Стохастическая оптимизация. SGD

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

30 ноября 2023

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:

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

Стохастическая оптимизация: постановка

- Рассматривали такую задачу:

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

- Теперь сформулируем задачу следующим образом:

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

- Чтобы понять суть, рассмотрим пример из машинного обучения:

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

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 <u>онлайн</u> режиме: поступают сэмплы, мы их обрабатываем (можем считать градиент).
- Естественное предположение, что данные поступают несмещенно:

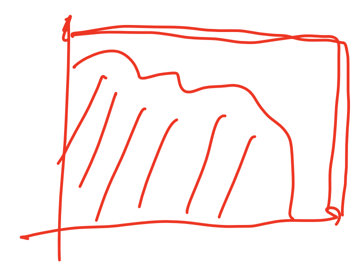
$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(x, \xi)] = \nabla f(x).$$

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 <u>оффлайн</u> (данные фиксированы, а не поступают в режиме реального времени).

- **Вопрос:** как связаны <u>онлайн</u> и <u>оффлайн</u>? <u>Оффлайн</u> – 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:

$$\nabla f(x, \xi_i), \quad \text{где} \quad \xi_i \text{ генерируется независимо и равномерно из } [n].$$

$\xi_i \sim \mathcal{D}$ $\qquad$ $\underline{\xi \sim [n]}$

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** $k = 0, 1, \ldots, K-1$ **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:

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

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

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

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

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

- 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:

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

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

- 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:
$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$
где $\mathcal{F}_k$ – σ-алгебра, порожденная x^0, $\xi^0, \ldots, \xi^{k-1}$.
- 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, $\xi^0, \ldots, \xi^{k-1}$.

Условное математическое ожидание

- В ходе доказательства сходимости потребуется ввести условное математическое ожидание:

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

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

- 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 $x^0, \xi^0, \ldots, \xi^{k-1}$.
- Также понадобится закон полного математического ожидания (tower property):

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

Сходимость: доказательство

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

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

$x^k \quad \xi^{k-1} \quad \ldots$

- Начинаем, как и раньше:

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

$\mathbb{E}[\cdot \mid x^k]$

$$\mathbb{E}[\|x^{k+1} - x^*\|_2^2 \mid x^k] = \|x^k - x^*\|_2^2 - 2\gamma_k \langle \mathbb{E}[\nabla f(x^k,\xi^k) \mid x^k]; x^k - x^* \rangle + \gamma_k^2 \mathbb{E}[\|\nabla f(x^k,\xi^k)\|_2^2 \mid x^k]$$

$$= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k); x^k - x^* \rangle$$

Сходимость: доказательство

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

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

- 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.):

$x^0, \dots \xi^{k-1}$ ξ^k

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

Сходимость: доказательство

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

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

$$= \langle \nabla f(x^k \cancel{, \xi^k}), x^k - x^* \rangle$$

$$\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|_2^2 \mid x^k\right] = \mathbb{E}\left[\|\nabla f(x^k, \xi^k) - \nabla f(x^k) + \nabla f(x^k)\|^2 \mid x^k\right]$$

$$= \mathbb{E}\left[\|\nabla f(x^k, \xi^k) - \nabla f(x^k)\|_2^2 \mid x^k\right] + \|\nabla f(x^k)\|_2^2$$

$$+ 2\,\mathbb{E}\left[\langle \nabla f(x^k, \xi^k) - \nabla f(x^k); \nabla f(x^k) \rangle \mid x^k\right]$$

$$\le \sigma^2 + \|\nabla f(x^k)\|_2^2$$

Сходимость: доказательство

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

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

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

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

Сходимость: доказательство

- Продолжаем:

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

Сходимость: доказательство

- Продолжаем:

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

- Предположение о стохастическом градиенте дает

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

Сходимость: доказательство

- Все, что получили:

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

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

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

Сходимость: доказательство

- Все, что получили:

