# Методы оптимизации
# Семинар 15. Безградиентные методы федеративного обучения с $l_1$ и $l_2$-рандомизацией для задач негладкой выпуклой стохастической оптимизации

Лобанов Александр Владимирович

Московский физико-технический институт
Факультет инноваций и высоких технологий

lobanov.av@mipt.ru

08 декабря 2022

# Содержание

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

**Вопрос**

Когда следует использовать безградиентные алгоритмы?

# Безгратиентные алгоритмы

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

## Вопрос

Когда следует использовать безградиентные алгоритмы?

## Критерии оптимальности

1. Число оракульных вызовов: $T$
2. Число последовательных итераций метода: $N$
3. Максимально допустимый уровень шума $\Delta$

# Безгратиентные алгоритмы

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

## Вопрос

Когда следует использовать безградиентные алгоритмы?

## Подходы для создания безградиентных методов

- Гладкий случай
  - Полная градиентная аппроксимация
  - Покоординатная рандомизация
  - Рандомизация с помощью ядра
- **Негладкий случай**
  - $l_1$ рандомизация
  - $l_2$ рандомизация

# Постановка задачи

## Постановка задачи

Рассматривается стохастическая негладкая выпуклая задача оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) := \mathbb{E}_\xi \left[ f(x, \xi) \right]$$

## Безградиентный оракул

Предполагается, что безградиентный оракул возвращает значение функции $f(x)$, возможно, с некоторым враждебным шумом $\delta(x)$:

$$f_\delta(x) := f(x) + \delta(x)$$

# Предположения

**Предположение 1. (Липшицева непрерывная функция).**

Функция $f(x, \xi)$ является $M$-Липшицевой непрерывной функцией в $l_p$-норме, то есть для всех $x, y \in Q$ имеем

$$|f(y, \xi) - f(x, \xi)| \leq M(\xi)\|y - x\|_p.$$

Более того, существует положительная константа $M$, которая определяется следующим образом: $\mathbb{E}\left[M^2(\xi)\right] \leq M^2$. В частности, для $p = 2$ используем обозначение $M_2$ для константы Липшица.

**Предположение 2. (Ограниченность шума).**

Для всех $x \in Q$ выполняется $|\delta(x)| \leq \Delta$, где $\Delta$ – уровень неточности (шума).

**Предположение 3.**

Для всех $x \in Q$ выполняется $\mathbb{E}_\xi\left[|f(x, \xi)|^2\right] \leq G^2$.

# Основная идея

**Негладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

$\longrightarrow$

**Гладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

# Схемы сглаживания [1, 2]

## Definition

Пусть $l_p$-шар определяется, как $B_p^d(r) := \{x \in \mathbb{R} : \|x\|_p \leq r\}$. Тогда гладкая аппроксимация негладкой функции $f(x)$ выглядит следующим образом

$$f_\gamma(x) := \mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right],$$

где $\gamma > 0$, $\tilde{e}$ — случайный вектор, равномерно распределенный на $B_p^d(1)$ (далее ограничимся рассмотрением случаев $p = 1$ и $p = 2$).

# Свойства функции $f_\gamma(x)$

## Случай, когда $\tilde{e} \in RB_2^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$;
2. $f_\gamma$ — $M$-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. $f_\gamma$ имеет $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$
где $q$ такой, что $1/p + 1/q = 1$.

## Первое свойство

$$f_\gamma(x) = \mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] \geq \mathbb{E}_{\tilde{e}}\left[f(x) + \langle\nabla f(x), \gamma\tilde{e}\rangle\right] = \mathbb{E}_{\tilde{e}}\left[f(x)\right] = f(x).$$
$$|f_\gamma(x) - f(x)| = \left|\mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] - f(x)\right| \leq \mathbb{E}_{\tilde{e}}\left[|f(x + \gamma\tilde{e}) - f(x)|\right] \leq \gamma M_2 \mathbb{E}_{\tilde{e}}\left[\|\tilde{e}\|_2\right] \leq \gamma M_2.$$

# Свойства функции $f_\gamma(x)$

## Случай, когда $\tilde{e} \in RB_2^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$;
2. $f_\gamma$ — $M$-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. $f_\gamma$ имеет $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$
где $q$ такой, что $1/p + 1/q = 1$.

## Второе свойство

$$|f_\gamma(y) - f_\gamma(x)| = |\mathbb{E}_{\tilde{e}}\left[f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})\right]| \leq \mathbb{E}_{\tilde{e}}\left[|f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})|\right] \leq M\|y - x\|_p$$

# Свойства функции $f_\gamma(x)$

## Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \le f_\gamma(x) \le f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. $f_\gamma$ — $M$-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \le M\|y - x\|_p,$$
3. $f_\gamma$ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \le L_{f_\gamma}\|y - x\|_p.$$

где $q$ такой, что $1/p + 1/q = 1$.

