# Оптимизация-1

## «Сложность и эффективность»*
## на пути к идеалу

* Название монографии А.С. Немировского

Сириус, Дмитрий Пасечнюк, июль 2021

# Пасечнюк Дмитрий

http://dmivilensky.ru

- Работаю в оптимизации с 2017 года (с 9 класса, 4 года)
- Студент МФТИ
- Сотрудник «Лаборатории продвинутой комбинаторики и сетевых приложений» А.М. Райгородского
- Автор 16 статей и препринтов

# Организационное

- Планируется 4 лекции на обширный круг вопросов в рамках современной теории оптимизации, и ещё несколько лекций по сопредельным тематикам
- В конце каждой лекции анонсируются Related topics, из них путём голосования определяются темы дополнительных семинаров
- Прошу сделать чат в Telegram со мной (@dmivilensky) для промежуточных отчётов и вопросов (флудилка пусть будет отдельно 🙃)
- После нескольких лекций начнём кодить по делу
- Прошу добровольно избрать «ответственного за романтику» 💕 для фотоотчётов и day-ликов
- Увлечённый и неравнодушный труд будет учтён и запомнен!

# Оптимизация

## План первой лекции

- Непрерывная оптимизация (сложность и эффективность)
- Градиентный спуск (классы функций, градиентный метод, оценка скорости сходимости)
- Оптимальность (ускоренный градиентный метод, переходы между классами)

# Непрерывная оптимизация

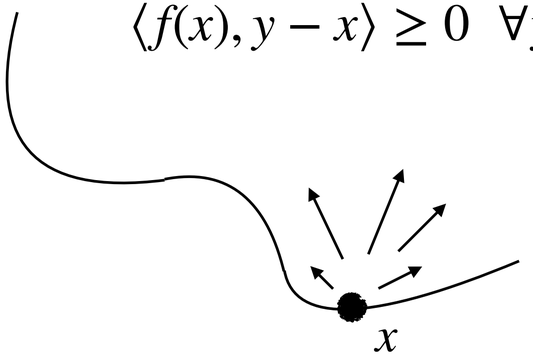
Если мы что-то и понимаем в оптимизации, то в непрерывной.
Но в ней мы – поэты!

**Принципиальная задача:** Найти глобально оптимальное решение (какой-бы то ни было задачи), располагая лишь локальной информацией.

# Оптимизационные задачи

## «Да нельзя здесь просто обнулить производную!»

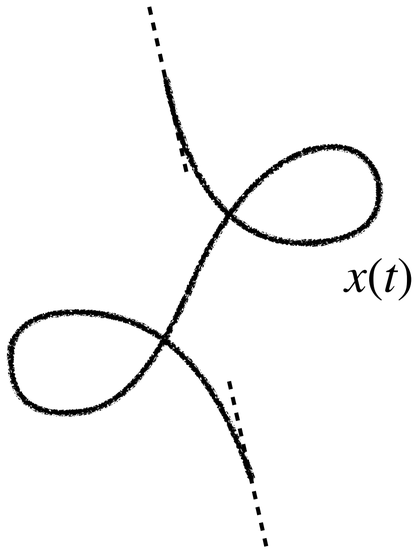
$$\langle f(x), y - x \rangle \geq 0 \quad \forall y$$




Вариационные неравенства

*Пример: теория игр*

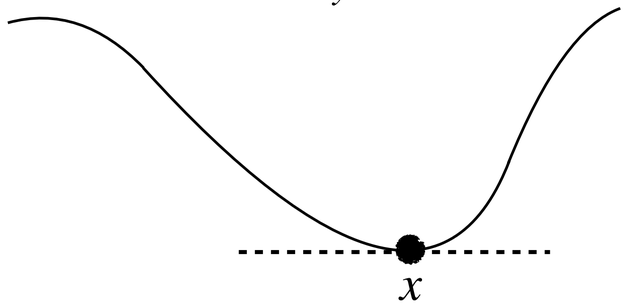
$$\int_{t_0}^{t_1} f(x(t), x'(t), t)dt = \min_{x(\cdot)}$$




Оптимальное управление

*Пример: ракеты*

$$f(x) = \min_{y} f(y)$$




Экстремальные задачи

*Пример: машинное обучение*

# Framework

## Как это стало математикой

Для начала, простая вещь: если мы располагаем только
локальной информацией, мы не сможем «тыкнуть
пальцем в небо» и угадать

Посему, «решить задачу» для нас не значит написать доказательство решения
(как в мат. логике) или прогнуть задачу силой личного убеждения (как в жизни)

Наш метод – постепенное улучшение
приблизительного решения

# Framework

## Что такое *эффективность*

Итак, мы имеем дело с итеративными алгоритмами вида

$$x_{t+1} = modify(x_t, t)$$

Формализуем понятие о локальной информации. Будем считать, что описание $f$ снабжено оракулом $Oracle$, возвращающим для точки $x$ всю информацию, которую мы можем использовать в $modify(x, \cdot)$

*Пример: оракул 1-го порядка* $Oracle(x) = (f(x), \nabla f(x))$

Эффективность алгоритма = число обращений к оракулу при фиксированных требованиях к решению и характеристиках задачи

# Framework

## Что такое *эффективность*

Описывать эффективность алгоритмов мы будем утверждениями вида

$$N = \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right),$$

где $N$ – число итераций метода,
$L, \mu$ – некоторые характеристики задачи,
$\varepsilon$ – необходимая точность решения

В большинстве случаев, под точностью мы имеем в виду условие $f(x_N) - \min\limits_x f(x) \leq \varepsilon$, характерное для экстремальных задач

$$f = \mathcal{O}(g) \Leftrightarrow \exists C > 0 : |f(x)| < C\,|g(x)|$$

# Framework

## Что такое *сложность*

Сложностью класса задач называется эффективность наилучшего метода, использующего оракул $Oracle$, на наихудшей для него задаче из класса

Такие оценки называются минимаксными. Их математическое обоснование весьма нетривиально, а сами их утверждения весьма глубоки

Вся эта теория была введена в книге:

Немировский А.С, Юдин Д.Б.
«Сложность задач и эффективность методов оптимизации»

# Градиентный спуск

## Простейший метод 1-го порядка

Быстро катится камень,
Скоро ляжет он у подножья.
Сколько ждёт его испытаний!

# Интуиция и метод

## Коши 1847 г. → ML 2021 г.

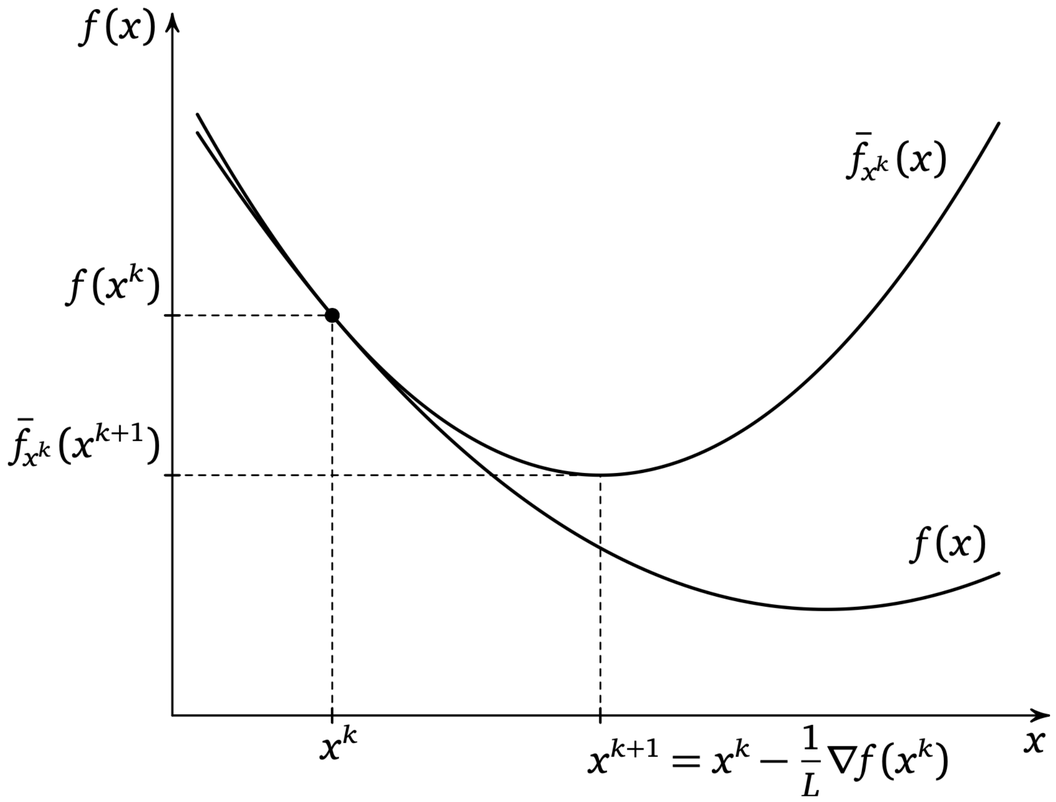
Проще не бывает: будем идти туда, где функция уменьшается быстрее всего

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

Квадратичная мажоранта, дискретизация градиентной динамики

Очень многое о методе написано в книге:

А.В. Гасников
«Современные численные методы оптимизации»



«The main lesson of the development of our field in the last few decades is that efficient optimization methods can be developed only by intelligently employing the structure of particular instances of problems».