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

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

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

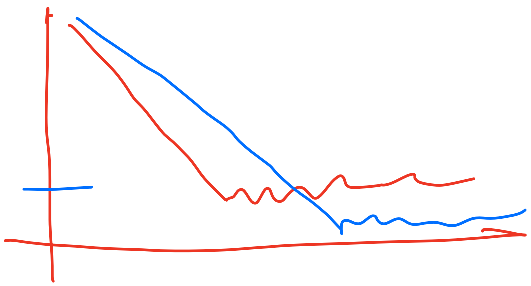
- Итого

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

Сходимость: доказательство

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

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



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

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

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

$$
\leq (1-\gamma\mu)^{k+1}\,\mathbb{E}\left[\|x^0-x^*\|_2^2\right] + \gamma^2\sigma^2 \underbrace{\sum_{i=0}^{k}(1-\gamma\mu)^i}_{=\frac{1}{\gamma\mu}}
$$

$$
\leq (1-\gamma\mu)^{k+1}\,\mathbb{E}\|x^0-x^*\|_2^2 + \boxed{\frac{\gamma\sigma^2}{\mu}}
$$

Сходимость: доказательство

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

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

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

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

- Взяв полное ожидание и применив tower property:

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

Сходимость SGD

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 насыщенности 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

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

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?

Сходимость SGD: анализ

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

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

- **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:

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

  похож на то, что мы уже видели для градиентного спуска.
- Первый член – линейная сходимость к решению

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

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

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?

$$\frac{1}{b} \sum_{j \in S_k} \nabla f(x, \xi_j) \qquad |S_k| = b$$

$$\mathbb{E} \left\| \frac{1}{b} \sum_{j \in S_k} \left(\nabla f(x, \xi_j) - \nabla f(x) \right) \right\|_2^2 = \frac{1}{b^2} \underbrace{\sum}_{b} \underbrace{\mathbb{E} \| \ldots \|}_{\sigma^2} \leq \frac{\sigma^2}{b}$$

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:

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

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

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

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:

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

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

Сходимость SGD

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

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

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

- Ускорение Нестерова возможно:

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

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости: $(0{,}99)^k$ $\quad \frac{\mu}{L} = \frac{1}{100}$

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

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

- Ускорение Нестерова возможно: $(0{,}9)^k$

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

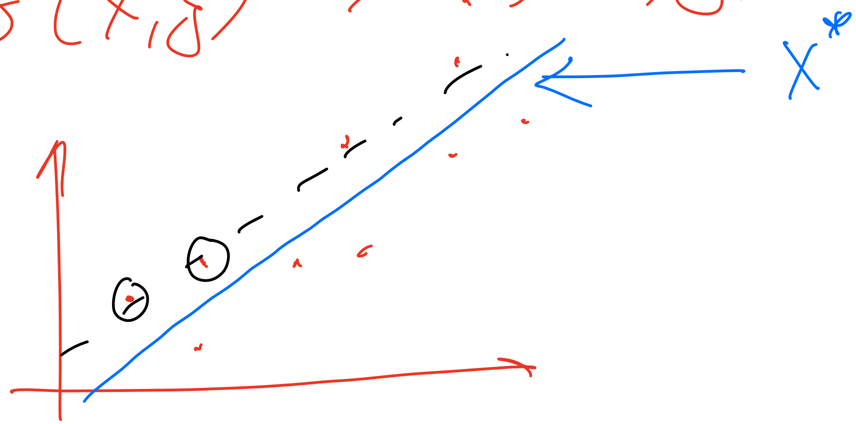
Важной деталью является улучшение/ускорение только первого члена, второй член (который и возникает из-за стохастики) остался прежним. Оказывается, его нельзя изменить и результат выше является оптимальным.

