

Стохастическая оптимизация. SGD

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

18 апреля 2025

Стохастическая оптимизация: постановка

- Рассматривали такую задачу:

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

Стохастическая оптимизация: постановка

- Рассматривали такую задачу:

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

- Теперь сформулируем задачу следующим образом:

$$\min_{x \in \mathbb{R}^d} [f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]] .$$

Стохастическая оптимизация: постановка

- Рассматривали такую задачу:

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

- Теперь сформулируем задачу следующим образом:

$$\min_{x \in \mathbb{R}^d} [f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [f(x, \xi)]] .$$

- Чтобы понять суть, рассмотрим пример из машинного обучения:

$$\min_{x \in \mathbb{R}^d} [f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [\ell(g(x, \xi_x), \xi_y)]] ,$$

где \mathcal{D} – распределение данных (природа данных), $\xi = (\xi_x, \xi_y)$ – элемент выборки: ξ_x – объект (картинка, текст) и ξ_y – метка (ответ), g – модель машинного обучения (линейная модель, нейросеть), принимает на вход объект и настраиваемые веса x , ℓ – функция потерь (штрафует модель за несовпадения с реальной меткой ξ_y).

Стохастическая оптимизация: постановка

- Мы хотим «подстроиться» под природу, и чтобы потери модели в среднем по всему распределению были наименьшими, т.е. модель наилучшим образом аппроксимировала зависимость ξ_y от ξ_x .

Стохастическая оптимизация: постановка

- Мы хотим «подстроиться» под природу, и чтобы потери модели в среднем по всему распределению были наименьшими, т.е. модель наилучшим образом аппроксимировала зависимость ξ_y от ξ_x .
- **Вопрос:** в чем проблема?

Стохастическая оптимизация: постановка

- Мы хотим «подстроиться» под природу, и чтобы потери модели в среднем по всему распределению были наименьшими, т.е. модель наилучшим образом аппроксимировала зависимость ξ_y от ξ_x .
- **Вопрос:** в чем проблема? Функцию f (а также градиенты и более старшие производные) не считаются, так как мы не знаем \mathcal{D} (это и суть аппроксимировать что-то сложное и неизвестное), да даже если и знаем, интеграл (мат. ожидание) часто не взять так просто.

Стохастическая оптимизация: постановка

- Мы хотим «подстроиться» под природу, и чтобы потери модели в среднем по всему распределению были наименьшими, т.е. модель наилучшим образом аппроксимировала зависимость ξ_y от ξ_x .
- **Вопрос:** в чем проблема? Функцию f (а также градиенты и более старшие производные) не считаются, так как мы не знаем \mathcal{D} (это и суть аппроксимировать что-то сложное и неизвестное), да даже если и знаем, интеграл (мат. ожидание) часто не взять так просто.
- Возникает потребность в методе, который может оперировать с $\nabla f(x, \xi)$ (градиентом по конкретному сэмплу из распределения данных). Т.е. хотим работать в онлайн режиме: поступают сэмплы, мы их обрабатываем (можем считать градиент).
- Естественное предположение, что данные поступают несмещенно:

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(x, \xi)] = \nabla f(x).$$

Стохастическая оптимизация: другая постановка

- Часто в машинном обучении мы стартуем не с «нуля» и дана обучающая выборка, тогда часто задачу обучения записывают в виде минимизации эмпирического риска:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^n [\ell(g(x, \xi_{x,i}), \xi_{y,i})] \right],$$

где $\{\xi_i\}_{i=1}^n$ – выборка из \mathcal{D} , g – модель, ℓ – функция. Такую постановку называют оффлайн (данные фиксированы, а не поступают в режиме реального времени).

Стохастическая оптимизация: другая постановка

- Часто в машинном обучении мы стартуем не с «нуля» и дана обучающая выборка, тогда часто задачу обучения записывают в виде минимизации эмпирического риска:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^n [\ell(g(x, \xi_{x,i}), \xi_{y,i})] \right],$$

где $\{\xi_i\}_{i=1}^n$ – выборка из \mathcal{D} , g – модель, ℓ – функция. Такую постановку называют оффлайн (данные фиксированы, а не поступают в режиме реального времени).

