Рогачев Александр. Заочно
Домашнее задание №4.
Проксимальные методы
SGD

Во всех задачах предполагается, что в задаче

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n} \tag{1}$$

функция $f(x)$ является μ-сильно выпуклой, $R(x)$ — правильная замкнутая выпуклая функция, x^* — точка минимума функции $F(x)$.

1. (1 балл) Докажите, что $x^* = \text{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$, где $\gamma > 0$.
   Решение :
   Пусть $x' = \text{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$.
   Воспользуемся определением.
   $x' = \underset{y}{\text{argmin}} \{R(y) + \frac{1}{2\gamma}\|y - x^* + \gamma \nabla f(x^*)\|^2\} = \underset{y}{\text{argmin}} \{R(y) + f(x^*) + \nabla f(x^*)^T(y - x^*) + \frac{1}{2\gamma}\|y - x^*\|^2\}$. Заметим, что при $y = x^*$ часть суммы $R(y) + f(x^*)$ достигает минимума, так как x^* — точка минимума функции $F(x)$, а в остальных слагаемых получаем минимум, так как $y - x^*$ обращается в ноль. Делаем вывод - интересующий нас аргминимум - x^*, т.е. $x' = x^*$, чтд.

2. (1 балл) Пусть $R(x) = \gamma g(x/\gamma)$ для некоторой замкнутой правильной выпуклой функции g и $\gamma > 0$. Докажите, что $\text{prox}_R(x) = \gamma \, \text{prox}_{g/\gamma}(x/\gamma)$.
   Решение:

   (a) $\text{prox}_R(x) = \underset{y}{\text{argmin}} \{\gamma g(y/\gamma) + \frac{1}{2}\|y - x\|^2\}$

   (b) $\gamma \, \text{prox}_{g/\gamma}(x/\gamma) = \gamma \, \underset{y}{\text{argmin}} \{g(y) + \frac{\gamma}{2}\|y - x/\gamma\|^2\}$

   Пусть y' - y, минимизирующий выражение (b). Тогда, если домножить y' на γ и подставить в (a), то мы получим под argmin $g(\frac{y'\gamma}{\gamma}) = g(y')$, аналогично выражению (b). Такие же рассуждения и для нормы (грубо говоря, для нормы нам нужно быть максимально близко к x, а y' максимально близок $\frac{x}{\gamma}$ в контексте минимизации выражения (b)). Получаем, что $y'_a = \gamma y'_b$, чтд.

3. (2 балла) Пусть $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n \mid x_i > 0, \ i = \overline{1, n}\}$ и

$$R(x) = \begin{cases} -\gamma \sum\limits_{i=1}^{n} \ln x_i, & x \in \mathbb{R}^n_{++}, \\ +\infty, & \text{иначе}, \end{cases} \quad \text{где } \gamma > 0.$$

Найдите $\text{prox}_R(x)$.
Решение:
$\text{prox}_R(x) = \underset{y}{\text{argmin}} \{-\gamma \sum_{i=1}^{n} \ln y_i + \frac{1}{2}\|y-x\|^2\}$. Пусть $\phi(x) = -\gamma \sum_{i=1}^{n} \ln y_i + \frac{1}{2}\|y-x\|^2$. Найдем $\frac{\partial \phi}{\partial y}$ и приравняем к нулю. Получим $\sum_{i=1}^{n}(-\frac{\gamma}{y_i} + y_i - x_i)$. Получим, что $y_i^2 - y_i x_i - \gamma = 0$. Решим уравнение относительно y_i, получим $y_i = \frac{x_i \pm \sqrt{x_i^2+4\gamma}}{2}$. Из ограничений в условии следует, что в данном выражении следует выбирать корень, соответствующий выражению с плюсом, $y_i = \frac{x_i+\sqrt{x_i^2+4\gamma}}{2}$, y_i - i-ая координата $\text{prox}_R(x)$.

