, может быть, будет полезно...

пристегнитесь

формул будет много

не все понятные

# Любопытные свойства нейросетей

- Глубокая сеть может выучить случайные метки
- Ядерные методы – тоже
- Бесконечномерную нейросеть можно интерпретировать как Гауссов процесс
- Ее параметры имеют нормальное распределение, описываемое ядром

# Пререквизиты

- Слои от 0 (вход) до L (выход)
- Активация: липшицева, дважды дифференцируема, ограниченность второй производной (ReLU $\odot$ )

# Пререквизиты

- Функция сети:  $f_\theta$
- Отображает вход в выход с параметрами  $\theta$
- Функция реализации сети  $F^{(L)} : \mathbb{R}^P \rightarrow \mathcal{F}$
- Отображает параметры сети  $\theta$  в функцию  $f_\theta$

$$P = \sum_{\ell=0}^{L-1} (n_\ell + 1)n_{\ell+1}$$

$$\{f : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L}\}$$

# ФУНКЦИЯ СЕТИ

- Активации –  $\alpha$
- Линейные слои –  $\tilde{\alpha}$

$$\alpha^{(0)}(x; \theta) = x$$

$$\tilde{\alpha}^{(\ell+1)}(x; \theta) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x; \theta) + \beta b^{(\ell)}$$

$$\alpha^{(\ell)}(x; \theta) = \sigma(\tilde{\alpha}^{(\ell)}(x; \theta)),$$

## 0.5 нормы

- Зафиксируем распределение  $p^{in}$  на пространстве входов
- Определим относительно него полуформу  $\|\cdot\|_{p^{in}}$  на пространстве функций:

$$\langle f, g \rangle_{p^{in}} = \mathbb{E}_{x \sim p^{in}} [f(x)^T g(x)]$$

- Распределение - эмпирическое, e.g. среднее по мерам Дирака:

$$\frac{1}{N} \sum_{i=0}^N \delta_{x_i}$$

# Обучаемся

- Функциональные потери  $C : \mathcal{F} \rightarrow \mathbb{R}$
- Оптимизируем функцию сети в пространстве функций
- По параметрам  $C \circ F^{(L)} : \mathbb{R}^P \rightarrow \mathbb{R}$
- Первое – м.б. выпукло, второе – точно нет
- При обучении функция сети градиентно спускается относительно градиента NTK – можем исследовать выпуклую оптимизацию



# Многомерное ядро

- Функция  $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L \times n_L}$  такая, что  $K(x, x') = K(x', x)^T$
- Она задает билинейное отображение на пространстве функций:

$$\langle f, g \rangle_K := \mathbb{E}_{x, x' \sim p^{in}} [f(x)^T K(x, x') g(x')]$$

- К положительно определено относительно полуформы, если:

$$\|f\|_{p^{in}} > 0 \implies \|f\|_K > 0$$

# Двойственность

- Двойственное пространство:  $\mathcal{F}^* = \{\mu : \mathcal{F} \rightarrow \mathbb{R} | \mu = \langle d, \cdot \rangle_{p_{in}}, d \in \mathcal{F}\}$
- Частичное применение ядра  $K_{i,\cdot}(x, \cdot)$  - функция в  $\mathcal{F}$
- Зададим новое отображение  $\Phi_K : \mathcal{F}^* \rightarrow \mathcal{F}$ , отображает  $\mu$  в  $f_\mu = \Phi_K(\mu)$

$$f_{\mu,i}(x) = \mu K_{i,\cdot}(x, \cdot) = \langle d, K_{i,\cdot}(x, \cdot) \rangle_{p^{in}}$$

# Ядерный градиент

- Функционал  $C$  зависит только от значения функции сети на эл-тах датасета
- Его производную в точке  $f_0 \in \mathcal{F}$  можно представить как элемент сопряженного пространства

$$\partial_f^{in} C|_{f_0} = \langle d|_{f_0}, \cdot \rangle_{p^{in}}$$

# Ядерный градиент

- Определим  $\nabla_K C|_{f_0} \in \mathcal{F}$  как  $\Phi_K \left( \partial_f^{in} C|_{f_0} \right)$
- Производная выше определена на датасете, а градиент везде!