## Первое свойство

$$f_\gamma(x) = \mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] \ge \mathbb{E}_{\tilde{e}}\left[f(x) + \langle\nabla f(x), \gamma\tilde{e}\rangle)\right] = \mathbb{E}_{\tilde{e}}\left[f(x)\right] = f(x).$$

$$|f_\gamma(x) - f(x)| = \left|\mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] - f(x)\right| \le \mathbb{E}_{\tilde{e}}\left[|f(x + \gamma\tilde{e}) - f(x)|\right]$$
$$\le \gamma M_2 \mathbb{E}_{\tilde{e}}\left[\|\tilde{e}\|_2\right] \le \frac{2}{\sqrt{d}}\gamma M_2.$$

# Свойства функции $f_\gamma(x)$

## Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. $f_\gamma$ — $M$-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. $f_\gamma$ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где $q$ такой, что $1/p + 1/q = 1$.

## Лемма 1 из [2] для первого свойства

Пусть $q \in [1, \infty)$ и пусть $v \in RB_1^d(1)$. Тогда

$$\mathbb{E}[\|v\|_q] \leq \frac{qd^{\frac{1}{q}}}{d+1}$$

# Свойства функции $f_\gamma(x)$

## Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \le f_\gamma(x) \le f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. $f_\gamma$ — $M$-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \le M\|y - x\|_p,$$
3. $f_\gamma$ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \le L_{f_\gamma}\|y - x\|_p.$$

где $q$ такой, что $1/p + 1/q = 1$.

## Второе свойство

$$|f_\gamma(y) - f_\gamma(x)| = |\mathbb{E}_{\tilde{e}}\left[f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})\right]| \le \mathbb{E}_{\tilde{e}}\left[|f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})|\right] \le M\|y - x\|_p$$

## Доказательство третьего свойства

Применяя Лемму 8 из [3] имеем

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q = \left\| \int_{Q_\gamma} \nabla f(z)\mu(z-y)dz - \int_{Q_\gamma} \nabla f(z)\mu(z-x)dz \right\|_q \leq$$

$$\leq M \underbrace{\int_{Q_\gamma} |\mu(z-y) - \mu(z-x)|dz}_{I_1} \leq \ldots \leq \frac{dM}{2\gamma}\|y-x\|_p,$$

где $\mu(x) = \begin{cases} \frac{1}{V(B_1^d(\gamma))}, & x \in B_1^d(\gamma) \\ 0, & \text{иначе} \end{cases}.$

# Основная идея

**Негладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

$\longrightarrow$

**Гладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

**Связь между задачами**

Если имеем $\frac{\varepsilon}{2}$-точность для функции $f_\gamma(x)$, то имеем $\varepsilon$-точность для функции $f(x)$:

$$f(x^{N+1}) - f(x_*) \leq f(x^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x^{N+1}) - f_\gamma(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

# Рандомизация с двухточечной обратной связью

## Следствие из формулы Стокса [2]

$$\nabla f_\gamma(x) = \mathbb{E}_{\tilde{e}}[\nabla f(x+\gamma\tilde{e})] = \frac{\mathsf{Vol}_{d-1}(\partial D)}{\mathsf{Vol}_d(D)} \cdot \mathbb{E}_e[f(x+\gamma e)n(e)]$$

# Рандомизация с двухточечной обратной связью

## Следствие из формулы Стокса [2]

$$\nabla f_\gamma(x) = \mathbb{E}_{\tilde{e}}[\nabla f(x + \gamma\tilde{e})] = \frac{\mathrm{Vol}_{d-1}(\partial D)}{\mathrm{Vol}_d(D)} \cdot \mathbb{E}_e[f(x+\gamma e)n(e)]$$

## Рандомизация с двухточечной обратной связью

### $l_1$-рандомизация

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{2\gamma}(f_\delta(x+\gamma e,\xi) - f_\delta(x-\gamma e,\xi))\mathrm{sign}(e), \quad (e \in RS_1^d(1))$$

### $l_2$-рандомизация

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{2\gamma}(f_\delta(x+\gamma e,\xi) - f_\delta(x-\gamma e,\xi))e, \quad (e \in RS_2^d(1))$$

# Рандомизация с двухточечной обратной связью

## Свойства $\nabla f_\gamma$ для $l_1$-рандомизации

1. Если $\Delta = 0$, то оценки будут несмещенными
$$E_{e,\xi}\left[\nabla f_\gamma(x,\xi,e)\right] = \nabla f_\gamma(x)$$
2. ("смещение") при $\Delta > 0$
$$E_e\langle[\nabla f_\gamma(x,\xi,e)] - \nabla f_\gamma(x), r\rangle \lesssim \frac{d\Delta R}{\gamma}, \quad \forall r : \|r\|_2 \le R.$$
3. (оценка второго момента)
$$E_e\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \le \kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{12(1+\sqrt{2})^2\gamma^2}\right),$$

где $1/p + 1/q = 1$ и $\kappa(p,d) = 48(1+\sqrt{2})^2 d^{2-\frac{2}{p}}$.