4. (2 балла) Видоизмените доказательство Теоремы 2 из лекции 6, чтобы получить доказательство сходимости проксимального градиентного спуска в терминах функциональных значений F в случае, когда функция f является μ-сильно выпуклой и L-гладкой. *Подсказка:* найдите место в доказательстве, в котором используется выпуклость функции f и воспользуйтесь тем, что функция теперь μ-сильно выпуклая.

   Решение:
   К сожалению, тут скорее размышления, чем строгое решение, ибо победить задачу не удалось. Рискну предположить, что если в случае теоремы, разобранной на лекции, мы убирали дивергенцию, то тут можем сказать, что в силу μ - сильной выпуклости, можем сказать, что $V_d \geq \frac{\mu}{2}\|x^k - x^{k+1}\|_2^2$, и тогда переписав огранечение на $F(x^k) - F(x^k + 1)$ помимо норм из-за L-гладкости останется еще это слагаемое. Далее, при сворачивании телескопической суммы у нас эта компонента никуда не уйдет. Рискну прежполодить, что тогда мы сможем ограничить $\sum(F(x^*) - F(x^{k+1})) \geq -\frac{L}{2}\|x^* - x^0\|_2^2 + \frac{(N-1)\mu}{2}\|x^* - x^{k*}\|_2^2$, где x^{k*} должен быть ближайшим к x^*, но есть ощущение, что я рассуждения неверны.

5. (1+1+2 балла) Рассмотрим задачу (1), в которой

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{2}$$

   где $f_i(x)$ — выпуклые и L-гладкие функции для всех $i = \overline{1,m}$. Пусть есть m компьютеров (их называют *рабочими*), причём все они соединены с одним и тем же сервером (который называют *мастером*), но между собой не соединены. Такая архитектура сети в оптимизации называется *параллельной*. Предположим, что мы хотим решать задачу (1) в такой архитектуре обычным проксимальным градиентным спуском с шагом γ. В таком случае, на каждой итерации i-й рабочий должен вычислить $\nabla f_i(x^k)$ (для всех $i = \overline{1,m}$), затем отправить $\nabla f_i(x^k)$ мастеру, который в свою очередь вычисляет среднее арифметическое от полученных векторов, то есть вычисляет $\nabla f(x^k)$, вычисляет $x^{k+1} = \text{prox}_{\gamma R}(x^k - \gamma \nabla f(x^k))$ и отправляет x^{k+1} всем рабочим. В описанной процедуре большая нагрузка ложится на мастера, поскольку m и n могут быть большими и мастеру придётся принять (а рабочим передать) за одну итерацию очень большой объём информации (битов).

   Разумная попытка по устранению этого недостатка состоит в том, чтобы вместо $\nabla f_i(x^k)$ отправлять несмещённую оценку g_i^k, которая будет иметь меньше битов информации, чем $\nabla f_i(x^k)$. Один из таких способов — это квативация/спарсификация.

Определение 1 (Квантизация/Спарсификация)**.** Будем называть стохастический оператор $Q(x)$ оператором квантизации или просто квантизацией, если для любого $x \in \mathbb{R}^n$ выполняется:

$$\mathbb{E}[Q(x)] = x, \quad \mathbb{E}[\|Q(x) - x\|_2^2] \leq \omega \|x\|_2^2, \tag{3}$$

где $\omega \geq 0$.

Легко заметить, что тождественный оператор $Q(x) = x$ удовлетворяет Определению 1 с константой $\omega = 0$.

(a) (1 балл) Покажите, что стохастический оператор

$$\text{rand}_t(x) = \frac{n}{t} \sum_{i \in S} x_i e_i,$$

где t — некоторое число из множества $\{1, \ldots, n\}$ (количество компонент вектора x, которые мы передаём), S — случайное подмножество множества $\{1, \ldots, n\}$ размера t (подмножество S выбирается случайно и равновероятно среди всех возможных подмножеств размера t), $(e_1, \ldots, e_n)$ — стандартный базис в $\mathbb{R}^n$. Покажите, что данный оператор удовлетворяет Определению 1 с константой $\omega = \frac{n}{t} - 1$. *Подсказка:* для этого воспользуйтесь $\mathbb{E}[\|Q(x)\|_2^2] = \mathbb{E}[\|Q(x) - x\|_2^2] + \|x\|_2^2$.

