# Стохастический градиентный спуск.
# 1 апреля 2020 г.

*Семинарист: Данилова М.*

## Стохастический градиентный спуск.

Чтобы формально все шаги в методе градиентного спуска были корректно определены, необходимо, чтобы

а) функция $f$ имела градиент во всех, генерируемых точках и

б) можно было бы *посчитать градиент* функции в любой указанной наперёд точке.

Иными словами, мы неявно предполагали, что вычислительная процедура способна вычислять градиент в любой точке, в которой мы захотим. Однако в реальных задачах с вычислением градиента функции может возникнуть ряд проблем. Например, возможна ситуация, когда вычислить градиент точно мы не можем, а можем лишь вычислить его с некоторыми ошибками, имеющими случайную природу.

Более того, часто возникают ситуации (особенно в машинном обучении), когда подсчёт градиента оптимизируемой функции занимает много времени, но хочется использовать что-то похожее на градиентный спуск.

Во всех перечисленных случаях можно рассматривать исходную задачу, как *задачу стохастической оптимизации со стохастическим оракулом первого порядка.*

**Определение 1** (Стохастический градиент с конечной дисперсией)**.** Случайный вектор $g(x)$, зависящий от параметра $x \in \mathbb{R}^n$, будем называть стохастическим градиентом функции $f$ с конечной дисперсией, если для любой точки $x \in \mathbb{R}^n$

$$\mathbb{E}\left[g(x)|x\right] = \nabla f(x) \tag{1}$$

и существует такое число $\sigma \geq 0$, что для всех $x \in \mathbb{R}^n$

$$\mathbb{E}\left[\|g(x) - \nabla f(x)\|_2^2 |x\right] \leq \sigma^2. \tag{2}$$

Отметим, что в определении стохастического градиента рассматриваются условные математические ожидания, т.к. в алгоритмах стохастической оптимизации обычно точка, в которой рассматривается стох. градиент, сама по себе является случайным вектором. То есть введённые определения характеризуют только ту случайность, которую превносит стох. градиент в указанной точке.

Часто методы стохастической оптимизации первого порядка получаются из детерминированных методов оптимизации заменой градиента на его стохастический аналог. Если в методе градиентного спуска, в формуле для $x^{k+1}$ градиент $\nabla f(x^k)$ заменить на стохастический градиент $g(x^k)$, то получится *стохастический градиентный спуск* (stochastic gradient descent, или SGD).

Итак, рассмотрим задачу
$$f(x) \to \min_{x \in \mathbb{R}^n},$$
где $f$ сильно выпукла с константой $\mu$ и $f$ имеет Липшицев градиент с константой $L$ в евклидовой норме. Ниже приведена процедура `SGD`. Здесь добавились новые параметры $\gamma^k$ — это размеры шагов.

**procedure** `SGD`$(f, x^0, N, \{\gamma^k\}_{k=1}^{N-1})$
  **for** $k = 0, 1, 2, \ldots, N-1$ **do**
    Сгенерировать $g(x^k)$ независимо от предыдущих шагов
    $x^{k+1} := x^k - \gamma^k g(x^k)$
  **end for**
  **return** $x^N$
**end procedure**