# Рандомизация с двухточечной обратной связью

## Свойства $\nabla f_\gamma$ для $l_2$-рандомизации

1. Если $\Delta = 0$, то оценки будут несмещенными

$$E_{e,\xi}\left[\nabla f_\gamma(x,\xi,e)\right] = \nabla f_\gamma(x)$$

2. ("смещение") при $\Delta > 0$

$$E_e\langle[\nabla f_\gamma(x,\xi,e)] - \nabla f_\gamma(x), r\rangle \lesssim \frac{\sqrt{d}\Delta R}{\gamma}, \quad \forall r : \|r\|_2 \le R.$$

3. (оценка второго момента)

$$E_e\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \le \kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{\sqrt{2}\gamma^2}\right),$$

где $1/p + 1/q = 1$ и $\kappa(p,d) = \sqrt{2}\min\{q, \ln d\}\, d^{2-\frac{2}{p}}$.

# Рандомизация с одноточечной обратной связью

## $l_1$-рандомизация

- Аппроксимация градиента

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{\gamma} f_\delta(x+\gamma e,\xi)\,\mathsf{sign}(e), \quad (e \in RS_1^d(1))$$

- Оценка второго момента

$$E_e\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \le d^{3-\frac{2}{p}}\left(\frac{G^2}{\gamma^2}+\frac{\Delta^2}{\gamma^2}\right),$$

## $l_2$-рандомизация

- Аппроксимация градиента

$$\nabla f_\gamma(x,\xi,e) = \frac{d}{\gamma} f_\delta(x+\gamma e, \xi) e, \quad (e \in RS_2^d(1))$$

- Оценка второго момента

$$E_e\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \leq \min\{q, \ln d\}\, d^{3-\frac{2}{p}} \left(\frac{G^2}{\gamma^2} + \frac{\Delta^2}{\gamma^2}\right),$$

# Основная идея

**Негладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

$\longrightarrow$

**Гладкая задача**

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

**Связь между задачами**

Если имеем $\frac{\varepsilon}{2}$-точность для функции $f_\gamma(x)$, то имеем $\varepsilon$-точность для функции $f(x)$:

$$f(x^{N+1}) - f(x_*) \leq f(x^{N+1}) - f(x_*(\gamma)) \leq f_\gamma(x^{N+1}) - f_\gamma(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

**Параметры**

$$\boxed{\gamma = \frac{\varepsilon}{2M_2}} \qquad \boxed{L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma} = \frac{2\sqrt{d}MM_2}{\varepsilon}} \qquad \boxed{\sigma^2 \leq 2\sqrt{2}\min\{q, \ln d\}\, d^{2-\frac{2}{p}} M_2^2}$$

# Федеративное обучение

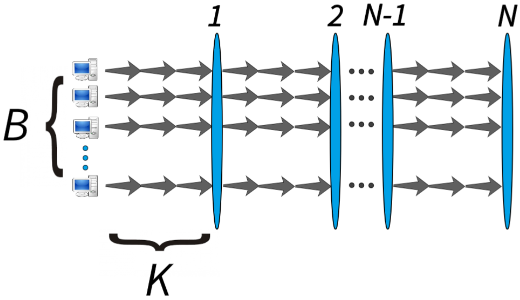



Рис.: Архитектура федеративного обучения

# Федеративное обучение

| Algorithm | $\mathbb{E}\left[f(\cdot)\right] - f^* \lesssim \ldots$ | Reference |
|---|---|---|
| Mb-Ac-SGD | $\frac{LR^2}{N^2} + \frac{\sigma R}{\sqrt{BNK}}$ | (Woodworth et al., 2021) [4] |
| SM-Ac-SGD | $\frac{LR^2}{N^2K^2} + \frac{\sigma R}{\sqrt{NK}}$ | (Woodworth et al., 2021) [4] |
| Local-AC-CA | $\frac{LR^2}{N^2K^2} + \frac{\sigma R}{\sqrt{BNK}}$ | (Woodworth et al., 2020) [5] |
| FedAc | $\frac{LR^2}{N^2K} + \frac{\sigma R}{\sqrt{BNK}} + \min\left\{\frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}R^{\frac{4}{3}}}{NK^{\frac{1}{3}}}, \frac{L^{\frac{1}{2}}\sigma^{\frac{1}{2}}R^{\frac{3}{2}}}{NK^{\frac{1}{4}}}\right\}$ | (Yuan, Ma, 2020) [6] |
| Mb-SMP | $\max\left\{\frac{LR^2}{N}, \frac{\sigma R}{\sqrt{BNK}}\right\}$ | |
| SM-SMP | $\max\left\{\frac{LR^2}{NK}, \frac{\sigma R}{\sqrt{NK}}\right\}$ | |