Nesterov Yu.E., 2018

# Класс функций

## Или условия эффективности

Разумеется, не существует (и не может существовать, что доказуемо) универсально эффективного метода оптимизации. Однако можно рассматривать достаточно общий, но всё-таки ограниченный класс задач

Введение класса задач ещё даёт нам возможность говорить об оптимальных методах, выполняющих минимаксную оценку. Обычно рассматривают класс гладких выпуклых функций

Замечательно, что методы, эффективные в теории только на этом классе, часто оказываются практически эффективны и в более сложных задачах

# Класс функций

## Условие прогресса

Назовём функцию липшицево гладкой, если

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$$

Или, иначе говоря, $\lambda_{max}(\nabla^2 f(x)) \leq L$

Это условие полезно в смысле гарантии прогресса каждого шага используемого метода оптимизации. Действительно:

$$f(x_t) - f(x_{t+1}) \geq \eta \langle \nabla f(x_t), d \rangle - \eta^2 \frac{L}{2}\|d\|^2 \geq \frac{\langle \nabla f(x_t), d \rangle^2}{2L\|d\|^2}$$

# Класс функций

## Условие оптимальности

Назовём функцию выпуклой, если

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$$

И сильно выпуклой, если

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2$$

Это же условие определяет то, в коей степени наш прогресс связан с приближением к реальному минимуму (заметим, что без выпуклости не гарантирована даже унимодальность)

Надо сказать, выпуклость не единственное подходящее свойство. Есть ещё условие Поляка-Лоясиевича и Hölder Error Bound

# Скорость сходимости

## Теоретическая оценка для выпуклого случая

$$f(x_t) - f(x_{t+1}) \geq \frac{\langle \nabla f(x_t), \nabla f(x_t) \rangle^2}{2L\|\nabla f(x_t)\|^2} \geq \frac{\langle \nabla f(x_t), x_t - x^* \rangle^2}{2L\|x_t - x^*\|^2}$$

$$(f(x_t) - f^*) - (f(x_{t+1}) - f^*) \geq \frac{\langle \nabla f(x_t), x_t - x^* \rangle^2}{2L\|x_t - x^*\|^2} \geq \frac{(f(x_t) - f^*)^2}{2L\|x_0 - x^*\|^2}$$

$$f(x_t) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{t+4} \qquad N = \mathcal{O}\left(\frac{LR^2}{\varepsilon}\right)$$

В сильно выпуклом случае, подобными же рассуждениями, можно прийти к оценке:

$$N = \mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$$

# Оптимальность

## Методы, достигающие минимаксной оценки

Там, где кончаются эвристики…

# Сложность классов

## Таблица

Рассматривается класс методов вида

$$x_{t+1} \in x_0 + span\{\nabla f(x_0), \ldots, \nabla f(x_t)\}$$

| | **Липшицева непрерывность** | **Липшицева гладкость** |
|---|---|---|
| **Выпуклость** | $\mathcal{O}\left(\frac{M^2R^2}{\varepsilon^2}\right)$ | $\mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$ |
| **Сильная выпуклость** | $\mathcal{O}\left(\frac{M^2}{\mu\varepsilon}\right)$ | $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{\mu R^2}{\varepsilon}\right)$ |

# Инвариантность оптимальности

## Относительно переходов между классами

Выпуклый случай → сильно выпуклый случай

Перейдём к регуляризованной функции $f(x) \to f_\mu(x) = f(x) + \frac{\varepsilon}{2R^2}\|x - x_0\|^2$

$$f_\mu(x_N) - \min_x f_\mu(x) \le \frac{\varepsilon}{2} \Rightarrow f(x_N) - \min_x f(x) \le \varepsilon$$

Сильно выпуклый случай → выпуклый случай

Запустим метод на $N = 2\sqrt{L/\mu}$, получим $\|x_N - x^*\|^2 \le \frac{1}{2}\|x_0 - x^*\|^2$.

Рестартуем метод логарифмическое по точности число раз

# Инвариантность оптимальности

## Относительно переходов между классами

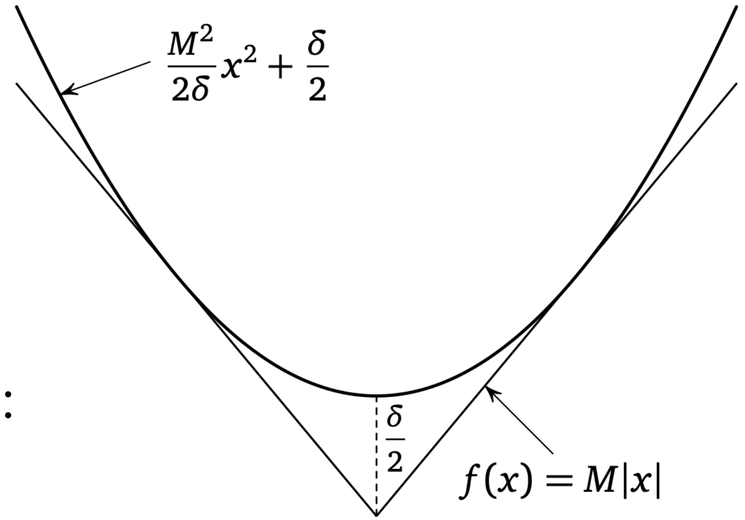
Гладкий случай → непрерывный случай

$$|f(y) - f(x)| \leq M\|y - x\|$$

$$\Rightarrow \forall \delta > 0 f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{M^2}{4\delta}\|y - x\|^2 + \delta$$

Для гладких функций с неточным оракулом имеем:

$$f(x_N) - \min_x f(x) \leq \frac{LR^2}{N^2} + N\delta \sim \frac{M^2R^2}{\delta N^2} + N\delta \to \min_{\delta > 0} = \frac{MR}{\sqrt{N}}$$

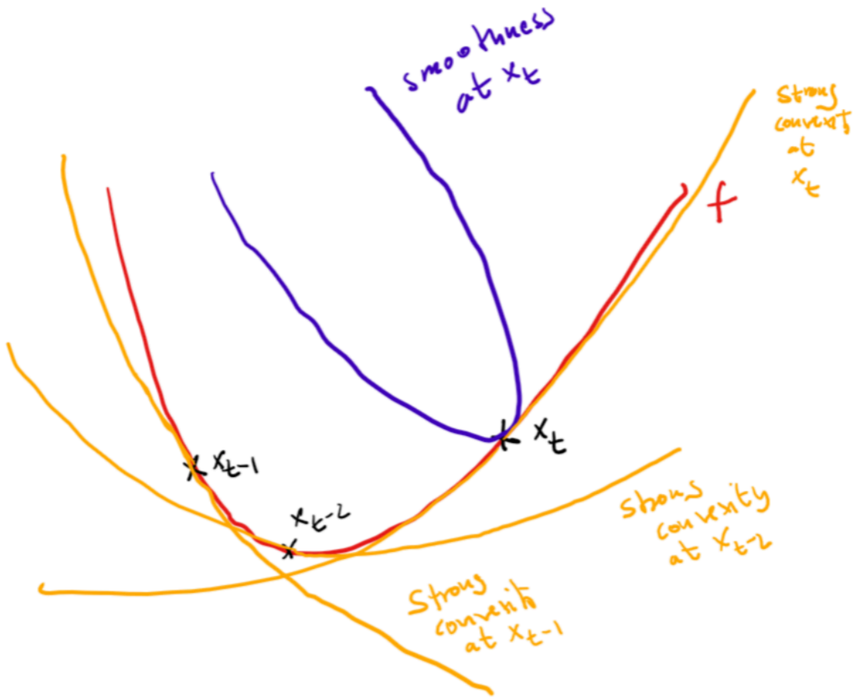


# Ускоренный градиентный метод

## Идея

Как можно видеть, градиентный метод не является оптимальным. Однако в нём мы используем информацию только с текущего шага, хотя у нас есть и прошлые

Конкретно, неплохо было бы использовать вместе все миноранты, которые у нас были



# Ускоренный градиентный метод

## Формальный вид

Выберем выпуклую комбинацию минорант и будем генерировать последовательность её минимумов

$$\frac{1}{A_t}\sum_{i=0}^{t} a_i[f(y_i) + \langle \nabla f(y_i), z - y_i\rangle + \frac{\mu}{2}\|y_i - z\|^2] \le f(z) \qquad w_t = \frac{1}{A_t}\sum_{i=0}^{t} a_i[y_i - \frac{1}{\mu}\nabla f(y_i)]$$

При сохранении вида основной последовательности и вытекающем из анализа виде смешанной последовательности

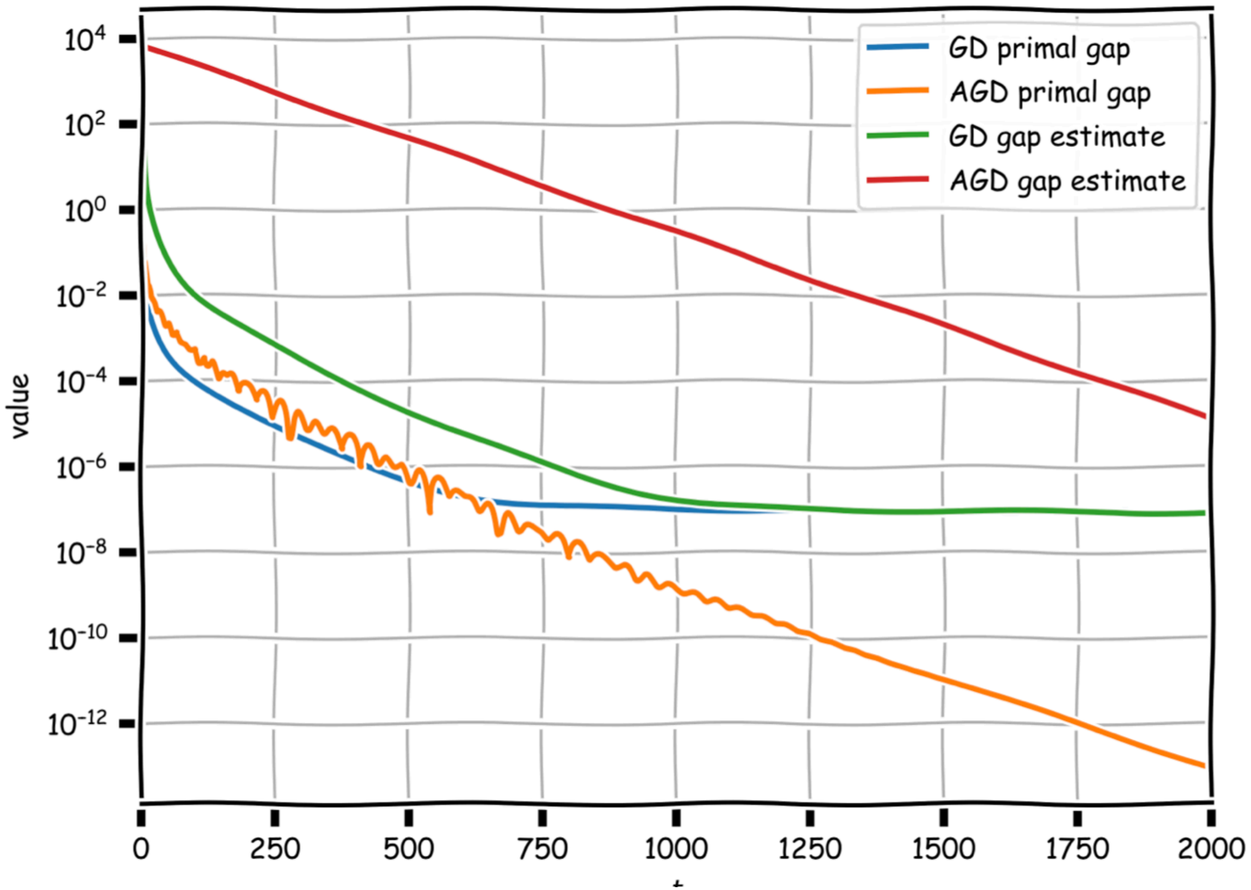
$$x_t = y_t - \frac{1}{L}\nabla f(y_t) \qquad y_t = \frac{A_t}{A_{t+1}}x_{t-1} + \frac{a_t}{A_{t+1}}w_{t-1} \qquad \frac{a_t}{A_t} = \sqrt{\frac{\mu}{L}}$$