- Вопрос:** как связаны онлайн и оффлайн?

Стохастическая оптимизация: другая постановка

- Часто в машинном обучении мы стартуем не с «нуля» и дана обучающая выборка, тогда часто задачу обучения записывают в виде минимизации эмпирического риска:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^n [\ell(g(x, \xi_{x,i}), \xi_{y,i})] \right],$$

где $\{\xi_i\}_{i=1}^n$ – выборка из \mathcal{D} , g – модель, ℓ – функция. Такую постановку называют оффлайн (данные фиксированы, а не поступают в режиме реального времени).

- Вопрос:** как связаны онлайн и оффлайн? Оффлайн – это Монте-Карло аппроксимация исходного интеграла (мат.ожидания). Насэмплив много, аппроксимация через конечную сумму будет стремиться к реальному интегралу (при определенных предположениях).

Стохастическая оптимизация: другая постановка

- **Вопрос:** в оффлайн постановке уже можно считать градиент?

Стохастическая оптимизация: другая постановка

- **Вопрос:** в оффлайн постановке уже можно считать градиент? Да! Получается, что проблема решена, и лекция закончена. Но почему-то в машинном обучении часто не используют полные честные градиенты. **Вопрос:** почему?

Стохастическая оптимизация: другая постановка

- **Вопрос:** в оффлайн постановке уже можно считать градиент? Да! Получается, что проблема решена, и лекция закончена. Но почему-то в машинном обучении часто не используют полные честные градиенты. **Вопрос:** почему? Дорого/долго считать полный градиент.

Стохастическая оптимизация: другая постановка

- **Вопрос:** в оффлайн постановке уже можно считать градиент? Да! Получается, что проблема решена, и лекция закончена. Но почему-то в машинном обучении часто не используют полные честные градиенты. **Вопрос:** почему? Дорого/долго считать полный градиент.
- Поэтому вместо полного градиента вызывают градиент по случайному сэмплу:

$\nabla f(x, \xi_i)$, где ξ_i генерируется независимо и равномерно из $[n]$.

Стохастический градиентный спуск

- Простая идея – вновь модифицировать градиентный спуск и посмотреть, что будет.

Алгоритм 1 Стохастический градиентный спуск (SGD)

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Сгенерировать независимо ξ^k
- 3: Вычислить стохастический градиент $\nabla f(x^k, \xi^k)$
- 4: $x^{k+1} = x^k - \gamma_k \nabla f(x^k, \xi^k)$
- 5: **end for**

Выход: x^K

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

$$\mathbb{E} [\cdot \mid x^k] = \mathbb{E} [\cdot \mid \mathcal{F}_k],$$

где \mathcal{F}_k – σ -алгебра, порожденная $x^0, \xi^0, \dots, \xi^{k-1}$.

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

$$\mathbb{E} [\cdot \mid x^k] = \mathbb{E} [\cdot \mid \mathcal{F}_k],$$

где \mathcal{F}_k – σ -алгебра, порожденная $x^0, \xi^0, \dots, \xi^{k-1}$.

- Суть – «фиксируем» всю случайность, которая произошла до k итерации и ожидаем только по случайности, которая осталась размороженной.

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

$$\mathbb{E} [\cdot \mid x^k] = \mathbb{E} [\cdot \mid \mathcal{F}_k],$$

где \mathcal{F}_k – σ -алгебра, порожденная $x^0, \xi^0, \dots, \xi^{k-1}$.

- Суть – «фиксируем» всю случайность, которая произошла до k итерации и ожидаем только по случайности, которая осталась размороженной. **Вопрос:** такое математическое ожидание, что дает на выходе: что-то детерминистическое или случайное?

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

$$\mathbb{E} [\cdot \mid x^k] = \mathbb{E} [\cdot \mid \mathcal{F}_k],$$

где \mathcal{F}_k – σ -алгебра, порожденная $x^0, \xi^0, \dots, \xi^{k-1}$.