Таблица: **Summary of results on convergence rates.**
Notation: $R : \|x^0 - x_*\|$; $B$: number of computers; $K$: number of local update; $N$: number of communication rounds; $L$: smoothness.

**Theorem**

*Схема сглаживания, применяемая к негладкой задаче, обеспечивает сходимость Minibatch Accelerated SGD [4]. Другими словами, для достижения $\varepsilon$ точности решения негладкой задачи необходимо проделать $NK$ итераций с максимально допустимым уровнем шума $\Delta$ и общим числом вызова безградиентного оракула $T$ в соответствии с выбранным методом и схемой сглаживания:*

- $l_1$-рандомизация

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)dM_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \ (q = 2) \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1 \ (q = \infty). \end{cases}$$

# Главный результат

## Theorem

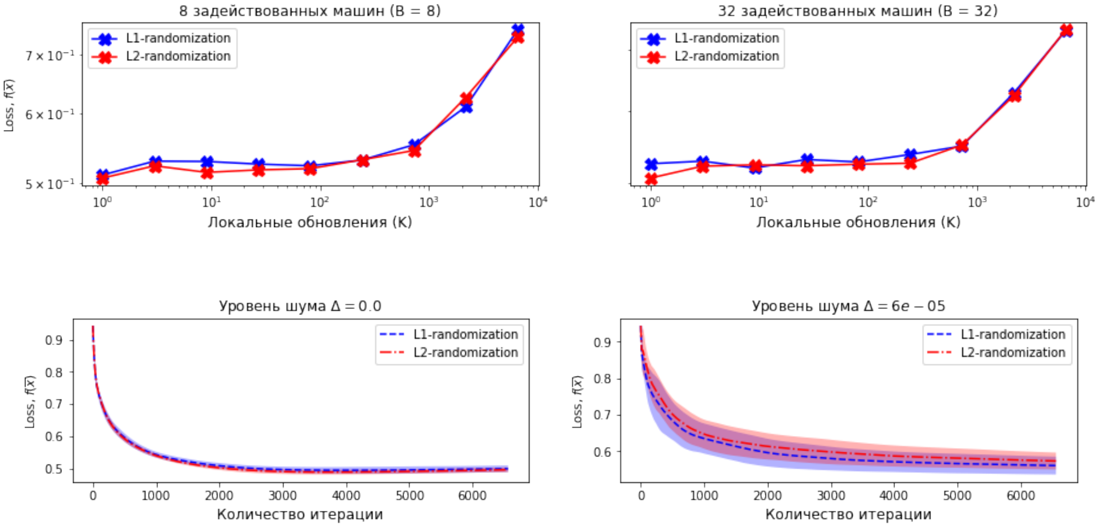
*Схема сглаживания, применяемая к негладкой задаче, обеспечивает сходимость Minibatch Accelerated SGD [4]. Другими словами, для достижения $\varepsilon$ точности решения негладкой задачи необходимо проделать $NK$ итераций с максимально допустимым уровнем шума $\Delta$ и общим числом вызова безградиентного оракула $T$ в соответствии с выбранным методом и схемой сглаживания:*

- $l_2$-рандомизация

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)dM_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2\ (q = 2) \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1\ (q = \infty). \end{cases}$$

# Эксперименты



# Reference


- *Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B.* The Power of First-Order Smooth Optimization for Black-Box Non-Smooth Problems // 2022. URL: https://arxiv.org/pdf/2201.12289.pdf.
- Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B Tsybakov. A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. *arXiv preprint arXiv:2205.13910*, 2022.
- Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.
- *Woodworth B., Bullins B., Shamir O., Srebro N.* The Min-Max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication. // Proceedings of Machine Learning Research. 2021. V. 134. P. 1–52.
- *Woodworth B., Patel K. K., Stich S. U., Dai Z., Bullins B., McMahan H. B., Shamir O., Srebro N.* Is Local SGD Better than Minibatch SGD? // Proceedings of the 37th