В очень доступном виде на анализ можно посмотреть в блоге:

Sebastian Pokutta
Cheat Sheet: Acceleration from First Principles

# Ускоренный градиентный метод

## Практическая эффективность



«One of the most important theoretical results in convex optimization was the development of accelerated optimization methods»

Популярное в статьях утверждение

# Related topics

- Ускоренные методы с точностью до констант
- Ускоренные проксимальные оболочки
- Градиентный слайдинг

# Related bibliography


- d’Aspremont «Acceleration Methods»
- Parikh, Boyd «Proximal Algorithms»
- Gasnikov et al. «Accelerated meta-algorithm for convex optimization»

# Оптимизация-2

## От малого до big data

Сириус, Дмитрий Пасечнюк, июль 2021

# Оптимизация

## План второй лекции

- Пара слов о машинном обучении (SVM, нейронные сети)
- Стохастическая оптимизация (online и finite sum)
- Батчинг (оптимальность, параллелизм, федеративность)
- Адаптивность

«One of the big tensions in Statistics, which is a mystery to me, is really big data sets. You can try to estimate huge numbers of parameters with very few data points. Now I understand sometimes there's a story that seeks to justify that, but it makes me very, very nervous».

Persi Diaconis

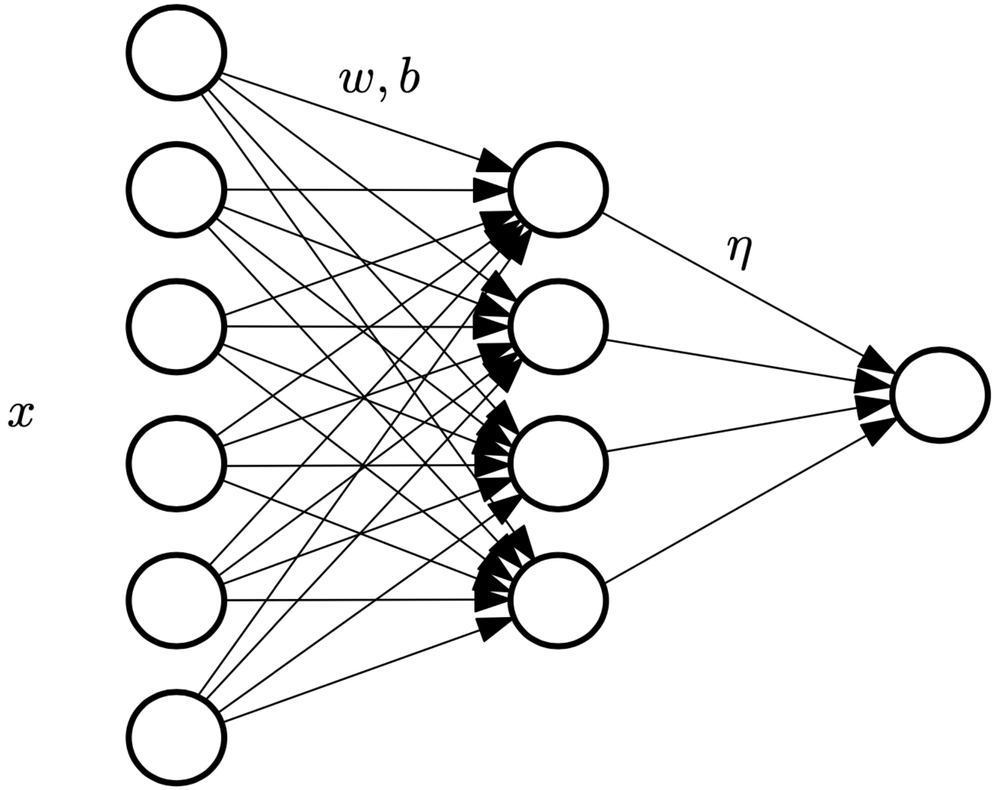
# Машинное обучение

## Гиперплоскости и нелинейные преобразования

Разделяющие котиков и собачек 😢

# Модели машинного обучения

## Нейронные сети



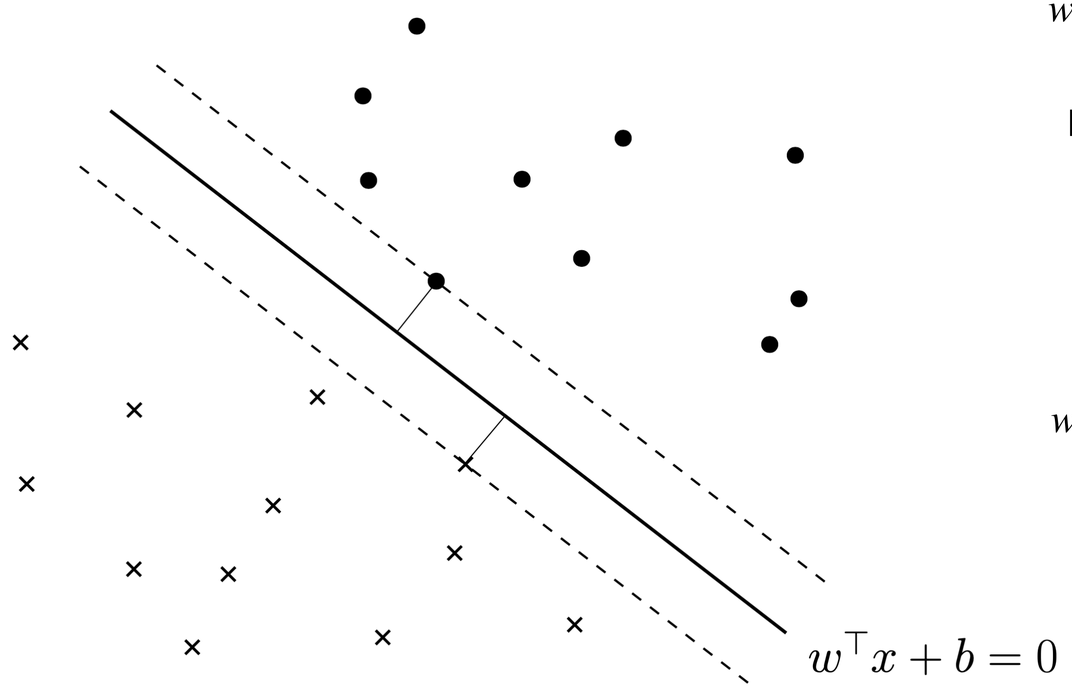

Один из двух магистральных подходов в машинном обучении – минимизация эмпирического риска

$$\min_{\theta \in \mathbb{R}^{d(m+1)}} \frac{1}{n} \sum_{i=1}^{n} \ell \left( y_i, \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x_i + b_j) \right) + \Omega(\eta, w, b)$$

В практике применения градиентных методов используется алгоритм обратного распространения ошибки и автоматическое дифференцирование

# Модели машинного обучения

## Support vector machines




$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

при $\quad \forall i \in \{1,\ldots,n\},\ \ y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0$

Что эквивалентно

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} (1 - y_i(w^\top x_i + b))_+ + \frac{\lambda}{2}\|w\|_2^2$$

Эти задачи и теория подробно разобраны в книге:

Francis Bach

«Learning theory from first principles»

# Риск

## Ожидаемый и эмпирический

Риск есть математическое ожидание ошибки по распределению данных

$$R(f_\theta) - R^* = \{R(f_\theta) - \inf_{\theta' \in \Theta} R(f_{\theta'})\} + \{\inf_{\theta' \in \Theta} R(f_{\theta'}) - R^*\}$$

ошибка оценки

ошибка аппроксимации

Эмпирический риск есть средняя ошибка по выборке

$$R(f_\theta) - R(f_{\theta'}) = \{R(f_\theta) - \hat{R}(f_\theta)\} + \{\hat{R}(f_\theta) - \hat{R}(f_{\theta'})\} + \{\hat{R}(f_{\theta'}) - R(f_{\theta'})\}$$

$$\leq 2 \sup_{\theta \in \Theta} |\hat{R}(f_\theta) - R(f_\theta)| + \text{ ошибка оптимизации}$$

# Стохастическая оптимизация

## Естественная случайность

Тише едешь – дальше будешь

# Стохастическая постановка

## Мотивация

Заметим, что в машинном обучении в постановке минимизации эмпирического риска мы имеем дело с функциями вида суммы

Суммы могут быть большими (сильно больше, чем размерность данных), слагаемые могут быть трудновычислимыми, в конце концов, они могут лежать на разных компьютерах

Вывод: нужны методы, работающие к отдельными слагаемыми.
«Разделяй и властвуй»

# Стохастическая постановка

## Формализм

$$f(x) = \sum_{i=1}^{n} f_i(x) \qquad \text{стохастический градиент} \quad \nabla f_i(x)$$

А положим теперь, что данных у нас потенциально бесконечно (имеем мы только их часть, хотя думать хотим обо всех). Пусть ещё наши объекты являются случайными объектами из некоторого неизвестного распределения $\xi \sim \mathscr{D}$

$$f(x) = \mathbb{E}[f(x, \xi)] \qquad \text{стохастический градиент} \quad \nabla f(x, \xi)$$