$$\nabla_K C|_{f_0}(x) = \frac{1}{N} \sum_{j=1}^N K(x, x_j) d|_{f_0}(x_j)$$

# Ядерный GD

- $f(t)$  – временная функция

$$\partial_t f(t) = -\nabla_K C|_{f(t)}$$

- Как меняются функциональные потери?

$$\partial_t C|_{f(t)} = -\langle d|_{f(t)}, \nabla_K C|_{f(t)} \rangle_{p^{in}} = -\|d|_{f(t)}\|_K^2$$

- Положительной определенности ядра, выпуклости и ограниченности снизу С достаточно для глобальной сходимости в пределе

# Случайные функции

- Апроксимируем  $\mathbb{P}$  функциями, сэмплированными из пространства

Ковариация:  $\mathbb{E}[f_k^{(p)}(x)f_{k'}^{(p)}(x')] = K_{kk'}(x, x')$

- Функции задают случайную реализацию  $F^{lin} : \mathbb{R}^P \rightarrow \mathcal{F}$

$$\theta \mapsto f_\theta^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^P \theta_p f^{(p)}$$

- Частные производные:  $\partial_{\theta_p} F^{lin}(\theta) = \frac{1}{\sqrt{P}} f^{(p)}$

# Диффуры

- Оптимизация параметра – решение ОДУ

$$\partial_t \theta_p(t) = -\partial_{\theta_p}(C \circ F^{lin})(\theta(t)) = -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{f_{\theta(t)}^{lin}} f^{(p)} = -\frac{1}{\sqrt{P}} \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}}$$

- Оптимизация функции:

$$\partial_t f_{\theta(t)}^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^P \partial_t \theta_p(t) f^{(p)} = -\frac{1}{P} \sum_{p=1}^P \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}} f^{(p)}$$

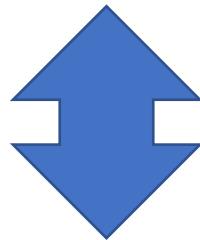
- Правая часть – ядерный градиент  $-\nabla_{\tilde{K}} C$  касательного ядра

$$\tilde{K} = \sum_{p=1}^P \partial_{\theta_p} F^{lin}(\theta) \otimes \partial_{\theta_p} F^{lin}(\theta) = \frac{1}{P} \sum_{p=1}^P f^{(p)} \otimes f^{(p)}$$

$$\tilde{K}_{ii'}(x, x') = \frac{1}{P} \sum_{p=1}^P f_i^{(p)}(x) f_{i'}^{(p)}(x')$$

# Собираем конструктор

Градиентный спуск на потерях по параметрам  $C \circ F^{lin}$



Ядерный градиентный спуск с ядром  $\tilde{K}$  на пространстве функций

- В бесконечной параметризации, случайное ядро по ЗБЧ сходится к фиксированному К (назовем предельным ядром)
- Обычная оптимизация аппроксимирует ЯГС относительно К

если вы что-то не поняли, самое время спросить

# NTK

- Теперь вернемся к нашей сети, оптимизируем  $C \circ F^{(L)}$
- Динамика:  $\partial_t f_{\theta(t)} = -\nabla_{\Theta^{(L)}} C|_{f_{\theta(t)}}$
- NTK:  $\Theta^{(L)}(\theta) = \sum_{p=1}^P \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta)$
- Проблемы: реализация нелинейна, производные зависят от параметров, ядро случайно на инициализации...

