Онлайн градиентный спуск

Онлайн градиентный спуск (online gradient descent, OGD) для начальной точки (базового решения) $x_1 \in \mathcal{D}$ и последовательности шагов $\{\alpha_t\}$ конструирует последовательность решений $\{x_t\}$ по следующим правилам:

- игрок в момент $t \in [1, T]$ принимает решение x_t и получает потери $f_t(x_t)$;
- осуществляется сдвиг и проектирование:

$$
\begin{aligned}
y_{t+1} &= x_t - \alpha_t \nabla f_t(x_t), \\
x_{t+1} &= \Pi_{\mathcal{D}}(y_{t+1}).
\end{aligned}
$$

Будем предполагать, что справедливы следующие утверждения:

- существует положительная постоянная D такая, что для любых $x', x'' \in \mathcal{D}$ справедливо неравенство $\|x' - x''\| \le D$;
- существует положительная константа G такая, что для любой функции потерь $f \in \mathcal{F}$ и допустимого решения $x \in \mathcal{D}$ справедлива оценка $\|\nabla f(x)\| \le G$.

Теорема 1. *Для переменного шага $\alpha_t = \frac{D}{G\sqrt{t}}$ онлайн градиентный спуск обеспечивает следующую оценку на регрет:*

$$
\mathrm{regret}_t(\mathrm{OGD}) = \sum_{s=1}^{t} f_s(x_s) - \min_{x \in \mathcal{D}} \sum_{s=1}^{t} f_s(x) \le \frac{3}{2} G D \sqrt{t}.
$$

Доказательство. Пусть

$$
x_* \in \arg\min_{x \in \mathcal{D}} \sum_{s=1}^{t} f_s(x).
$$

В силу выпуклости функции $f_s(x)$ имеем неравенство

$$
f_s(x_*) \ge f_s(x_s) + \langle \nabla f_s(x_s), x_* - x_s \rangle,
$$

которое перепишем в виде

$$
f_s(x_s) - f_s(x_*) \le \langle \nabla f_s(x_s), x_s - x_* \rangle. \tag{1}
$$

Далее,

$$
\begin{aligned}
\|x_{s+1} - x_*\|^2 = \|\Pi_{\mathcal{D}}(x_s - \alpha_s \nabla f_s(x_s)) - x_*\|^2 \le \|x_s - \alpha_s \nabla f_s(x_s) - x_*\|^2 = \\
= \|x_s - x_*\|^2 + \alpha_s^2 \|\nabla f_s(x_s)\|^2 - 2\alpha_s \langle \nabla f_s(x_s), x_s - x_* \rangle,
\end{aligned}
$$

откуда

$$
\begin{aligned}
\langle \nabla f_s(x_s), x_s - x_* \rangle &\le \frac{1}{2\alpha_s} \left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2 \right) + \frac{\alpha_s}{2} \|\nabla f_s(x_s)\|^2 \le \\
&\le \frac{1}{2\alpha_s} \left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2 \right) + \frac{\alpha_s}{2} G^2.
\end{aligned}
$$

Суммируем теперь (1) по s от 1 до t (полагая $\alpha_0 = +\infty$ в последующих выкладках):

$$\begin{aligned}
\mathrm{regret}_t(\mathrm{OGD}) &= \sum_{s=1}^{t} \left(f_s(x_s) - f_s(x_*)\right) \le \sum_{s=1}^{t} \langle \nabla f_s(x_s), x_s - x_* \rangle \le \\
&\le \sum_{s=1}^{t} \frac{1}{2\alpha_s} \left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2\right) + \frac{G^2}{2} \sum_{s=1}^{t} \alpha_s = \\
&= \frac{1}{2} \sum_{s=1}^{t} \|x_s - x_*\|^2 \left(\frac{1}{\alpha_s} - \frac{1}{\alpha_{s-1}}\right) - \frac{\|x_{t+1} - x_*\|^2}{2\alpha_t} + \frac{G^2}{2} \sum_{s=1}^{t} \alpha_s \le \\
&\le \frac{D^2}{2} \sum_{s=1}^{t} \left(\frac{1}{\alpha_s} - \frac{1}{\alpha_{s-1}}\right) + \frac{G^2}{2} \sum_{s=1}^{t} \alpha_s = \frac{D^2}{2\alpha_t} + \frac{G^2}{2} \sum_{s=1}^{t} \alpha_s = \\
&= \frac{DG\sqrt{t}}{2} + \frac{DG}{2} \sum_{s=1}^{t} \frac{1}{\sqrt{s}} \le \frac{3DG\sqrt{t}}{2}. \qquad \square
\end{aligned}$$

Теорема 2. *Если семейство $\mathcal{F}$ состоит из μ-сильно выпуклых функций потерь, то онлайн градиентный спуск с шагом $\alpha_t = 1/\mu t$ гарантирует следующую оценку на регрет:*

$$\mathrm{regret}_T(\mathrm{OGD}) \le \frac{G^2}{2\mu}(1 + \log T).$$

Доказательство. Поскольку функция $f_s(x)$ является μ-сильно выпуклой, то имеет место неравенство

$$f_s(x_*) \ge f_s(x_s) + \langle \nabla f_s(x_s), x_* - x_s \rangle + \frac{\mu}{2}\|x_* - x_s\|^2,$$