- Суть – «фиксируем» всю случайность, которая произошла до k итерации и ожидаем только по случайности, которая осталась размороженной. **Вопрос:** такое математическое ожидание, что дает на выходе: что-то детерминистическое или случайное? Случайное, зависящее от случайных величин $x^0, \xi^0, \dots, \xi^{k-1}$.

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

$$\mathbb{E} [\cdot \mid x^k] = \mathbb{E} [\cdot \mid \mathcal{F}_k],$$

где \mathcal{F}_k – σ -алгебра, порожденная $x^0, \xi^0, \dots, \xi^{k-1}$.

- Суть – «фиксируем» всю случайность, которая произошла до k итерации и ожидаем только по случайности, которая осталась размороженной. **Вопрос:** такое математическое ожидание, что дает на выходе: что-то детерминистическое или случайное? Случайное, зависящее от случайных величин $x^0, \xi^0, \dots, \xi^{k-1}$.
- Также понадобится закон полного математического ожидания (tower property):

$$\mathbb{E} [\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Сходимость: доказательство

- Будем доказывать в случае, когда f является L -гладкой и μ -сильно выпуклой.
- Введем также новое предположение, касающиеся стохастического градиента:

$$\mathbb{E}_{\xi}[\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}_{\xi} [\|\nabla f(x, \xi) - \nabla f(x)\|_2^2] \leq \sigma^2.$$

- Начинаем, как и раньше:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k, \xi^k)\|^2.$$

Сходимость: доказательство

- Будем доказывать в случае, когда f является L -гладкой и μ -сильно выпуклой.
- Введем также новое предположение, касающееся стохастического градиента:

$$\mathbb{E}_{\xi}[\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}_{\xi}[\|\nabla f(x, \xi) - \nabla f(x)\|_2^2] \leq \sigma^2.$$

- Начинаем, как и раньше:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k, \xi^k)\|^2.$$

- Берем условное мат.ожидание по случайности только на итерации k (важно, что x^k — это неслучайная величина относительно условного м.о.):

$$\begin{aligned} \mathbb{E}[\|x^{k+1} - x^*\|^2 \mid x^k] &= \|x^k - x^*\|^2 - 2\gamma_k \mathbb{E}[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k] \\ &\quad + \gamma_k^2 \mathbb{E}[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k]. \end{aligned}$$

Сходимость: доказательство

- Работаем с $\mathbb{E} [\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k]$:

$$\begin{aligned} \mathbb{E} [\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k] &= \langle \mathbb{E} [\nabla f(x^k, \xi^k) \mid x^k], x^k - x^* \rangle \\ &= \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \end{aligned}$$

Сходимость: доказательство

- Работаем с $\mathbb{E} [\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k]$:

$$\begin{aligned}\mathbb{E} \left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k \right] &= \langle \mathbb{E} \left[\nabla f(x^k, \xi^k) \mid x^k \right], x^k - x^* \rangle \\ &= \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle\end{aligned}$$

- Работаем с $\mathbb{E} [\|\nabla f(x^k, \xi^k)\|^2 \mid x^k]$:

$$\begin{aligned} \mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right] &= \mathbb{E} \left[\left\| \nabla f(x^k, \xi^k) - \nabla f(x^k) + \nabla f(x^k) \right\|^2 \mid x^k \right] \\ &= \mathbb{E} \left[\left\| \nabla f(x^k, \xi^k) - \nabla f(x^k) \right\|^2 \mid x^k \right] \\ &\quad + \mathbb{E} \left[\left\| \nabla f(x^k) \right\|^2 \mid x^k \right] \\ &\quad + 2\mathbb{E} \left[\langle \nabla f(x^k, \xi^k) - \nabla f(x^k), \nabla f(x^k) \rangle \mid x^k \right]. \end{aligned}$$

Сходимость: доказательство

- Продолжаем:

$$\begin{aligned}\mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right] &= \mathbb{E} \left[\left\| \nabla f(x^k, \xi^k) - \nabla f(x^k) \right\|^2 \mid x^k \right] + \left\| \nabla f(x^k) \right\|^2 \\ &\quad + 2 \langle \mathbb{E} \left[\nabla f(x^k, \xi^k) \mid x^k \right] - \nabla f(x^k), \nabla f(x^k) \rangle.\end{aligned}$$