Инсайт: ядро показывает,  
как изменение  
поведения функции в  
точке X отразится на  
поведении в точке Y

# ДГТВ

- В пределе, каждая  $f_{\theta,k}$  - функция сети до слоя  $k$  – сходится к центрированному Гауссову процессу с ковариацией  $\Sigma^{(k)}$

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$$

$$\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(L)})} [\sigma(f(x)) \sigma(f(x'))] + \beta^2$$

# Сходимость

- НТК:  $\Theta^{(L)} \rightarrow \Theta_\infty^{(L)} \otimes Id_{n_L}$  при бесконечной размерности

$$\Theta_\infty^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \rightarrow \mathbb{R}$$

$$\Theta_\infty^{(1)}(x, x') = \Sigma^{(1)}(x, x')$$

$$\Theta_\infty^{(L+1)}(x, x') = \Theta_\infty^{(L)}(x, x') \dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x')$$

$$\dot{\Sigma}^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(L)})} [\dot{\sigma}(f(x)) \dot{\sigma}(f(x'))]$$

$\dot{\sigma}$  - производная

# Константность

- Параметры обновляются через направление  $d_t \in \mathcal{F}$  ( $d_t = -d|_{f_{\theta(t)}}$ )

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}$$

- Если интеграл  $\int_0^T \|d_t\|_{p^{in}} dt$  ограничен для  $T$ , то для  $t \in [0, T]$  верно следующее задание динамики в пределе:

$$\partial_t f_{\theta(t)} = \Phi_{\Theta_\infty^{(L)} \otimes Id_{n_L}} \left( \langle d_t, \cdot \rangle_{p^{in}} \right)$$

это последний слайд с формулами, честно

# Fun facts

- Предел NTK положительно определен относительно  $p^{in}$ , если линейная оболочка над частными производными  $\partial_{\theta_p} F^{(L)}$  задает функциональное пространство
- В общем-то это верно для очень большого числа распределений на входе и активаций
- Гарантированная глобальная сходимость ЯГС!

# Еще более fun facts

- ЯГС учит модель на ядерные главные компоненты входа, причем чем главнее, тем быстрее
- Ранняя остановка: успеваем выучить только значимые компоненты

- При случайной инициализации сети из заданного нормального распределения, один шаг ГС будет иметь примерно одинаковый эффект для всех инициализаций  
(с точки зрения разности функций сети на шагах 1 и 0)
- «Однаковость» растет с ростом ширины сети
- NTK объясняет это наблюдение

# Заглянем внутрь

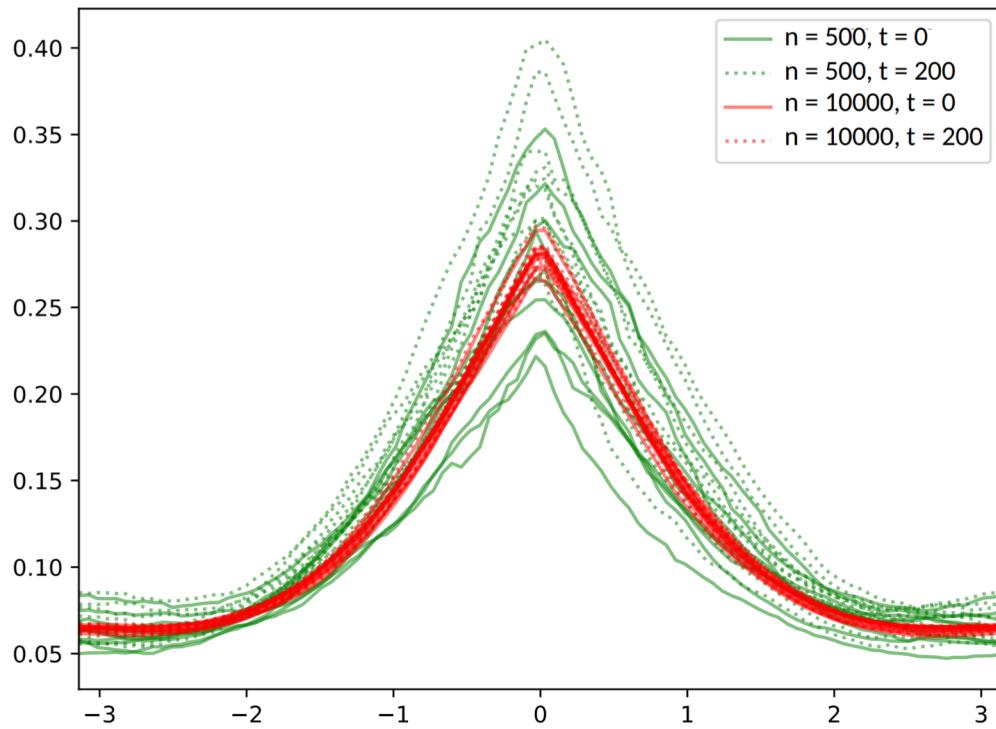


Рис. 1. Сходимость НТК. Сеть глубиной 4, точка  $(1,0)$ , базис – косинус и синус, аппроксимируем их произведение