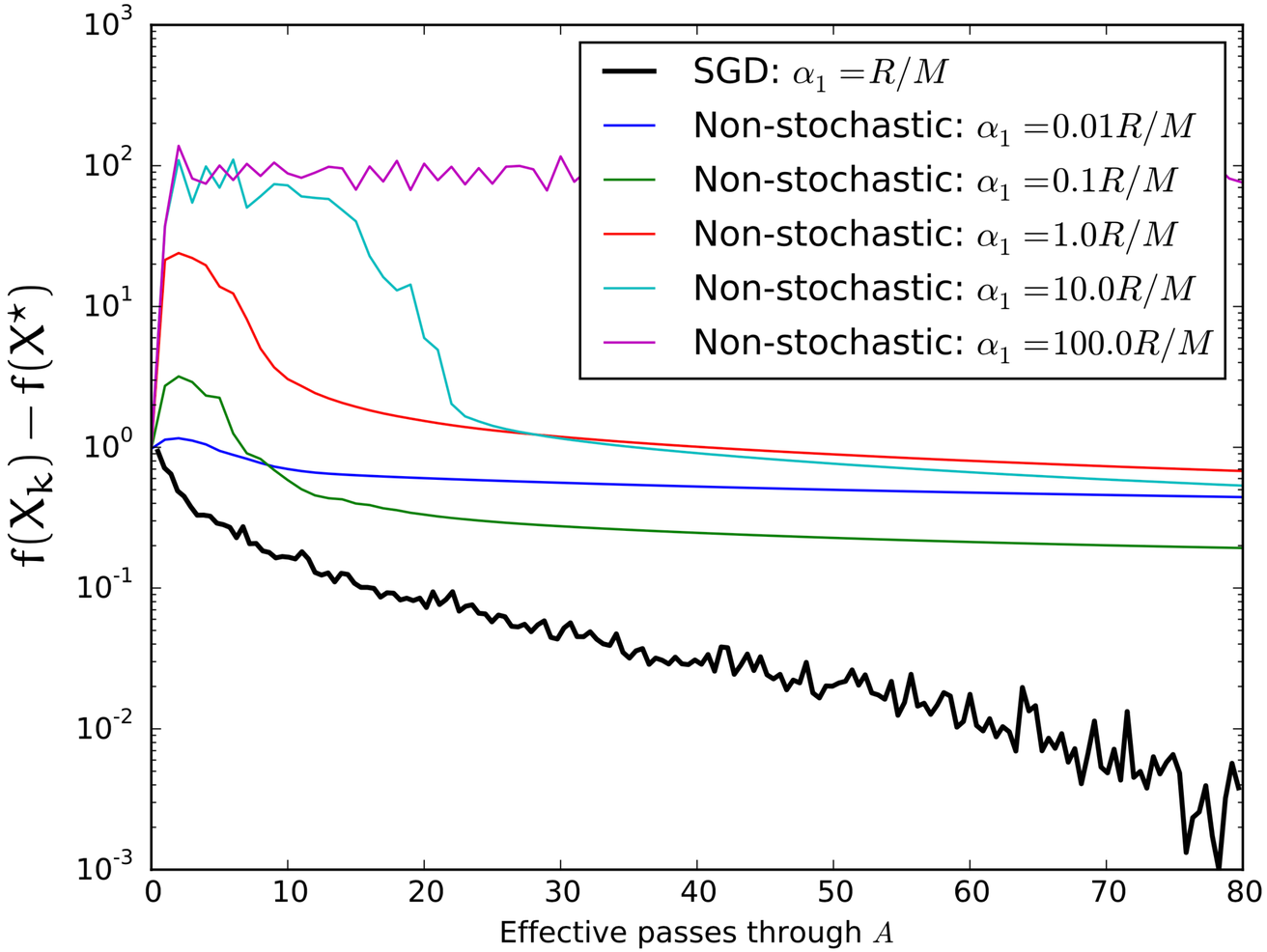
Более общо: $$\mathbb{E}[g(x, \xi)] = \int g(x, \xi)\, dp(\xi) \in \partial f(x)$$

Методы и теория для такого оракула обозреваются в книге:

John C. Duchi
«Introductory lectures on stochastic optimization»

# Стохастическая постановка

## Практическая эффективность



# Батчинг

Новые задачи, старая песня

# Сложность стохастических задач

Таблица

| | $\mathbb{E}[\|g(x,\xi)\|_2^2] \le M^2$ | Липшицева гладкость | Липшицева гладкость + $\mathbb{E}[\|g(x,\xi) - \nabla f(x)\|_2^2] \le D$ |
|---|---|---|---|
| Выпуклость | $\dfrac{M^2R^2}{\varepsilon^2}$ | $\sqrt{\dfrac{LR^2}{\varepsilon}}$ | $\max\left\{\sqrt{\dfrac{LR^2}{\varepsilon}}, \dfrac{DR^2}{\varepsilon^2}\right\}$ |
| Сильная выпуклость | $\dfrac{M^2}{\mu\varepsilon}$ | $\sqrt{\dfrac{L}{\mu}}\log\dfrac{\mu R^2}{\varepsilon}$ | $\max\left\{\sqrt{\dfrac{L}{\mu}}\log\dfrac{\mu R^2}{\varepsilon}, \dfrac{D}{\mu\varepsilon}\right\}$ |

# Сложность стохастических задач

## Достижимость нижних оценок

Просто пробатчим один оптимальный градиентный метод

$$g(x,\xi) = \frac{1}{b}\sum_{i=1}^{b} \nabla f(x,\xi_i) \qquad b = \mathcal{O}\left(\frac{D\alpha_{k+1}\log N}{(1+A_{k+1}\mu)\varepsilon}\right)$$

**Algorithm 1** Similar Triangles Method $\texttt{STM}(L,\mu,x^0)$, $Q=\mathbb{R}^n$

**Input:** $\tilde{x}^0 = z^0 = x^0$, number of iterations $N$, $\alpha_0 = A_0 = 0$, $L$, $\mu$

1: **for** $k = 0,\ldots,N$ **do**

2: Set $\alpha_{k+1} = \frac{1+A_k\mu}{2L} + \sqrt{\frac{1+A_k\mu}{4L^2} + \frac{A_k(1+A_k\mu)}{L}}$, $A_{k+1} = A_k + \alpha_{k+1}$

3: $\tilde{x}^{k+1} = (A_k x^k + \alpha_{k+1} z^k)/A_{k+1}$

4: $z^{k+1} = z^k - \frac{\alpha_{k+1}}{1+A_{k+1}\mu}\left(\nabla f(\tilde{x}^{k+1}) + \mu\left(z^k - \tilde{x}^{k+1}\right)\right)$

5: $x^{k+1} = (A_k x^k + \alpha_{k+1} z^{k+1})/A_{k+1}$

6: **end for**

**Output:** $x^N$

# Вычислительные преимущества

## Параллелизация

Вспомним оценки из таблицы:

$$\min\left\{\mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)+\mathcal{O}\left(\frac{DR^2}{\varepsilon^2}\right), \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{\mu R^2}{\varepsilon}\right)+\mathcal{O}\left(\frac{D}{\mu\varepsilon}\right)\right\}$$

Давайте распараллелим вычисления, связанные с батчингом, на несколько процессоров, а именно

$$\mathcal{O}\left(\frac{\frac{DR^2}{\varepsilon^2}}{\sqrt{\frac{LR^2}{\varepsilon}}}\right) \quad \text{или} \quad \mathcal{O}\left(\frac{\frac{D}{\mu\varepsilon}}{\sqrt{\frac{L}{\mu}}\log\frac{\mu R^2}{\varepsilon}}\right)$$

# Жизнь в пределе

## Связь online и finite sum постановок

Мы выписали некоторую малопрактичную постановку

$$\min_x f(x) = \mathbb{E}[f(x, \xi)]$$

Однако в реальности мы можем оперировать лишь задачами вида

$$\min_x \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i) + \frac{\varepsilon}{2R^2} \|x - x_0\|_2^2$$

А можем ли мы взять столько слагаемых,
чтобы решить бесконечную задачу?

Да! $$m = \min \left\{ \mathcal{O}\left( \frac{M^2 R^2}{\varepsilon^2} \right), \mathcal{O}\left( \frac{M^2}{\mu \varepsilon} \right) \right\}$$

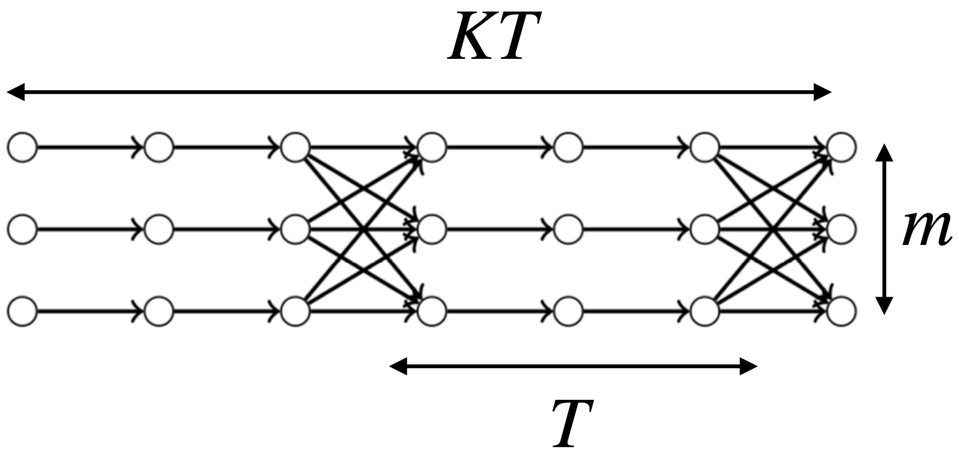
# Федеративное обучение

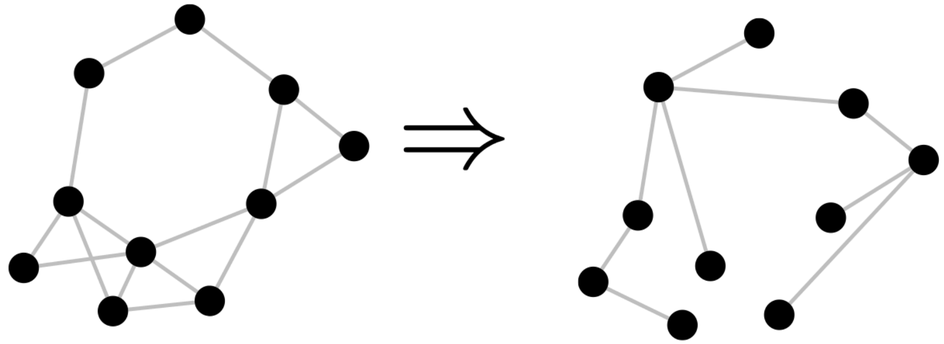
## Лучше смартфон в руках

Чем супер-компьютер в облаке

# Распределённость и персонализованность

Очень актуальные задачи



Строго говоря, результаты метода подобных треугольников и при допущении локальных шагов остаются оптимальными