- Предположение о стохастическом градиенте дает

$$\mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right] \leq \sigma^2 + \left\| \nabla f(x^k) \right\|^2.$$

Сходимость: доказательство

- Все, что получили:

$$\begin{aligned} \mathbb{E} \left[\|x^{k+1} - x^*\|^2 \mid x^k \right] &= \|x^k - x^*\|^2 - 2\gamma_k \mathbb{E} \left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k \right] \\ &\quad + \gamma_k^2 \mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right]. \end{aligned}$$

$$\mathbb{E} \left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k \right] = \langle \nabla f(x^k), x^k - x^* \rangle.$$

$$\mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right] \leq \sigma^2 + \|\nabla f(x^k)\|^2.$$

Сходимость: доказательство

- Все, что получили:

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] &= \|x^k - x^*\|^2 - 2\gamma_k \mathbb{E}\left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k\right] \\ &\quad + \gamma_k^2 \mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right].\end{aligned}$$

$$\mathbb{E} \left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k \right] = \langle \nabla f(x^k), x^k - x^* \rangle.$$

$$\mathbb{E} \left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k \right] \leq \sigma^2 + \|\nabla f(x^k)\|^2.$$

- Итого

$$\begin{aligned} \mathbb{E} \left[\|x^{k+1} - x^*\|^2 \mid x^k \right] &\leq \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\ &\quad + \gamma_k^2 \|\nabla f(x^k)\|^2 + \gamma_k^2 \sigma^2. \end{aligned}$$

Сходимость: доказательство

- Дальше уже привычно: L -гладкость и μ -сильная выпуклость

$$\begin{aligned}\mathbb{E} \left[\|x^{k+1} - x^*\|^2 \mid x^k \right] &\leq \|x^k - x^*\|^2 - 2\gamma_k \left(f(x^k) - f(x^*) + \frac{\mu}{2} \|x^k - x^*\|_2^2 \right) \\ &\quad + 2\gamma_k^2 L(f(x^k) - f(x^*)) + \gamma_k^2 \sigma^2 \\ &= (1 - \gamma_k \mu) \|x^k - x^*\|^2 + \gamma_k^2 \sigma^2 \\ &\quad - 2\gamma_k(1 - \gamma_k L)(f(x^k) - f(x^*)).\end{aligned}$$

- Если $\gamma_k \leq \frac{1}{L}$, то

$$\mathbb{E} \left[\|x^{k+1} - x^*\|^2 \mid x^k \right] \leq (1 - \gamma_k \mu) \|x^k - x^*\|^2 + \gamma_k^2 \sigma^2.$$

Сходимость SGD

Теорема сходимость SGD в случае ограниченной дисперсии

Пусть задача безусловной стохастической оптимизации с L -гладкой, μ -сильно выпуклой целевой функцией f решается с помощью SGD с $\gamma_k \leq \frac{1}{L}$ в условиях несмещенности и ограниченности дисперсии стохастического градиента. Тогда справедлива следующая оценка сходимости

$$\mathbb{E} \left[\|x^{k+1} - x^*\|^2 \right] \leq (1 - \gamma_k \mu) \mathbb{E} \left[\|x^k - x^*\|^2 \right] + \gamma_k^2 \sigma^2.$$

Сходимость SGD: анализ

- Постоянный шаг $\gamma_k \equiv \gamma$, тогда

$$\begin{aligned} \mathbb{E} \left[\|x^k - x^*\|^2 \right] &\leq (1 - \gamma\mu) \mathbb{E} \left[\|x^{k-1} - x^*\|^2 \right] + \gamma^2 \sigma^2 \\ &\leq (1 - \gamma\mu)^2 \mathbb{E} \left[\|x^{k-2} - x^*\|^2 \right] \\ &\quad + (1 - \gamma\mu) \gamma^2 \sigma^2 + \gamma^2 \sigma^2 \\ &\leq \dots \\ &\leq (1 - \gamma\mu)^k \mathbb{E} \left[\|x^0 - x^*\|^2 \right] + \gamma^2 \sigma^2 \sum_{i=0}^{k-1} (1 - \gamma\mu)^i. \end{aligned}$$