**Пример 2** (`SGD` в задачах машинного обучения)**.** Зачастую задачи машинного обучения (задачи классификации) сводятся к поиску оптимальных параметров модели, которые ищутся путём минимизации некоторой функции потерь на обучающей выборке:
$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) \to \min_{x \in \mathbb{R}^n},$$
где $x \in \mathbb{R}^n$ — это вектор параметров, который задаёт модель, а $f_i(x)$ — это потери на $i$-м элементе обучающей выборки, при использовании модели с параметрами $x$. Например, если $\{(z_i, y_i)\}_{i=1}^m$ — обычающая выборка, $z_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$, функция потерь в методе `SVM` представима в указанном виде:
$$f(x) = \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle x, z_i\rangle\} = \frac{1}{m}\sum_{i=1}^{m} f_i(x),$$
$$\text{где } f_i(x) = \max\{0, 1 - y_i\langle x, z_i\rangle\} + \frac{\mu}{2}\|x\|_2^2.$$
Чтобы посчитать градиент функции $f(x)$ в таких задачах, нужно посчитать градиенты всех функций $f_i(x)$, что может быть достаточно затратно по времени. В таком случае можно искуственно ввести стох. градиент. Вместо градиента функции $f(x)$ будем рассматривать вектор $g(x)$, который равен градиенту случайного слагаемого $f_i(x)$ в записанной выше сумме (каждое из слагаемых выбираем с вероятностью $\frac{1}{m}$). Тогда
$$\mathbb{E}[g(x)] = \sum_{i=1}^{m} \nabla f_i(x) \cdot \underbrace{\mathbb{P}\{g(x) = \nabla f_i(x)\}}_{\frac{1}{m}} = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x) = \nabla f(x),$$
то есть для введённого случайного вектора $g(x)$ выполняется условие (1). Обычно предполагается, что условие (2) выполняется для всех точек $x^k$, которые генерирует алгоритм `SGD`, и

это предположение на самом деле не обременительно, поскольку во многих реальных задачах машинного обучения оно оказывается выполненным (по крайней мере с большой вероятностью; конечно, есть ненулевая вероятность, уйти сколь угодно далеко через конечное число итераций, но вопрос о том, как эту сложность миновать, требует чуть более серьёзного погружения в стохастическую оптимизацию, а эта сложность была впервые преодолена в статье 2018 года).

Теперь займёмся вопросом о сходимости метода SGD. Чаще всего в методах стохастической оптимизации исследуется вопрос о сходимости в среднем (хотя, строго говоря, в таком случае может и не существовать сходимости почти неаверное). Иными словами, вместо невязки по функции $f(x^N) - f(x^*)$ или по аргументу $\|x^N - x^*\|_2^2$ исследуются средние невязки по функции $\mathbb{E}\left[f(x^N) - f(x^*)\right] = \mathbb{E}[f(x^N)] - f(x^*)$ и по аргументу $\mathbb{E}\left[\|x^N - x^*\|_2^2\right]$.

**Лемма 1** (Основная лемма о сходимости SGD)**.** Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ имеет Липшицев градиент с константой $L$ в норме $\|\cdot\|_2$ и функция $f$ сильно выпукла с константой $\mu$ в норме $\|\cdot\|_2$. Пусть последовательность $\{\gamma^k\}_{k=1}^N$ удовлетворяет условию $0 < \gamma^k \leq \frac{1}{L}$ и процедура SGD имеет доступ к стохастическому градиенту, удовлетворяющему условиям (1) и (2). Тогда для всех точек $x^k, k = 0, 1, 2, \ldots, N-1$, генерируемых процедурой SGD$(f, x^0, N, \{\gamma\}_{k=1}^N)$, будет выполнено следующее неравенство:

$$\mathbb{E}[f(x^{k+1})] - f(x^*) \leq (1 - \mu\gamma^k)\left(\mathbb{E}[f(x^k)] - f(x^*)\right) + \frac{L\left(\gamma^k\right)^2}{2}\sigma^2. \tag{3}$$

*Доказательство.* Запишем неравенство для Липшицевых функций для точек $y = x^{k+1}$ и $x = x^k$:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^{k+1}), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\ &= f(x^k) - \gamma^k \langle \nabla f(x^k), g(x^k) \rangle + \frac{L\left(\gamma^k\right)^2}{2}\|g(x^k)\|_2^2. \end{aligned}$$

Возьмём от предыдущего неравенства математическое ожидание от левой и правой частей и вычтем из обеих частей $f(x^*)$:

$$\mathbb{E}[f(x^{k+1})] - f(x^*) \leq \mathbb{E}[f(x^k)] - f(x^*) - \gamma^k \mathbb{E}\left[\langle \nabla f(x^k), g(x^k) \rangle\right] + \frac{L\left(\gamma^k\right)^2}{2}\mathbb{E}\left[\|g(x^k)\|_2^2\right]. \tag{4}$$