Забавный, но поучительный комикс от Google:

https://federated.withgoogle.com

# Адаптивность

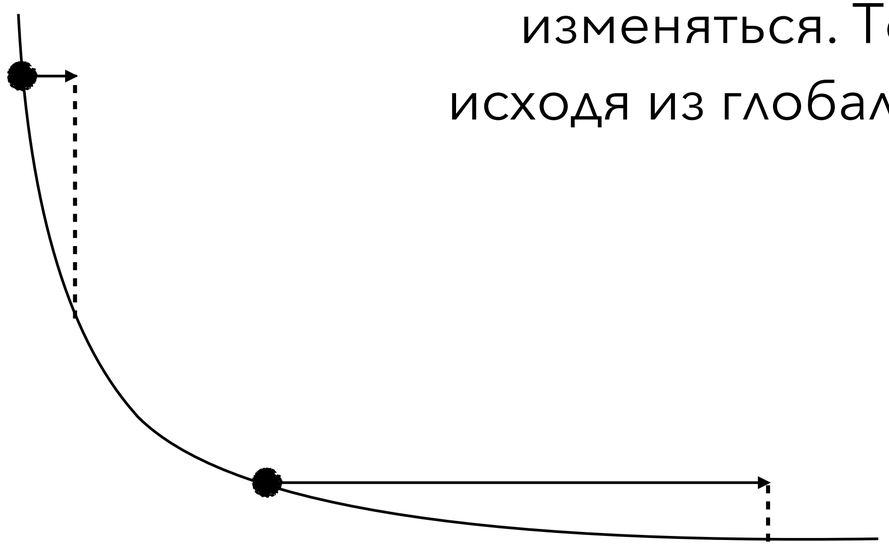
«Выживает не сильнейший, а реагирующий на изменения»*

… Там, где эвристики начинаются снова

* Чарльз Дарвин

# Адаптивность размера шага

## Идея



Характеристики функции могут локально изменяться. Тогда размер шага, выбранный исходя из глобальных свойств, может быть очень невыгоден

Что ж, придумаем стратегии изменения шага, или лучше уточнения локальных свойств функции

# Адаптивность размера шага

## Градиентный метод

Простейшая модификация: новая формула размера шага

$$\eta_t = \frac{\varepsilon}{L^2}, \quad \eta_t = \frac{R}{L\sqrt{N}}, \quad \eta_t = \frac{\varepsilon}{\|\nabla f(x_t)\|_2^2}, \quad \eta_t = \frac{R}{\|\nabla f(x_t)\|_2^2\sqrt{N}}$$

Более продвинутые схемы: самонастраивающиеся на гладкость

$$L_{t+1} = L_t/2, \quad x_{t+1} = x_t - \frac{1}{L_{t+1}}\nabla f(x_t)$$

Если $f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L_{t+1}}{2}\|x_{t+1} - x_t\|_2^2$

$$t = t + 1$$

Иначе

$$L_{t+1} = 2L_t$$

# Адаптивность размера шага

## Стохастические методы

Особенно подобные модификации важны в практических стохастических задачах, поскольку их свойства могут меняться разительно

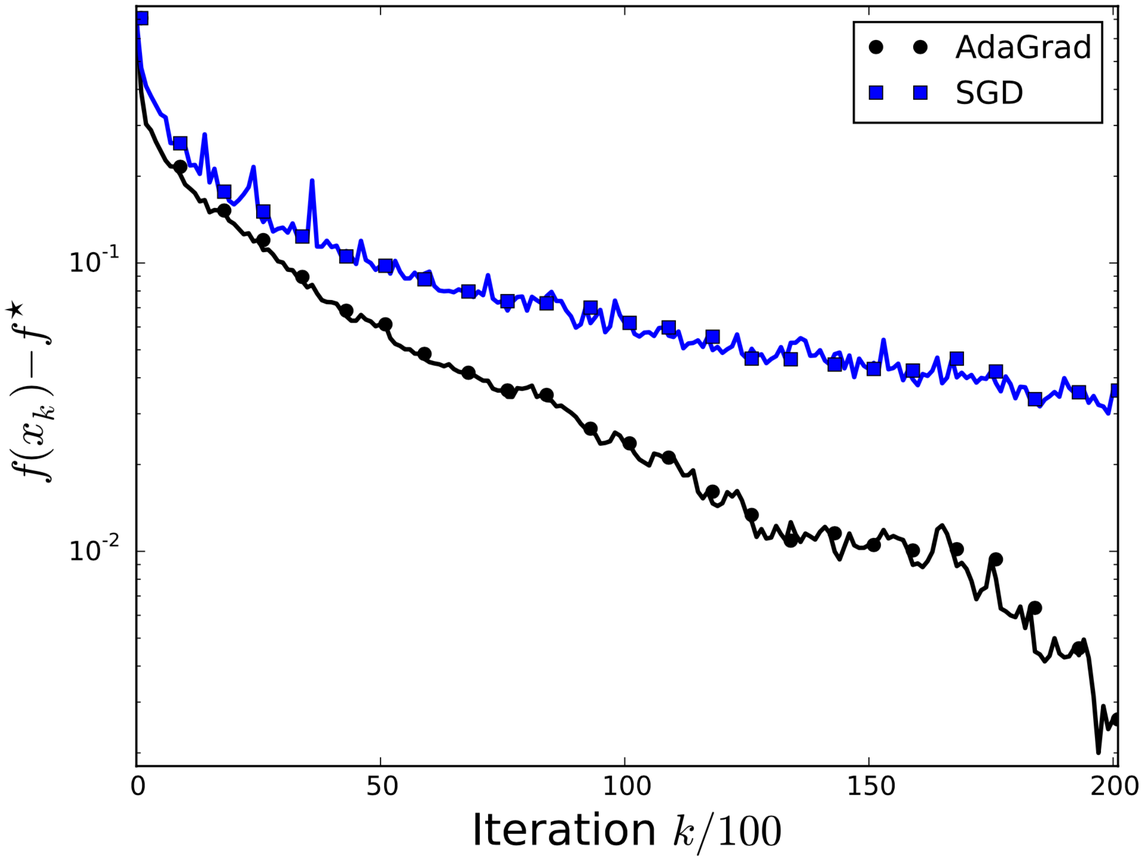
Методы переменной метрики: $$x_{t+1} = \arg\min_x \left\{ \langle g(x, \xi), x \rangle + \frac{1}{2}\|x - x_t\|^2_{H_t} \right\}$$

Где $H_t$ – положительно полуопределённые матрицы, $\|x\|_H = \sqrt{x^\top H^{-1} x}$ – расстояние Махаланобиса

Метод AdaGrad: $$x_{t+1} = x_t - \frac{\eta}{\sqrt{S_{t+1} + \epsilon}} g(x_t, \xi), \quad S_{t+1} = S_t + \langle g(x_t, \xi), g(x_t, \xi) \rangle$$

# Адаптивность размера шага

## Стохастические методы




Существует огромное разнообразие адаптивно-моментных стохастических методов. Все они очень актуальны в глубоком обучении

Почитать идейные объяснения и посмотреть на картинки можно в обзоре:

Sebastian Ruder
«An overview of gradient descent optimization algorithms»

# Related topics

- Редукция дисперсии
- Неточная модель функции
- Относительная гладкость
- Мета-обучение

# Related bibliography


- Yuxin Chen «Variance reduction for stochastic gradient methods»
- Воронцова, Хильдебранд, Гасников, Стонякин «Выпуклая оптимизация»
- Lu, Freund, Nesterov «Relatively-Smooth Convex Optimization by First-Order Methods, and Applications»

# Оптимизация-3

## Алиса и Боб в зазеркалье

Сириус, Дмитрий Пасечнюк, июль 2021

# Оптимизация

## План третьей лекции

- Линейное программирование (постановка, оптимальный транспорт, двойственность)
- Метод Лагранжа (условия Каруша–Куна–Таккера, Слейтера)
- Двойственность (зазор, прямо-двойственные методы, стохастический случай)

# Линейное программирование

От фронтов войны до границ политопов

# Постановка задачи

## И её история

$$\min_x \langle c, x \rangle \quad \text{при} \quad Ax \leq b, x \geq 0$$

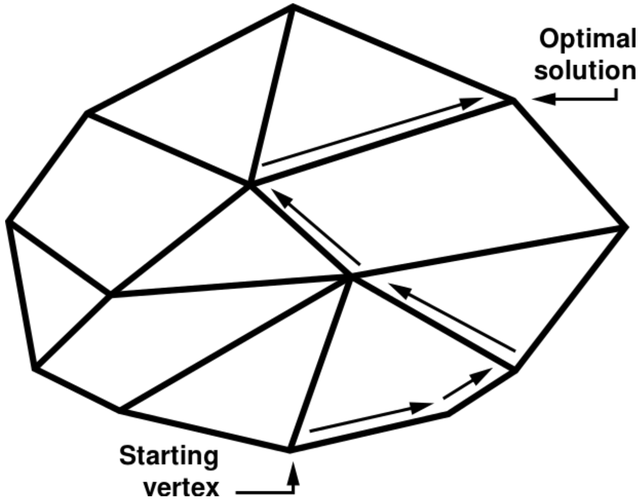
Активная работа над этой задачей была инициирована WWII, в виде задач о перемещении войск и продовольствия. Над задачей работали ведущие мировые математики:

Л.В. Канторович

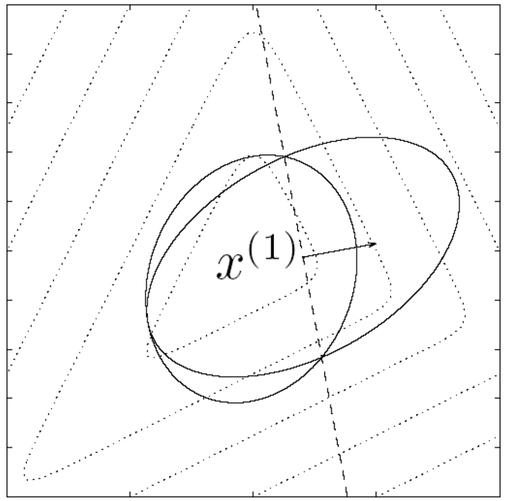
Джон фон Нейман

# Кратко о методах решения

Предлагаемые методы делились на базисные методы и методы внутренней точки, то есть точные полиномиальные или приблизительные и эффективные оптимизационные




Применение симплекс-метода для нахождения вершины




Применение метода эллипсоидов к linear feasibility problem

# Актуальность

## Релаксации комбинаторных задач

Одно из вечно востребованных приложений линейного программирования состоит в инструментах приближения задач на конечных пространствах: целочисленных задач, задач на графах и т.д.