из которого

$$f_s(x_s) - f_s(x_*) \le \langle \nabla f_s(x_s), x_s - x_* \rangle - \frac{\mu}{2}\|x_* - x_s\|^2. \tag{2}$$

Далее,

$$\begin{aligned}
\|x_{s+1} - x_*\|^2 = \|\Pi_{\mathcal{D}}(x_s - \alpha_s \nabla f_s(x_s)) - x_*\|^2 \le \|x_s - \alpha_s \nabla f_s(x_s) - x_*\|^2 = \\
= \|x_s - x_*\|^2 + \alpha_s^2 \|\nabla f_s(x_s)\|^2 - 2\alpha_s \langle \nabla f_s(x_s), x_s - x_* \rangle,
\end{aligned}$$

откуда

$$\langle \nabla f_s(x_s), x_s - x_* \rangle \le \frac{1}{2\alpha_s}\left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2\right) + \frac{\alpha_s}{2} G^2.$$

Таким образом,

$$\begin{aligned}
\mathrm{regret}_t(\mathrm{OGD}) &= \sum_{s=1}^{t} \left(f_s(x_s) - f_s(x_*)\right) \le \sum_{s=1}^{t} \left[\langle \nabla f_s(x_s), x_s - x_* \rangle - \frac{\mu}{2}\|x_* - x_s\|^2\right] \le \\
&\le \frac{G^2}{2} \sum_{s=1}^{t} \alpha_s + \frac{1}{2} \sum_{s=1}^{t} \|x_s - x_*\|^2 \left(\frac{1}{\alpha_s} - \frac{1}{\alpha_{s-1}} - \mu\right) = \\
&= \frac{G^2}{2} \sum_{s=1}^{t} \alpha_t = \frac{G^2}{2\mu} \sum_{s=1}^{t} \frac{1}{t} = \frac{G^2}{2\mu}(1 + \log T). \qquad \square
\end{aligned}$$

Приложение к стохастической оптимизации. SGD. Будем рассматривать задачу условной оптимизации

$$\min_{x \in \mathcal{D}} f(x). \tag{3}$$

Однако, в отличие от стандартной выпуклой оптимизации, нам неизвестен точный градиент $\nabla f(x)$, а только его зашумленное представление $\tilde{\nabla}(x)$, которое удовлетворяет условиям

$$\mathbb{E}[\tilde{\nabla}(x)] = \nabla f(x), \qquad \mathbb{E}[\|\tilde{\nabla}(x)\|^2] \leq G^2.$$

Таким образом, в каждой точке $x \in \mathcal{D}$ мы получаем случайный вектор, представляющий градиент $\nabla f(x)$, причем математическое ожидание этого вектора и есть градиент $\nabla f(x)$.

Метод стохастического градиента (SGD, stochastic gradient descent) устроен следующим образом:

- Задается начальное приближение $x_1 \in \mathcal{D}$, а также последовательность шагов $\{\alpha_t\}$.
- Для каждого раунда $t \in [1, T]$ итеративно выполняются следующие шаги:
  - наблюдается случайный вектор $\tilde{\nabla}(x_t)$ и по определению $f_t(x) \triangleq \langle \tilde{\nabla}(x_t), x \rangle$,
  - осуществляется сдвиг и проектирование в соответствии с OGD:

$$\begin{aligned} y_{t+1} &= x_t - \alpha_t \nabla f_t(x_t), \\ x_{t+1} &= \Pi_{\mathcal{D}}(y_{t+1}). \end{aligned}$$

- В качестве приближенного решения задачи (3) возвращается среднее значение от всех предсказаний

$$\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t.$$

Теорема 3. SGD *для переменного шага* $\alpha_t = \frac{D}{G\sqrt{t}}$ *обеспечивает оценку*

$$\mathbb{E}[f(\bar{x}_T)] \leq \min_{x \in \mathcal{D}} f(x) + \frac{3GD}{2\sqrt{T}}.$$

Доказательство.

$$\begin{aligned} \mathbb{E}[f(\bar{x}_T)] - \min_{x \in \mathcal{D}} f(x) = \mathbb{E}[f(\bar{x}_T)] - f(x_*) &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(x_t)\right] - f(x_*) = \\ &= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(f(x_t) - f(x_*))\right] \leq \frac{1}{T}\mathbb{E}\left[\langle \nabla f(x_t), x_t - x_* \rangle\right] = \\ &= \frac{1}{T}\mathbb{E}\left[\langle \tilde{\nabla}(x_t), x_t - x_* \rangle\right] = \frac{1}{T}\mathbb{E}\left[f_t(x_t) - f_t(x_*))\right] \leq \\ &\leq \frac{\mathrm{regret}_T(\mathrm{OGD})}{T} \leq \frac{3GD}{2\sqrt{T}}. \end{aligned}$$

□

Упражнение 1. *Покажите, что* SGD *для сильно выпуклых функций с соответствующим выбором переменного шага* α_s *сходится со скоростью* $O(\log T/T)$.