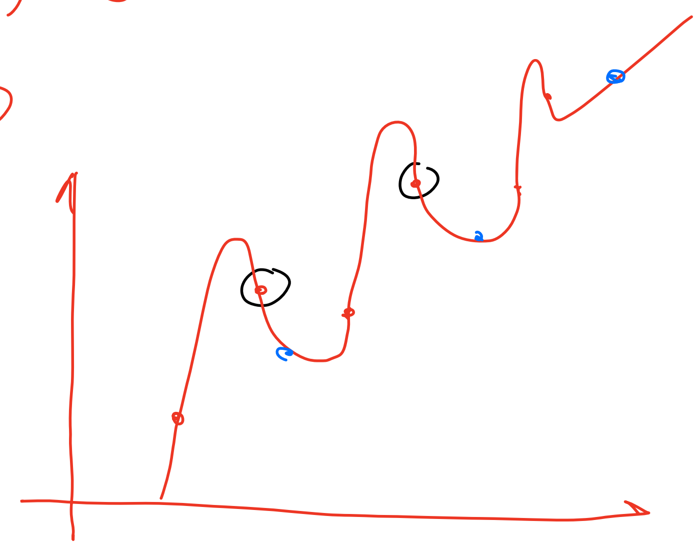
Почему SGD не сходится?

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

$\nabla f(x) \to \nabla f(x^*) = 0$

$\nabla f(x, \xi) \to \nabla f(x^*, \xi) \neq 0$





Почему SGD не сходится?

- Изначально у SGD с постоянным наблюдается поведение, как и градиентного спуска: $x \to x^*$, но потом начинаются осцилляции. **Вопрос:** с чем это связано? что такого неприятного появилось в физике метода? В градиентном спуске $\nabla f(x) \to \nabla f(x^*) = 0$. Сейчас никто этого не гарантирует: $\nabla f(x, \xi)$ может не стремится к 0.

Почему SGD не сходится?

- Изначально у SGD с постоянным наблюдается поведение, как и градиентного спуска: $x \to x^*$, но потом начинаются осцилляции. **Вопрос:** с чем это связано? что такого неприятного появилось в физике метода? В градиентном спуске $\nabla f(x) \to \nabla f(x^*) = 0$. Сейчас никто этого не гарантирует: $\nabla f(x, \xi)$ может не стремится 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.

Почему SGD не сходится?

- Изначально у SGD с постоянным наблюдается поведение, как и градиентного спуска: $x \to x^*$, но потом начинаются осцилляции. **Вопрос:** с чем это связано? что такого неприятного появилось в физике метода? В градиентном спуске $\nabla f(x) \to \nabla f(x^*) = 0$. Сейчас никто этого не гарантирует: $\nabla f(x, \xi)$ может не стремится 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

$$g^k \to \nabla f(x^*) = 0, \quad \text{при} \quad x^k \to x^*.$$

По возможности, чтобы

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

Модифицируем SGD

- Идея – взять метод на подобии SGD:

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

где

$$g^k \to \nabla f(x^*) = 0, \quad \text{при} \quad x^k \to x^*.$$

По возможности, чтобы

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

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

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

$\xi \sim D$

где, генерируя равномерно и независимо i_k, получаем

$i_k \in [n]$

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

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** $k = 0, 1, \ldots, K-1$ **do**

$\nabla f_{i_k}(x^*)$ $\quad \nabla f_{i_k}(x^*)$ $\quad 0$

2: Сгенерировать независимо i_k

3: Вычислить $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^{n} y_j^k$ $\quad \to 0$

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!
- $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «запаздывающая» версия $\nabla f(x^k)$.

SAGA

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

SAGA

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

  А значит $g^k \to 0$.

SAGA

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

  А значит $g^k \to 0$.
- Из минусов: лишняя $\mathcal{O}(nd)$ память.

SVRG

Алгоритм 3 SVRG

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций в эпохе K, количество эпох S

for $s = 0, 1, \ldots, S-1$ **do**
  Обновить $w^s = x^{s-1,K}$
  Посчитать и сохранить $\nabla f(w^s)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $x^{s,k+1} = x^{s,k} - \gamma g^k$
    Сгенерировать независимо i_k
    Вычислить $g^{k+1} = \boxed{\nabla f_{i_k}(x^{s,k+1})} - \nabla f_{i_k}(w^s) + \boxed{\nabla f(w^s)}$
  end for
end for

Выход: $x^{S-1,K}$

$\nabla f_{i_k}(x^*)$ $\nabla f(x^{s,k+1})$

$\nabla f_{i_k}(x^*)$ 0

$x^k \to x^*$ $w^k \to x^*$

SVRG

- Идея – редко считать полный градиент в некоторой референсной точке!