Задача о минимальном рёберном покрытии:

$$\min \sum_{e \in E} x_e \quad \text{при} \quad \sum_{e \in incident(v)} x_e \geq 1,\ x_e \in \{0,1\}$$

её релаксация:

$$\min \sum_{e \in E} x_e \quad \text{при} \quad \sum_{e \in incident(v)} x_e \geq 1,\ x_e \geq 0$$

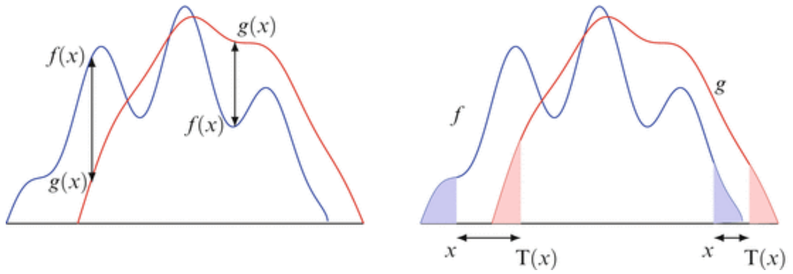
(поскольку при оптимума при $x_e > 1$ быть не может)

# Развитие

## Барицентры Вассерштейна

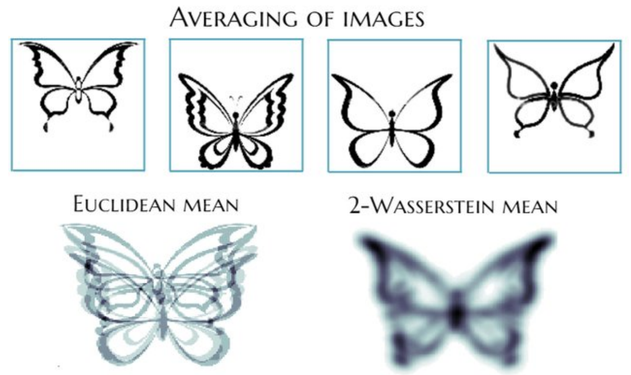
Развитием главной задачи ЛП стала общая задача ОТ Монжа–Канторовича

$$\min_x \left\{ \langle C, X \rangle + \gamma \sum_{i,j=1,1}^{n,n} X_{ij} \ln \frac{X_{ij}}{X^0_{ij}} \right\}$$




На них основаны методы нахождения барицентров Вассерштейна

$$W(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times X} d(x, y)\, d\gamma(x, y)$$




Благодаря энтропийной регуляризации стали доступны новые методы решения

Последние также имеют приложения в медицинской диагностике

# Двойственность

## «Зачем опять меняемся местами...»*

Сформулируем новую задачу ЛП, опираясь на нашу исходную:

$$\min_x \langle c, x \rangle \quad \text{при} \quad Ax \leq b, x \geq 0 \quad \longrightarrow \quad \max_y \langle b, y \rangle \quad \text{при} \quad A^\top y \geq c, y \geq 0$$

Математическим смыслом эту процедуру наделяют теоремы

Слабой двойственности $\min_x \langle c, x \rangle \geq \max_y \langle b, y \rangle$

и Сильной двойственности $\min_x \langle c, x \rangle = \max_y \langle b, y \rangle$

Кроме того, обе задачи (не)выполнимы одновременно

* И.А. Бродский

# Двойственность

## Особо педантичные читатели наверное заметили…

Основной смысл двойственности безусловно скрыт за завесами выпуклого анализа (теоремы отделимости и т.п.), но бывает приятно понимать интерпретацию

В нашем случае, $y$ есть суть оценка дефицитности товара и коэффициент пропорциональности в зависимости прибыли от величины ресурса $b$

Есть математические следствия, которые помогают понять фокус. Например, записывая совместно прямую и двойственную задачи, мы можем свести поиск оптимального решения к feasibility problem

Иными словами, ограничения и целевая функция имеют одну природу и могут рассматриваться единообразно, вопрос только в ролях

# Множители Лагранжа

## Двойственность по ограничениям

Как в ЛП, но не ЛП

# Двойственность

## Из ограничений в функцию

Пусть имеется задача с ограничениями:

$$\min_x f_0(x) \quad \text{при} \quad \begin{aligned} f_i(x) &\le 0, i \in \{1,\dots,m\} \\ h_i(x) &= 0, i \in \{1,\dots,p\} \end{aligned}$$

Определим Лагранжиан как

$$L(x,\lambda,\mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

А двойственную функцию и двойственную задачу соответственно как

$$g(\lambda,\mu) = \inf_x L(x,\lambda,\mu) \qquad \text{и} \qquad \max_{\lambda \ge 0,\mu} g(\lambda,\mu)$$

# Двойственность

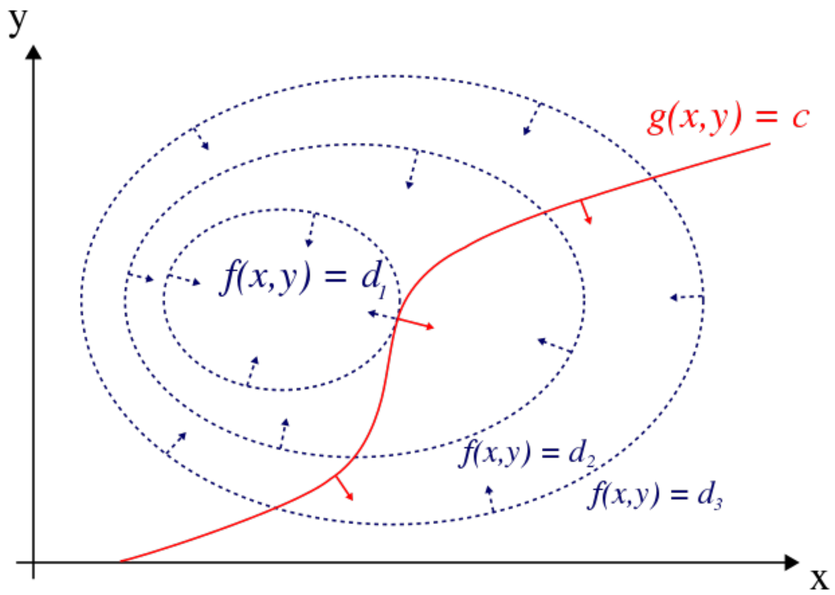
## Свойства

Для произвольных задач имеем слабую двойственность:

$g$ вогнута и $\forall \lambda \geq 0, \mu$ имеется нижняя оценка $g(\lambda, \mu) \leq \inf_{x,\ldots} f_0(x)$

Для выпуклых задач (+ условия ККТ) справедлива сильная двойственность:

$$\max_{\lambda \geq 0, \mu} g(\lambda, \mu) = \inf_{x,\ldots} f_0(x)$$




Для $m = 0, p = 1$ верно также правило множителей Лагранжа:

$$\exists! \lambda^* : \; Df_0(x^*) = \lambda^* Dh(x^*)$$

# Обобщение

## Условия Каруша–Куна–Таккера

Необходимое условие экстремума:

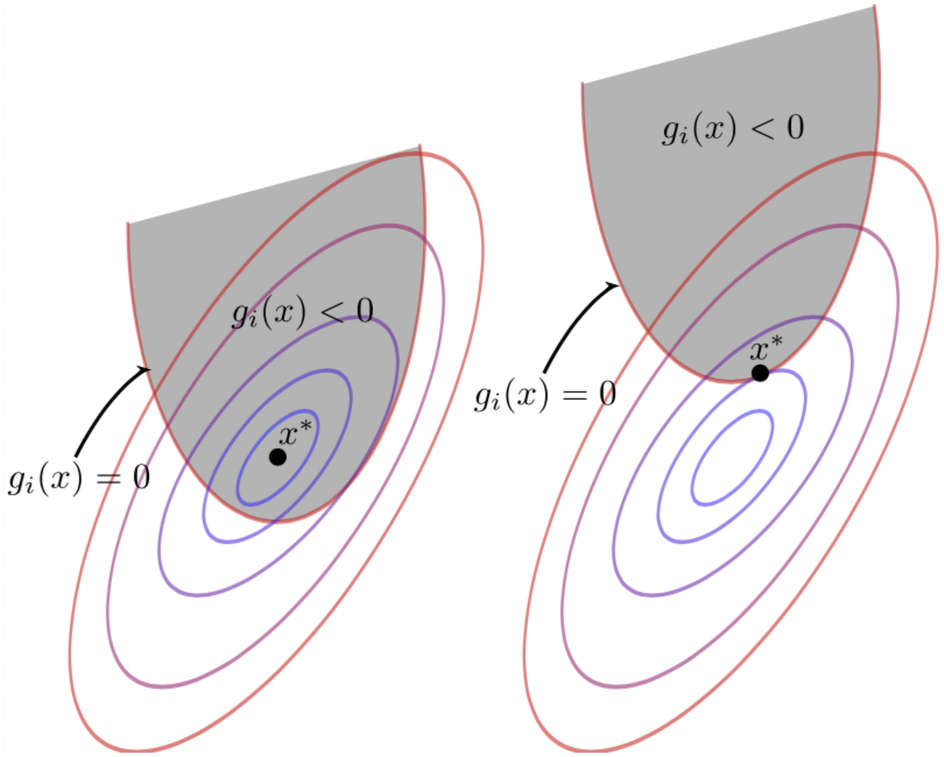
$$-\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x^*) + \sum_{i=1}^{p} \mu_i \nabla h_i(x^*) = 0$$

Прямая и двойственная выполнимости:

$$f_i(x^*) \le 0, \;\; h_i(x^*) = 0, \;\; \lambda_i \ge 0$$

Дополняющая нежёсткость:

