Рогачев Александр. Заочно
Домашнее задание №2.
Стохастическая оптимизация. Методы маломерной оптимизации. Градиентный спуск

Основные задачи

1. (2 балла) Пусть η — случайный n-мерный вектор (например, стохастический градиент). Предположим, что $\mathbb{E}\left[\|\eta\|_2^2\right] < \infty$ (второй момент η ограничен). Пусть $\mathrm{Var}[\eta] = \mathbb{E}\left[\|\eta - \mathbb{E}\eta\|_2^2\right]$ (дисперсия случайного вектора η). Докажите, что $\mathbb{E}\left[\|\eta\|_2^2\right] = \mathrm{Var}\left[\eta\right] + \|\mathbb{E}\eta\|_2^2$.

   Решение:

   $$\mathbb{E}\left[\|\eta\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}\eta_i^2\right] = \sum_{i=1}^{n}\mathbb{E}[\eta_i^2] = \sum_{i=1}^{n}[(\mathbb{E}\eta_i)^2 + \mathrm{Var}[\eta_i]] = \sum_{i=1}^{n}(\mathbb{E}\eta_i)^2 + \sum_{i=1}^{n}\mathrm{Var}[\eta_i] = \|\mathbb{E}\eta\|_2^2 + \mathrm{Var}\left[\eta\right]$$

   .

2. (2 балла) Рассмотрим задачу минимизации суммы функций:

   $$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) \longrightarrow \min_{x\in\mathbb{R}^n}. \tag{1}$$

   Пусть ξ — это случайная величина, которая случайно равновероятно принимает значения из множества $\{1, 2, \ldots, m\}$. Как было показано на лекции, случайный вектор $\nabla f_\xi(x)$ является несмещённой оценкой градиента f, т.е. $\mathbb{E}\left[\nabla f_\xi(x)\right] = \nabla f(x)$. Кроме того, были озвучены результаты о сходимости стохастического градиентного спуска (`SGD`) в предположении ограниченности второго момента $\nabla f_\xi(x)$, т.е. в предположении, что существует такое число $M > 0$, что для всех $x \in \mathbb{R}^n$ выполняется $\mathbb{E}\left[\|\nabla f_\xi(x)\|_2^2\right] \leq M^2$. Покажите, что существует такая задача оптимизации (1), которая имеет решение, но для которой величина $\mathbb{E}\left[\|\nabla f_\xi(x)\|_2^2\right]$ не является ограниченной, в то время как дисперсия $\nabla f_\xi(x)$ ограничена некоторой константой[1].

   Решение: Возьмем значение $m = 1$ и функцию $f(x) = x^2$. Данная функция является выпуклой, решение задачи минимизации (1) имеет решение. В данном случае градиент - просто напросто производная ($f'(x) = x$), модуль которой не ограничен сверху, а дисперсия в точности равна 0.

[1] Вообще говоря, во многих практически важных задачах второй момент стохастического градиента не ограничен константой и даже дисперсия не ограничена некоторой константой. Тем не менее долгое время для `SGD` существовал анализ сходимости только в таких предположениях.

Комментарий: в условии в явном виде не указано, должно ли выполняться условие, что $m > 1$ или должны ли быть функции разные, может я это упустил где-то. Если такие ограничения действительно есть, то можно взять $m = 2$, $f_1 = x^2, f_2 = x^2 + 1$, получим похожую ситуацию.

3. (1 балл) Как сводить одномерный поиск на полуинтервале, возникающий при использовании метода наискорейшего спуска для выпуклой функции, к одномерному поиску на отрезке? Предложите алгоритм, решающий эту задачу и оцените число итераций, необходимое процедуре одномерного поиска, если необходимая точность по аргументу равняется ε.