SVRG

- Идея – редко считать полный градиент в некоторой референсной точке!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.

SVRG

- Идея – редко считать полный градиент в некоторой референсной точке!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- При $x^k \to x^*$ имеем, что $w^k \to x^*$, $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)) \to 0$, и $\nabla f(w^*) \to \nabla f(x^*) = 0$. А значит $g^k \to 0$.

SVRG

- Идея – редко считать полный градиент в некоторой референсной точке!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- При $x^k \to x^*$ имеем, что $w^k \to x^*$, $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)) \to 0$, и $\nabla f(w^*) \to \nabla f(x^*) = 0$. А значит $g^k \to 0$.
- Из минусов: нужно иногда считать полный градиент и каждую итерацию вычислять два раза ∇f_{i_k}.

x^k $\quad$ w^s

Алгоритм 4 SARAH

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций в эпохе K, количество эпох S

$x \to x^*$

for $s = 0, 1, \ldots, S-1$ **do**

  Посчитать $g^0 = \nabla f(x^{s-1,K})$

  for $k = 0, 1, \ldots, K-1$ **do**

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

    Сгенерировать независимо i_k

    Вычислить $g^{k+1} = \nabla f_{i_k}(x^{s,k+1}) - \nabla f_{i_k}(x^{s,k}) + g^k$

  end for

end for

$g^k = \text{const}$

Выход: $x^{S-1,K}$

SARAH

- Идея – более «плавно» по сравнению с SVRG считать референсный градиент!

SARAH

- Идея – более «плавно» по сравнению с SVRG считать референсный градиент!
- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, но $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$

SARAH

- Идея – более «плавно» по сравнению с SVRG считать референсный градиент!

$\mathbb{E}[\quad]$

- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, но $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$
- При $x^k \to x^*$ имеем, что $(\nabla f_{i_k}(\cancel{x^k}) - \nabla f_{i_k}(\cancel{x^{k-1}})) \to 0$, и $g^k \to \text{const}$ в пределах одной эпохи (запуска внутреннего цикла), но в силу обновления $g^k = \nabla f(x^{s-1,K})$: $g^k \to 0$.

SARAH

- Идея – более «плавно» по сравнению с SVRG считать референсный градиент!
- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, но $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$
- При $x^k \to x^*$ имеем, что $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-1})) \to 0$, и $g^k \to \text{const}$ в пределах одной эпохи (запуска внутреннего цикла), но в силу обновления $g^k = \nabla f(x^{s-1,K})$: $g^k \to 0$.
- Из минусов: нужно иногда считать полный градиент и каждую итерацию вычислять два раза ∇f_{i_k}.

Методы редукции диспресии: итог

- Предназначены для стохастических задач вида конечной суммы (оффлайн минимизация эмпирического риска).

Методы редукции диспресии: итог

- Предназначены для стохастических задач вида конечной суммы (оффлайн минимизация эмпирического риска).
- Обеспечивают сходимость, как у градиентного спуска,

  Суммарно $\mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right)$ итераций для SAGA/SVRG/SARAH.

  но в n раз дешевле (считаем не полный градиент, а только 1 слагаемое).

Методы редукции диспресии: итог

- Предназначены для стохастических задач вида конечной суммы (оффлайн минимизация эмпирического риска).
- Обеспечивают сходимость, как у градиентного спуска,

  Суммарно $\mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right)$ итераций для SAGA/SVRG/SARAH.

  но в n раз дешевле (считаем не полный градиент, а только 1 слагаемое).
- Обладают недостатками: траты памяти, подсчет полного градиента.

Методы редукции диспресии: итог

- Предназначены для стохастических задач вида конечной суммы (оффлайн минимизация эмпирического риска).
- Обеспечивают сходимость, как у градиентного спуска,

  Суммарно $\mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right)$ итераций для SAGA/SVRG/SARAH.

  но в n раз дешевле (считаем не полный градиент, а только 1 слагаемое).
- Обладают недостатками: траты памяти, подсчет полного градиента.
- Могут быть ускорены (SVRG $\to$ Katyusha).