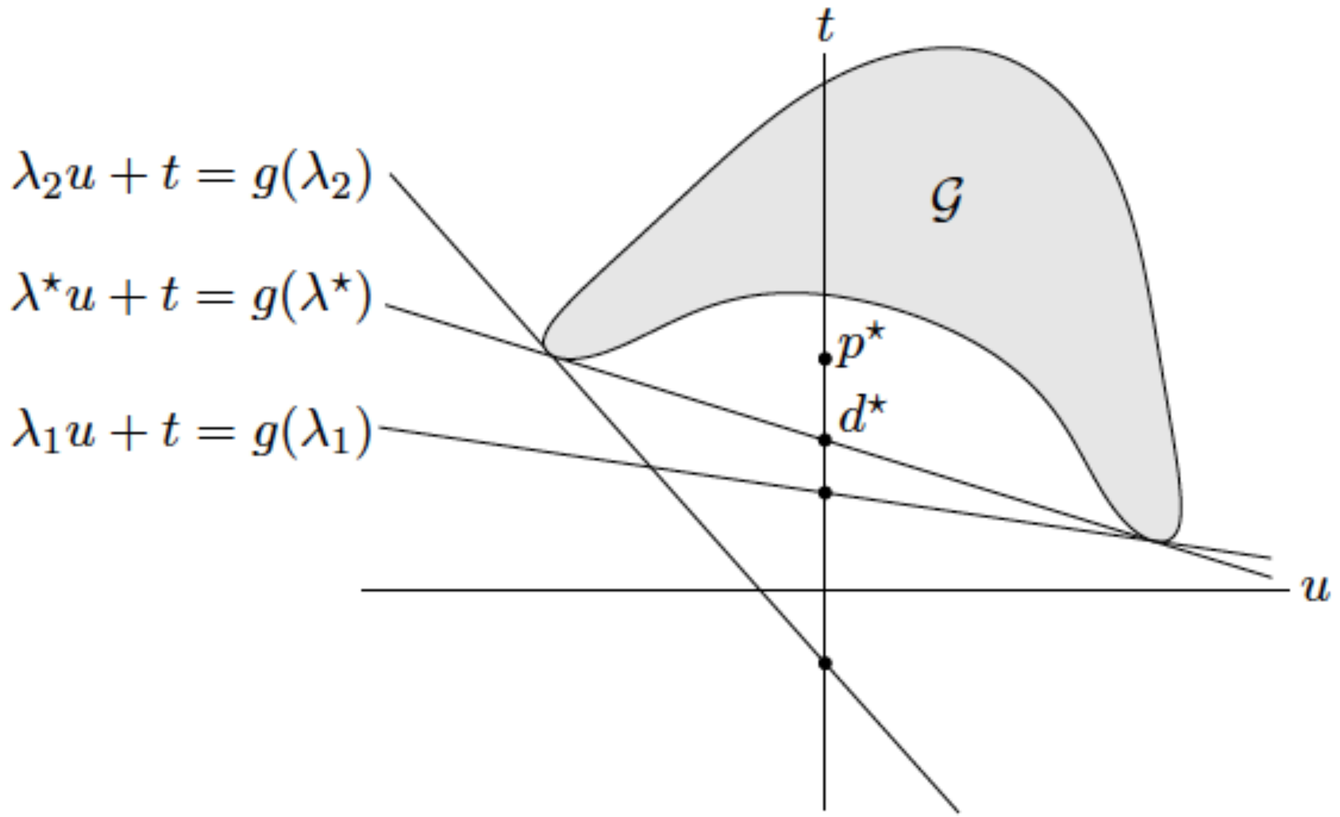
$$\mu_i f_i(x^*) = 0$$



# Двойственность

## Геометрическая интерпретация

Этот тип двойственности продолжает идею об оптимуме на границе



# Условие Слейтера

## Достаточное условие сильной двойственности

Если у feasible region существует внутренняя точка, то зазор двойственности в оптимуме равен нулю

Это также только одно из так называемых условий регулярности.
Однако его полезность есть ещё в том, что оно помогает ограничить задачу

$$g(0) \geq g(\lambda^*) \geq -\sum_{i=1}^{m} \lambda_i f_i(x) - f_0(x)$$

$$\|\lambda^*\| \leq \frac{1}{\min_{1,\dots,m} - f_i(x)} (f(x) - \min_{x,\dots} f_0(x))$$

# Двойственный оракул

## Двойственность в смысле Фенхеля

Чтобы что-то упростить, можно попробовать это сначала усложнить

# Преобразование Лежандра

## Переход в двойственное пространство

Производная и переменная меняются местами, значение функции заменяется на характеристику опорной гиперплоскости (касательного подпространства)

$$f^*(y) = \sup_{x \in X} \left\{ \langle y, x \rangle - f(x) \right\}$$

**Свойства:**

Для выпуклой полунепрерывной $f^{**}(x) \equiv f(x)$

Сохраняется выпуклость

Неравенство Юнга–Фенхеля $f(x) + f^*(y) \geq \langle y, x \rangle$

# Двойственность Фенхеля

Этот тип двойственности продолжает идею зеркальной замены переменных задачи

Теорема Фенхеля

Для выпуклых $\inf_x f(x) = \sup_y f^*(y)$

Существует идейная связь между концепциями Фенхеля и Лагранжа

$$g(0) = \inf_{x,\dots} L(x,0), \quad g^{**}(0) = \sup_y \{-L^*(0,y)\}$$

**Интересные свойства:**

Сильная выпуклость ↔ гладкость сопряжённой

Из этого свойства, в частности, вытекает техника двойственного сглаживания

# Вычислительная эффективность

## Задача OT

$$\varphi_l(y) = \varphi_{w^l}(y) = \min_{\substack{\sum_{j=1}^d \pi_{ij}=y_i,\ \sum_{i=1}^d \pi_{ij}=w_j^l \\ \pi_{ij}\geqslant 0,\ i,j=1,\ldots,d}} \left\{ \sum_{i,j=1}^d c_{ij}\pi_{ij} + \mu \sum_{i,j=1}^d \pi_{ij} \ln \pi_{ij} \right\} =$$

$$= \max_{x\in\mathbb{R}^d} \min_{\substack{\sum_{i=1}^d \pi_{ij}=w_j^l \\ \pi_{ij}\geqslant 0,\ i,j=1,\ldots,d}} \left\{ \sum_{i,j=1}^d c_{ij}\pi_{ij} + \sum_{i=1}^d x_i \cdot \left( y_i - \sum_{j=1}^d \pi_{ij} \right) + \mu \sum_{i,j=1}^d \pi_{ij} \ln \pi_{ij} \right\} =$$

$$= \max_{x\in\mathbb{R}^d} \Bigg\{ \langle y, x\rangle \underbrace{- \mu \sum_{j=1}^d w_j^l \ln \left( \frac{1}{w_j^l} \sum_{i=1}^d \exp\left( \frac{-c_{ij} + x_i}{\mu} \right) \right)}_{\varphi_l^*(x)=\varphi_{w^l}^*(x)} \Bigg\},$$

# Прямо-двойственность

## Градиентный метод

Использование градиентного метода по двойственным переменным с критерием останова по зазору двойственности позволяет найти пару прямо-двойственных решений, используя лишь двойственный оракул

$$f(x) = \max_y \{\langle x, b - Ay\rangle - \varphi(y)\} \to \min_x$$

$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t) \qquad y_{t+1} = \arg\max_y \{\dots\}$$

$$f(\bar{x}_N) + \varphi(\bar{y}_N) \le \varepsilon \to \varphi(\bar{y}_N) - \varphi(y^*) \le \varepsilon$$

Метод гарантированно остановится спустя $\frac{2LR^2}{\varepsilon}$ итераций

Фокус анализа в следующей формуле: $f(\bar{x}_N) + \varphi(\bar{y}_N) + 2R\|A\bar{y}_N - b\|_2 \le \varepsilon$

# Прямо-двойственность

## Распределённый градиентный метод, стохастический

| | $f_k$ is $\mu$-strongly convex, and $L$-smooth | $f_k$ is $\mu$-strongly convex |
|---|---|---|
| # of communication rounds | $\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$ | $O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$ |
| # of $\nabla\varphi_k(\lambda_k)$ oracle calls per node $k$ | $\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$ | $O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$ |
| Algorithm | ROGM-G or STM, $Q = \mathbb{R}^n$ | OGM-G or PDSTM |

| | $f_k$ is $\mu$-strongly convex, and $L$-smooth | $f_k$ is $\mu$-strongly convex |
|---|---|---|
| # of communication rounds | $\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$ | $O\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$ |
| # of $\nabla\varphi_k(\lambda_k, \xi_k)$ oracle calls per node $k$ | $\widetilde{O}\left(\max\left\{\frac{M^2\sigma_\phi^2}{\varepsilon^2}\chi, \sqrt{\frac{L}{\mu}}\chi\right\}\right)$ | $O\left(\max\left\{\frac{M^2\sigma_\phi^2}{\varepsilon^2}\chi, \sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right\}\right)$ |
| Algorithm | R-RRMA+AC-SA$^2$, $Q = \mathbb{R}^n$ | SPDSTM |

# Related topics

- Зеркальный спуск
- Дивергенция Брэгмана
- Барицентры Вассерштейна
- Выпуклый анализ

# Related bibliography


- Немировский, Юдин «Сложность задач и эффективность методов оптимизации»
- Статьи Д. Двинских (Darina Dvinskikh)
- Рокафеллар «Выпуклый анализ»

# Оптимизация-4

## Citius, Altius, Fortius!

Сириус, Дмитрий Пасечнюк, июль 2021

# Оптимизация

## План четвёртой лекции

- Суровая реальность (негладкость, неточность, мультимодальность, вид функции потерь)
- Методы с историей (квази-ньютон, сопряжённые градиенты)
- Немонотонные методы (Барзилаи–Борвейна, Б.Т. Поляка)
- Глобальные методы (симуляция отжига, биоинспирированные)
- Одномерный поиск (золотое сечение, Брента, квадратичные интерполяции)