Условное математическое ожидание обладает замечательным свойством, которое часто в англоязычных источниках называют tower property: $\mathbb{E}[\cdot] = \mathbb{E}\left[\mathbb{E}\left[\cdot | x^k\right]\right]$. Используя tower property, получим

$$\begin{aligned} \mathbb{E}\left[\langle \nabla f(x^k), g(x^k) \rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle \nabla f(x^k), g(x^k) \rangle | x^k\right]\right] \\ &= \mathbb{E}\left[\langle \nabla f(x^k), \mathbb{E}\left[g(x^k) | x^k\right] \rangle\right] \\ &\overset{(1)}{=} \mathbb{E}\left[\|\nabla f(x^k)\|_2^2\right] \end{aligned}$$

и

$$\begin{aligned}
\mathbb{E}\left[\|g(x^k)\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|g(x^k)\|_2^2 | x^k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|g(x^k) - \nabla f(x^k)\|_2^2 + 2\langle g(x^k) - \nabla f(x^k), \nabla f(x^k)\rangle + \|\nabla f(x^k)\|_2^2 | x^k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|g(x^k) - \nabla f(x^k)\|_2^2 | x^k\right] + 2\langle \mathbb{E}\left[g(x^k) - \nabla f(x^k) | x^k\right], \nabla f(x^k)\rangle + \|\nabla f(x^k)\|_2^2\right] \\
&\overset{(1),(2)}{\leq} \sigma^2 + \mathbb{E}\left[\|\nabla f(x^k)\|_2^2\right].
\end{aligned}$$

Подставляя полученные оценки в неравенство (4), получим

$$\mathbb{E}[f(x^{k+1})] - f(x^*) \leq \mathbb{E}[f(x^k)] - f(x^*) - \gamma^k \left(1 - \frac{L\gamma^k}{2}\right) \mathbb{E}\left[\|\nabla f(x^k)\|_2^2\right] + \frac{L\left(\gamma^k\right)^2}{2}\sigma^2.$$

Заметим, что в силу $0 < \gamma^k \leq \frac{1}{L}$, множитель $-\gamma^k \left(1 - \frac{L\gamma^k}{2}\right)$ перед $\mathbb{E}\left[\|\nabla f(x^k)\|_2^2\right]$ отрицательный, а значит, используя неравенство для сильно выпуклых функций

$$\forall x \in \mathbb{R}^n \hookrightarrow \|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*)),$$

мы получаем неравенство

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

Чтобы получить (3), осталось заметить, что $0 < \gamma^k \leq \frac{1}{L} \implies 2 - L\gamma^k \geq 1 \implies \left(1 - \mu\gamma^k \left(2 - L\gamma^k\right)\right) \leq (1 - \mu\gamma^k)$. □

**Теорема 1.** Пусть выполнены все условия Леммы 1, а шаги $\gamma^k$ выбираются постоянными: $\gamma^k \equiv \gamma \leq \frac{1}{L}$. Тогда для точки $x^N$, генерируемой процедурой $\texttt{SGD}(f, x^0, N, \{\gamma\}_{k=1}^N)$, будет выполнено следующее неравенство:

$$\mathbb{E}[f(x^N)] - f(x^*) \leq (1 - \mu\gamma)^N \left(\mathbb{E}[f(x^0)] - f(x^*)\right) + \frac{\sigma^2}{2\mu}. \tag{5}$$