   Решение: Попробуем свести указанную задачу к задаче поиска на отрезке. Пусть у нас зафиксированно a - начало полуинтервала $[a; +\infty)$, которое будет одним из концов будущего отрезка. Пусть в точке a значение производной меньше нуля, из этого можем сделать вывод, что минимум лежит правее (рассуждения для ситуации, когда ограничивает исходный полуотрезок с другой стороны и значение производной положительное абсолютно аналогичны). Выберем $\forall b \in [a; +\infty)$, $b = a + step, step > 0$. Если знак производной изменился - мы нашли целевой конец b отрезка $[a; b]$, и задача сводится к поиску минимума на отрезке . Иначе - возьмем новую точку, увеличив наш шаг *step*. Будем повторять данные действия это до тех пор, пока не поменяется знак производной. Оценим количество итераций, которое может нам потребоваться. Мы точно можем сказать, что нам нужно не меньше итераций, чем в случае поиска минимума на отрезке, следовательно, снизу мы ограничены значением, которое было приведено на лекции. В случае поиска минимума на полуотрезке мы, очевидно, зависим от того, как быстро мы найдем второй конец отрезка b. А это уже зависит от того, как далеко расположена точка a от точки минимума x_{min} ($|x_{min} - a|$), и того, как мы выбираем шаг *step* и увеличиваем его, в случае, если мы не смогли найти b. Однако, мы не можем заранее "почувствовать" и понять, какой шаг выбрать и как далеко мы от минимума, следовательно, вообще говоря, в данном случае мы никак не ограничены сверху для $\forall \epsilon$. Пример: пусть мы получили за N шагов точность ϵ, а за M шагов мы нашли точку b, причем b - минимальная из возможных точек, которая может быть концом отрезка. Возьмем вместо $step_i$ значение $\frac{step_i}{2}$, $i \in 1, ..., M$. Очевидно, что шагов нам потребуется больше, чем раньше, следовательно $M' > M$. (в худшем случае мы попадем в минимум раньше b). Однако, очевидно, что мы так же могли и сделать шаг $step = \sum_{i=1}^{M} step_i$, решив задачу поиска b за одну итерацию. Таким образом, ограничений на количество числа итераций нет из-за неопределенности выбора шага, хотя решение и существует.

4. (1 балл) Объясните, почему метод центров тяжести трудно эффективно реализовать на практике.
   Решение: На данный момент не существует эффективного способа вычисления центра тяжести c_t[2]. На лекции говорилось о существовании алгоритма сложности порядка $O(n^6)$. Следовательно, имеет место применимость данного алгоритма для случаев, когда значение n мало. В целом же, поиск объема множества вообще является сложной операцией[3].

[2] Convex Optimization: Algorithms and Complexity,Sébastien Bubeck
[3] Computing the Volume is Difficult, Imre Bfirfiny and Zoltfin Fiiredi

5. (2+2 балла) Рассмотрим задачу

$$f(x) \longrightarrow \min_{x \in Q \subseteq \mathbb{R}^n}, \tag{2}$$

где Q — выпуклое замкнутое подмножество $\mathbb{R}^n$, функция $f(x)$ выпукла и дифференцируема на Q, причём градиент f — Липшицева на множестве Q функция с константой Липшица $L > 0$, т.е. для всех $x, y \in Q$ выполняется неравенство

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

(a) Докажите[4], что для всех $x, y \in Q$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2.$$

Решение:
$\langle \nabla f(x+t(y-x)), y-x\rangle = \langle \nabla f(x)+z(t), y-x\rangle$, $\|z(t)\| \leq Lt\|y-x\| \Rightarrow \langle \nabla f(x+t(y-x)), y-x\rangle = \langle \nabla f(x), y-x\rangle + \langle \nabla z(t), y-x\rangle \leq \langle \nabla f(x), y-x\rangle + \langle \frac{y-x}{\|y-x|} Lt\|y-x\|, y-x\rangle = \langle \nabla f(x), y-x\rangle + Lt\|y-x\|^2$, получаем, что: $f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + t(y-x)), y-x\rangle\, dt \leq \int\limits_0^1 \langle \nabla f(x), y-x\rangle + Lt\|y-x\|^2 dt = \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|x-y\|^2$

(b) Для решения (2) на семинаре был рассмотрен метод проекции градиента

$$x^{k+1} = \pi_Q\left(x^k - \frac{1}{L}\nabla f(x^k)\right),$$

где $\pi_Q(\cdot)$ определяется как

$$\pi_Q(y) = \operatorname*{argmin}_{x \in Q} \|x - y\|_2.$$

Докажите, что

$$\operatorname*{argmin}_{x \in Q}\left\{f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|_2^2\right\} = \pi_Q\left(x^k - \frac{1}{L}\nabla f(x^k)\right).$$

Решение:
$\|x - (x^k - \frac{1}{L}\nabla f(x^k))\|^2 = \|(x - x^k) + \frac{1}{L}\nabla f(x^k))\|^2 = \|(x-x^k)\|^2 + \frac{1}{L^2}\|\nabla f(x^k))\|^2 + 2\langle x - x^k, \frac{1}{L}\nabla f(x^k)\rangle = [(f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|^2) f(x^k)]\frac{2}{L}$

$F(x) = \|x - (x^k - \frac{1}{L}\nabla f(x^k))\|^2, G(x) = f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|^2, \pi_Q\left(x^k - \frac{1}{L}\nabla f(x^k)\right) = \operatorname*{argmin}_{y \in Q} F(y) = \operatorname*{argmin}_{y \in Q} G(y) = \operatorname*{argmin}_{y \in Q}\left\{f(x^k) + \langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}\|y - x^k\|^2\right\}$

[4] *Подсказка:* воспользуйтесь представлением $f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle\, dt$.