# Реальные задачи

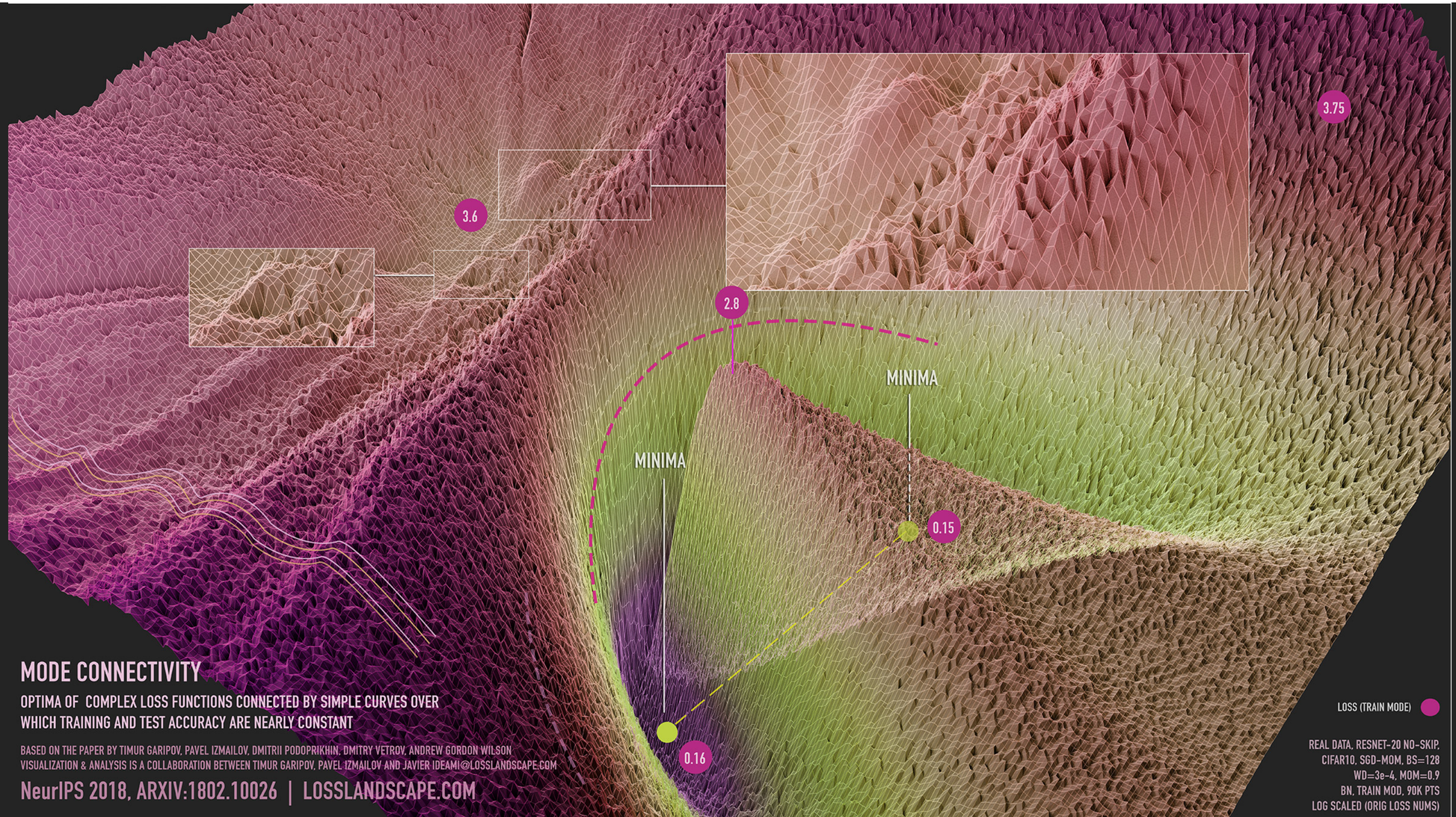
## Сложнее, чем нам хотелось бы

Но «непостижимая эффективность нашей математики»*…

* Из названия статьи Вигнера

# Поверхность целевой функции

## Глубокое обучение

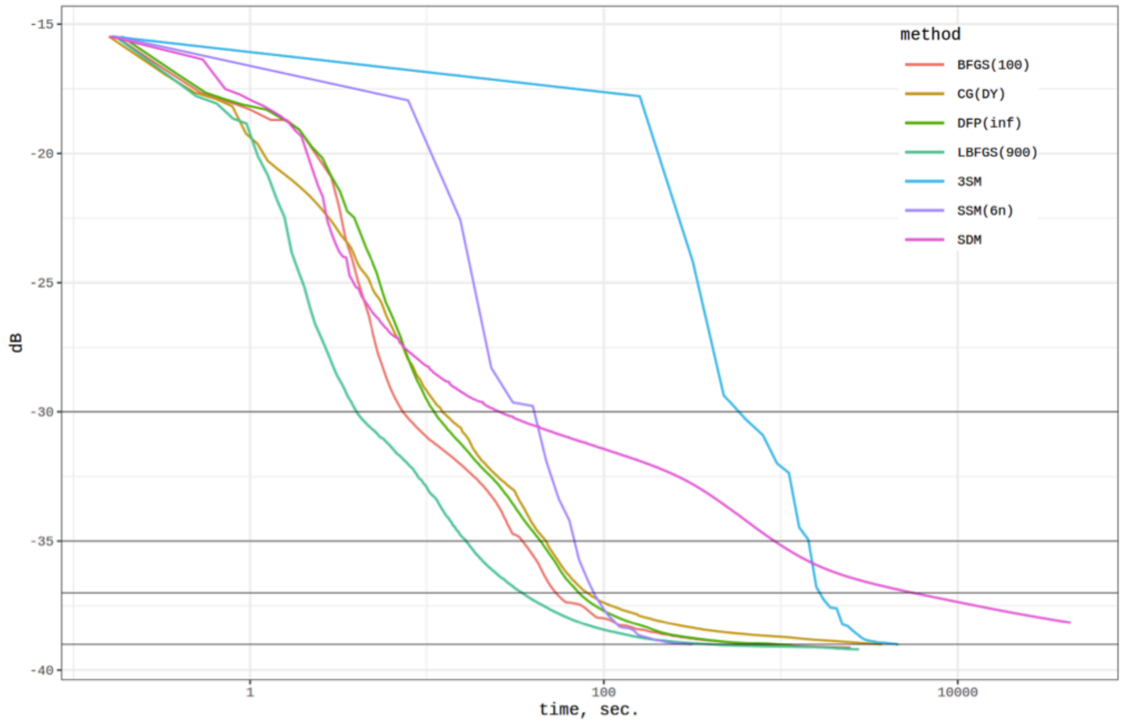


# Свойства целевых функций

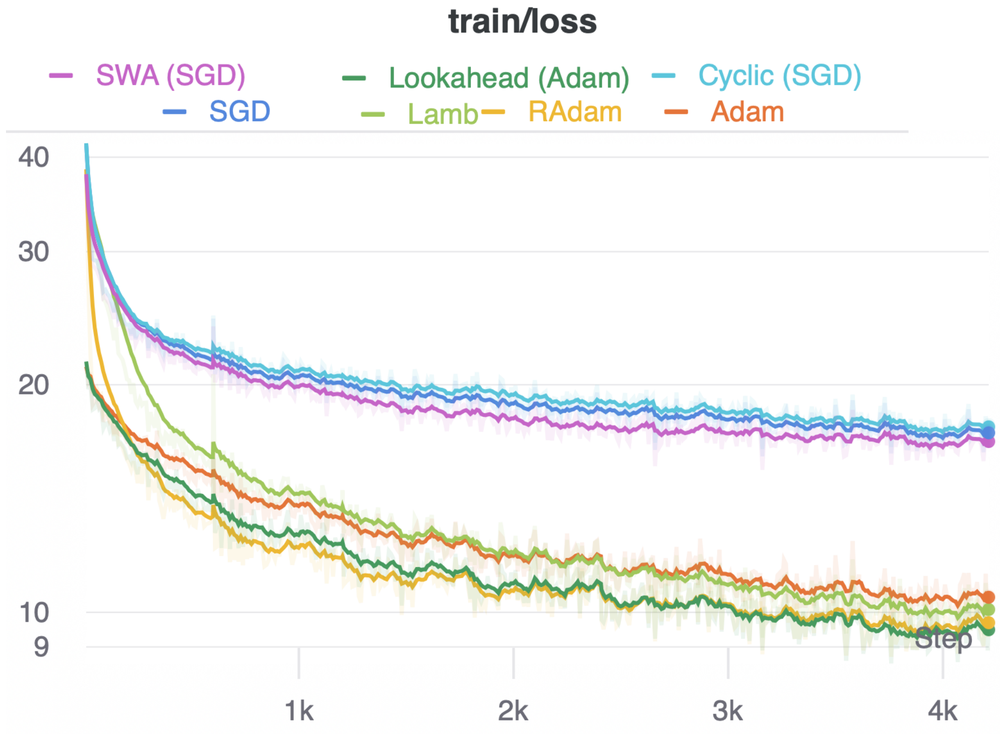
## Возникающих в реальных задачах

- Глобальная негладкость (не-непрерывность, невыпуклость, области регулярности). Локальные характеристики тем не менее могут оцениваться эмпирическими Монте–Карловскими процедурами. Адаптивность сохраняет эффективность
- Многоэкстремальность (локальные минимумы). Часто эта проблема нивелируется хорошими аппроксимирующими свойствами стохастических динамик
- Неточность вычислений (зашумлённость процедуры autograd, ограниченная точность типов). Следствия: ухудшение оценок до стохастических, область «застревания» методов в окрестности минимума

# Практика




Сходимость методов оптимизации
на задаче DPD




Сходимость методов оптимизации
на задаче NLP

# Практика

## Сумка с инструментами

Есть обширный список критериев, по которым делятся задачи:
математика это всё-таки экспериментальная наука
http://dmivilensky.ru/opt/Gornov.pdf

Но основной критерий – размерность задачи. При малой размерности допустимо применять градиентные и переборные методы, при большой – только стохастика

В этой лекции разговор будет о методах для среднего размера задач ( $\sim 10^5$).
Как минимум, они вдохновляют на нужные поиски

# Методы с историей

Скорость. Глубина. ~Устойчивость

# Метод Ньютона

Очень быстрый, но очень трудный

# Квази-ньютоновские методы

Limited-memory BFGS

# Метод сопряжённых градиентов

# Немонотонные методы

explore vs exploit

# Метод Б.Т. Поляка

## Глобально-ориентированная адаптивность

# Метод Барзилаи–Борвейна

Продолжая Ньютона

# Практика

## Попытка монотонизации

# Глобальные методы

## Глобальные эвристики

# Метод имитации отжига

# Практика

Потенциал, задачи, гиперпараметры

# Метод дифференциальной эволюции

# Одномерный поиск

«Коли будете иметь веру хоть с горчичное зерно...»*

* Евангелие от Матфея, 17:20

# Важность маломерных задач

Алгоритмы и приложения

# Метод дихотомии

# Метод золотого сечения

# Метод Брента

# Метод квадратичных интерполяций

# Мир оптимизации глубок и прекрасен!

Многие разные и сложные результаты можно получать простыми приёмами

Всё существует в тесной связи с дискретным анализом, выпуклым анализом, дифференциальной геометрией, теорией алгоритмов, теорией вероятностей, асимптотической статистикой и многими другими науками

И каждый может совершить как маленький шаг, так и решительный прорыв

Главное – любить науку

# Дерзайте!