*Доказательство.* Если применять неравенство (3) к своей правой части, то получится следующее:

$$\begin{aligned}
\mathbb{E}[f(x^N)] - f(x^*) &\leq (1 - \mu\gamma)^N \left(f(x^0) - f(x^*)\right) + \frac{L\gamma^2\sigma^2}{2}\sum_{k=0}^{N-1}(1 - \mu\gamma)^k \\
&\leq (1 - \mu\gamma)^N \left(f(x^0) - f(x^*)\right) + \frac{L\gamma^2\sigma^2}{2}\sum_{k=0}^{\infty}(1 - \mu\gamma)^k \\
&= (1 - \mu\gamma)^N \left(f(x^0) - f(x^*)\right) + \frac{L\gamma^2\sigma^2}{2} \cdot \frac{1}{\gamma\mu} \\
&\leq (1 - \mu\gamma)^N \left(f(x^0) - f(x^*)\right) + \frac{\sigma^2}{2\mu},
\end{aligned}$$

где последнее неравенство следует из $\gamma \leq \frac{1}{L}$. □

Заметим, что если $\sigma = 0$, то стохастический градиент — это с вероятностью 1 просто градиент функции $f$, то есть $g(x) = \nabla f(x)$ с вероятностью 1. Иными словами, `SGD` совпадает с градиентным спуском, если $\sigma = 0$. Но тогда мы можем применить результат предыдущей теоремы и для градиентного спуска.

**Следствие 1.** Пусть функция $f$ сильно выпуклая с константой $\mu$ в евклидовой норме и имеет Липшицев на $\mathbb{R}^n$ градиент с константой $L$ в евклидовой норме. Тогда процедура `Gradient_Descent`$(f, x^0, N)$ вернёт точку $x^N$, для которой

$$f(x^N) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^N \left(f(x^0) - f(x^*)\right). \tag{6}$$

Иными словами, чтобы

*Доказательство.* Указанное неравенство следует из (5) с шагом $\gamma = \frac{1}{L}$ и $\sigma = 0$. □

**Следствие 2.** Чтобы получить приближение решения с точностью $\varepsilon$ по значению функции для функции $f$, удовлетворяющей описанным выше свойствам, достаточно сделать $O\left(\frac{L}{\mu} \ln \frac{1}{\varepsilon}\right)$ шагов градиентного спуска.

Итак, мы доказали, что при использовании `SGD` с постоянным шагом через $N$ шагов будет выполнено неравенство (5). Недостаток такой оценки в том, что она не гарантирует, что мы сможем получить решение со сколь угодно большой точностью по функции, если сделаем сколь угодно большое число шагов. И действительно, на практике `SGD` с постоянным шагом не будет сходиться к сколь угодно малой окрестности решения. Всему виной слагаемое в правой части (5), пропорциональное дисперсии.

Существуют разные способы борьбы с дисперсией, но мы рассмотрим 2 таких способа.

1. **Мини-батчинг.** Оказывается, если на шаге 3 процедуры `SGD` независимо сэмплировать не один градиент, а сразу $l$ штук, и в качестве стох. градиента использовать их среднее арифметическое $\tilde{g}^l(x) = \frac{1}{l}\sum\limits_{i=1}^{l} g^i(x)$, то получим:

$$\mathbb{E}\left[\tilde{g}^l(x)\right] = \frac{1}{l}\sum_{i=1}^{l} \mathbb{E}[g^i(x)] \overset{(1)}{=} \nabla f(x),$$

и, пользуясь тем, что дисперсия суммы независимых случайных величин, равна сумме дисперсий, получим

$$\mathbb{E}\left[\|\tilde{g}^l(x) - \nabla f(x)\|_2^2\right] = \frac{1}{l^2}\mathbb{E}\left[\left\|\sum_{i=1}^{l}(g^i - \nabla f(x))\right\|_2^2\right] \overset{(2)}{\leq} \frac{\sigma^2}{l}.$$

Тогда в итоговой оценке (5) при использовании стох. градиента $\tilde{g}^l(x)$ в правой части вместо $\sigma^2$ будет фигурировать $\frac{\sigma^2}{l}$. Беря достаточно большой размер батча $l$, мы можем уменьшить дисперсию настолько, насколько захотим.