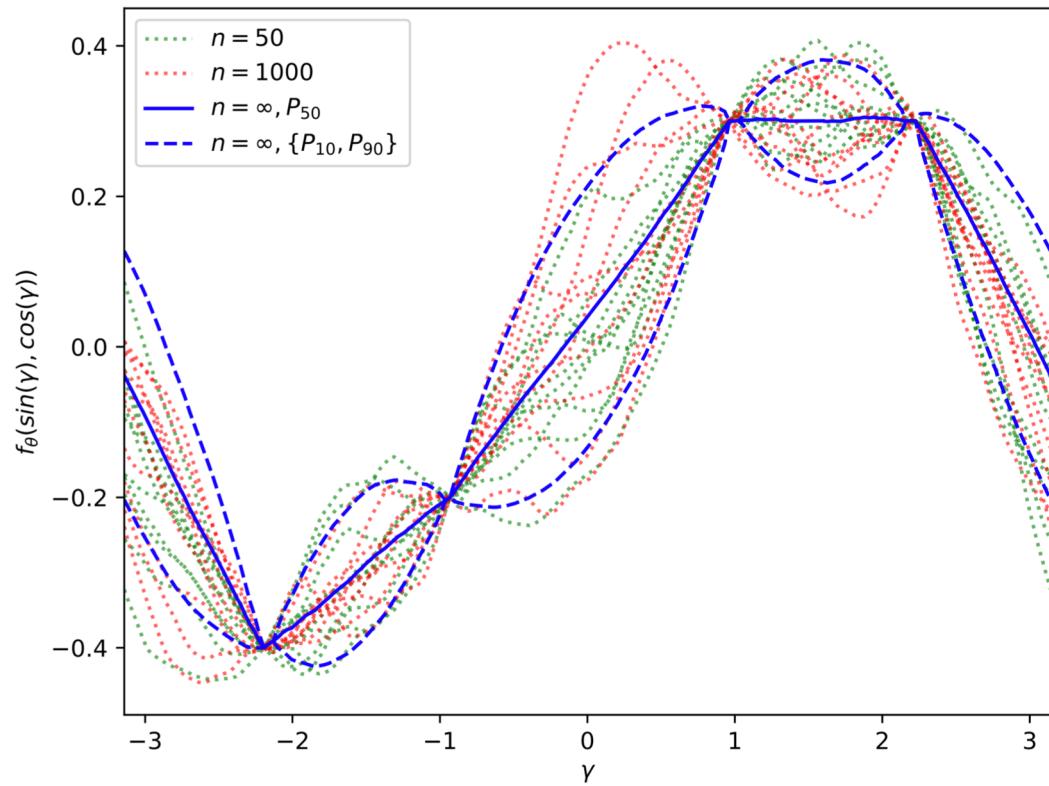


Рис. 2. Регрессия на 4 точки. Сравнение теоретического нормального распределения бесконечной сети с реальными

# Ложка дегтя

- Даже лучшие сети, для которых удается вычислить NTK, работают на MNIST и CIFAR где-то на 7% хуже, чем с обычным обучением
- Тем не менее, постепенно удается уменьшать разрыв



# Вопросы

- Приведите определение функции сети и функции реализации. В чем их различия?
- Определите НТК. Расшифруйте все обозначения.
- Выпишите ДУ, характеризующее динамику обучения сети в пространстве функций

# Источники

- <https://arxiv.org/pdf/1806.07572.pdf> - NTK
- <https://arxiv.org/abs/1802.01396> – о связи ядерных методов и сетей
- <https://rajatvd.github.io/NTK/> - интересные визуализации

# Neural Tangent Kernel

Федоров Игорь, БПМИ171

07.10.2020