- **Вопрос:** как оценить второе слагаемое? Геометрическая прогрессия: $\sum_{i=0}^{k-1} (1 - \gamma\mu)^i \leq \sum_{i=0}^{+\infty} (1 - \gamma\mu)^i = \frac{1}{\gamma\mu}$:

$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma\mu)^k \mathbb{E} \left[\|x^0 - x^*\|^2 \right] + \frac{\gamma \sigma^2}{\mu}.$$

Сходимость SGD: анализ

- Результат вида:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq (1 - \gamma\mu)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\gamma\sigma^2}{\mu},$$

похож на то, что мы уже видели для градиентного спуска.

- Первый член – линейная сходимость к решению

Сходимость SGD: анализ

- Результат вида:

$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma\mu)^k \mathbb{E} \left[\|x^0 - x^*\|^2 \right] + \frac{\gamma\sigma^2}{\mu},$$

похож на то, что мы уже видели для градиентного спуска.

- Первый член – линейная сходимость к решению
- Второй член – говорит о том, что некоторую точность (зависящую от γ , σ и μ) метод преодолеть не может и начинает осциллировать, больше не приближаясь к решению.

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить шаг. Например, брать $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.

Вопрос: какой видно плюс и минус?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить шаг. Например, брать $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.

Вопрос: какой видно плюс и минус? Плюс – точнее сходимость, минус – потеря линейной сходимости в начале.

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить шаг. Например, брать $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.

Вопрос: какой видно плюс и минус? Плюс – точнее сходимость, минус – потеря линейной сходимости в начале.

- Уменьшить σ . **Вопрос:** а как?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить шаг. Например, брать $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.

Вопрос: какой видно плюс и минус? Плюс – точнее сходимость, минус - потеря линейной сходимости в начале.

- Уменьшить σ . **Вопрос:** а как? С помощью техники батчинга/батчирования:

$$\nabla f(x^k, \xi^k) \rightarrow \frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j),$$

где S^k – набор индексов из $[n]$, $|S^k| = b$, и все индексы генерируются независимо друг от друга.

Сходимость SGD: батчинг

- Вопрос: что можем сказать про

Сходимость SGD: батчинг

- **Вопрос:** что можем сказать про

$$\mathbb{E} \left[\frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j) \mid x^k \right], \quad \mathbb{E} \left[\left\| \frac{1}{b} \sum_{j \in S^k} (\nabla f(x, \xi_j) - \nabla f(x)) \right\|_2^2 \mid x^k \right] ?$$

- Независимость дает

$$\mathbb{E} \left[\frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j) \mid x^k \right] = \nabla f(x),$$

$$\mathbb{E} \left[\left\| \frac{1}{b} \sum_{j \in S^k} (\nabla f(x, \xi_j) - \nabla f(x)) \right\|_2^2 \mid x^k \right] \leq \frac{\sigma^2}{b}$$

- Получается дисперсию можно уменьшить в b раз, но тогда и вычисление стохастического градиента подорожает.

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\sigma^2}{\mu^2 b k}.$$

Линейная по «детерминистической» части и сублинейная по «стохастической».

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\sigma^2}{\mu^2 b k}.$$

Линейная по «детерминистической» части и сублинейная по «стохастической».

- Ускорение Нестерова возможно:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\sigma^2}{\mu^2 b k}.$$

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\sigma^2}{\mu^2 b k}.$$

Линейная по «детерминистической» части и сублинейная по «стохастической».

- Ускорение Нестерова возможно:

$$\mathbb{E} [\|x^k - x^*\|^2] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E} [\|x^0 - x^*\|^2] + \frac{\sigma^2}{\mu^2 b k}.$$

Важной деталью является улучшение/ускорение только первого члена, второй член (который и возникает из-за стохастики) остался прежним. Оказывается, его нельзя изменить и результат выше является оптимальным.