У этого подхода есть очевидная проблема. Если дисперсия $\sigma^2$ велика, то приёдтся брать слишком большой размер батча и пропадёт выгода от того, что мы считаем не градиент функции $f(x)$ целиком (если вспомнить задачу оптимизацию, которая часто возникает в машинном обучении, то при больших $l$ мы будем считать практически все градиенты функций $f_i$, что не даёт существенного выигрыша в скорости по сравнению с подсчётом просто градиента функции $f$). Поэтому в машинном обучении только лишь минибатчингом при использовании `SGD` проблему не решить. Так мы приходим ко второму методу борьбы с дисперсией.

2. **Уменьшающийся шаг.** Оказывается, что с дисперсией можно бороться путём уменьшения шагов $\gamma^k$. Покажем это строго.

**Лемма 2.** Пусть последовательность неотрицательных чисел $\{a^k\}_{k=0}^{\infty}$, удовлетворяет неравенству

$$a^{k+1} \leq (1-\mu\gamma^k)a^k + (\gamma^k)^2 T,$$

где $0 < \gamma^k < \gamma^0$, $\theta = \frac{2}{\gamma^0}$ и $C$ — такая константа, что $C \geq \max\left\{a^0, \frac{4T}{\mu\theta}\right\}$. Тогда, если выбрать $\gamma^k = \frac{2}{\mu k + \theta}$, то

$$a^k \leq \frac{C}{\frac{\mu}{\theta}k + 1}. \tag{7}$$

*Доказательство.* Докажем (7) по индукции. Для $k = 0$ неравенство очевидно. Пусть теперь оно выполнено для всех $k = 1, 2, \ldots, N$, и докажем его для $k = N$. Из условия мы имеем:

$$\begin{aligned} a^{k+1} &\leq (1-\gamma^k\mu)a^k + (\gamma^k)^2 T \\ &\leq \left(1 - \frac{2\mu}{\mu k + \theta}\right) \cdot \frac{\theta C}{\mu k + \theta} + \theta\mu\frac{C}{(\mu k + \theta)^2}. \end{aligned}$$

Нам осталось показать, что

$$\left(1 - \frac{2\mu}{\mu k + \theta}\right) \cdot \frac{\theta C}{\mu k + \theta} + \theta\mu\frac{C}{(\mu k + \theta)^2} \leq \frac{\theta C}{\mu(k+1) + \theta}.$$

Домножим обе части на $\frac{\mu(k+1)+\theta}{\theta C} \cdot (\mu k + \theta)$:

$$\left(1 - \frac{2\mu}{\mu k + \theta}\right)(\mu k + \mu + \theta) + \mu \cdot \frac{\mu k + \mu + \theta}{\mu k + \theta} \leq \mu k + \theta \Longleftrightarrow \mu - \mu \cdot \frac{\mu k + \mu + \theta}{\mu k + \theta} \leq 0,$$

где последнее неравенство выполнено. □

**Теорема 2.** Пусть выполнены все условия Леммы 1, а шаги $\gamma^k$ выбираются уменьшающимися: $\gamma^k = \frac{2}{\mu k + 2L} \leq \frac{1}{L}$. Тогда для точки $x^N$, генерируемой процедурой `SGD`$(f, x^0, N, \{\gamma\}_{k=1}^N)$, будет выполнено следующее неравенство:

$$\mathbb{E}[f(x^N)] - f(x^*) \leq \frac{2L}{\mu k + 2L} \max\left\{f(x^0) - f(x^*), \frac{\sigma^2}{\mu}\right\}. \tag{8}$$

*Доказательство.* Заметим, что из (3) следует, что условия Леммы 2 выполнены для $a^k = \mathbb{E}[f(x^k)] - f(x^*), T = \frac{L\sigma^2}{2}, \theta = 2L, C = \max\left\{f(x^0) - f(x^*), \frac{\sigma^2}{\mu}\right\}$. Применяя результат леммы, получим неравенство (8). □