Почему SGD не сходится?

- Изначально у SGD с постоянным наблюдается поведение, как и градиентного спуска: $x \rightarrow x^*$, но потом начинаются осцилляции. **Вопрос:** с чем это связано? что такого неприятного появилось в физике метода?

Почему SGD не сходится?

- Изначально у SGD с постоянным наблюдается поведение, как и градиентного спуска: $x \rightarrow x^*$, но потом начинаются осцилляции. **Вопрос:** с чем это связано? что такого неприятного появилось в физике метода? В градиентном спуске $\nabla f(x) \rightarrow \nabla f(x^*) = 0$. Сейчас никто этого не гарантирует: $\nabla f(x, \xi)$ может не стремиться к 0.
- Это объяснимо на пример машинного обучения: x^* — минимизирует потери по всей выборке/по всему распределению. $f(x, \xi)$ отражает только потери на сэмпле ξ . Никто не гарантирует, что x^* — лучшая настройка модели для конкретного сэмпла ξ .
- Из-за того, что в общем случае $\nabla f(x^*, \xi) \neq 0$ для некоторых ξ и возникает осциллирующий эффект.

Модифицируем SGD

- Идея – взять метод на подобии SGD:

$$x^{k+1} = x^k - \gamma g^k,$$

где $g^k \rightarrow \nabla f(x^*) = 0$, при $x^k \rightarrow x^*$.

По возможности, чтобы

$$\mathbb{E} [g^k | x^k] = \nabla f(x^k) \quad \text{или} \quad \mathbb{E} [g^k] = \nabla f(x^k).$$

Модифицируем SGD

- Идея – взять метод на подобии SGD:

$$x^{k+1} = x^k - \gamma g^k,$$

где $g^k \rightarrow \nabla f(x^*) = 0$, при $x^k \rightarrow x^*$.

По возможности, чтобы

$$\mathbb{E} [g^k | x^k] = \nabla f(x^k) \quad \text{или} \quad \mathbb{E} [g^k] = \nabla f(x^k).$$

- В общем онлайн случае это нереализуемо. Но возможно в оффлайн вида:

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x),$$

где, генерируя равномерно и независимо i_k , получаем

$$\nabla f(x^k, \xi^k) = \nabla f_{i_k}(x^k).$$

SAGA

Алгоритм 2 SAGA

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: Сгенерировать независимо i_k
- 3: Вычислить $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k$
- 4: Обновить $y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k \\ y_i^k, & \text{иначе} \end{cases}$
- 5: $x^{k+1} = x^k - \gamma g^k$

6: **end for**

Выход: x^K

SAGA

- Идея – если я считал когда-то градиент для f_i , то зачем его забывать? Сохраним!

SAGA

- Идея – если я считал когда-то градиент для f_i , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$.

SAGA

- Идея – если я считал когда-то градиент для f_i , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$.
- $\mathbb{E} [g^k \mid x^k] = \nabla f(x^k)$.

SAGA

- Идея – если я считал когда-то градиент для f_i , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$.
- $\mathbb{E}[g^k | x^k] = \nabla f(x^k)$.
- При $x^k \rightarrow x^*$ имеем, что $y_j^k \rightarrow \nabla f_j(x^*)$, и $\frac{1}{n} \sum_{j=1}^n y_j^k \rightarrow \nabla f(x^*) = 0$.
А значит $g^k \rightarrow 0$.

SAGA

- Идея – если я считал когда-то градиент для f_i , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$.
- $\mathbb{E}[g^k | x^k] = \nabla f(x^k)$.
- При $x^k \rightarrow x^*$ имеем, что $y_j^k \rightarrow \nabla f_j(x^*)$, и $\frac{1}{n} \sum_{j=1}^n y_j^k \rightarrow \nabla f(x^*) = 0$.

А значит $g^k \rightarrow 0$.

- Из минусов: лишняя $\mathcal{O}(nd)$ память. Будем решать ее на следующей лекции.