Решение:
Пусть $y = rand_t(x)$. Тогда $y_i = 0$, если $i \notin S$, и $y_i = x_i \frac{n}{t}$, если $i \in S$.S - случайное подмножество размера t, вероятность, что данный индекс i входит в подмножество случайного размера t - t/n. Т.е. $\mathbb{E}[y_i] = \frac{t}{n}(x_i \frac{n}{t}) + (1 - \frac{t}{n})0 = x_i$. Это верно для $\forall i$, т.е. $\mathbb{E}[rand_t(x)] = x$.
Далее, $\mathbb{E}[\|Q(x) - x\|_2^2] = \mathbb{E}[\|Q(x)\|_2^2] - \|x\|_2^2 \leq \|\mathbb{E}Q(x)\|_2^2 - \|x\|_2^2 \leq \frac{n}{t}\|x\|_2^2 - \|x\|_2^2 = (\frac{n}{t} - 1)\|x\|_2^2, \to \omega = \frac{n}{t} - 1$

(b) (1 балл) Рассмотрим следующий оператор, который называют трёхуровневой квантизацией:

$$[Q(x)]_i = \|x\|_2 \text{sign}(x_i)\xi_i, \quad i = 1, \ldots, n,$$

где $[Q(x)]_i$ — i-я компонента вектора $Q(x)$ и ξ_i — случайная величина, имеющая распределение Бернулли с параметром $\frac{|x_i|}{\|x\|_2}$, то есть

$$\xi_i = \begin{cases} 1 & \text{с вероятностью } \frac{|x_i|}{\|x\|_2}, \\ 0 & \text{с вероятностью } 1 - \frac{|x_i|}{\|x\|_2}. \end{cases}$$

Таким образом, если мы хотим передать вектор $Q(x)$, то нам нужно передать вектор, состоящий из нулей и ± 1, и вещественное число $\|x\|_2$, причём вероятность обнуления компоненты тем больше, чем компонента меньше по модулю по сравнению с остальными компонентами. Покажите, что данный оператор удовлетворяет Определению 1 с параметром $\omega = \sqrt{n} - 1$.

Решение:
Аналогично, рассмотрим математическое ожидание i-ой компоненты $Q(x)_i$. ξ_i —

случайная величина, имеющая распределение Бернулли с параметром $\frac{|x_i|}{\|x\|_2}$, то есть $\mathbb{E}\xi_i = \frac{|x_i|}{\|x\|_2}$. Тогда $\mathbb{E}[Q(x)_i] = \|x\|_2 \text{sign}(x_i)\frac{|x_i|}{\|x\|_2} = \text{sign}(x_i)|x_i| = x_i$. Далее, $\mathbb{E}[\|Q(x) - x\|_2^2] = \mathbb{E}[\|Q(x)\|_2^2] - \|x\|_2^2 \leq \sqrt{n}|x| - \|x\|_2^2 = (\sqrt{n} - 1)\|x\|_2^2$. $\to \omega = \sqrt{n} - 1$

(c) (2 балла) Рассмотрим проксимальный квантизированный градиентный спуск.

Algorithm 1 Проксимальный квантизированный градиентный спуск (`prox-QGD`)

Require: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N

for $k = 0, 1, \ldots, N-1$ **do**
  Отправить x^k всем рабочим ▷ выполняется мастером
  for $i = 1, \ldots, m$ параллельно **do**
    Принять x^k от мастера ▷ выполняется рабочими
    Вычислить $\nabla f_i(x^k)$ ▷ выполняется рабочими
    Независимо от других рабочих сгенерировать $g_i^k = Q(\nabla f_i(x^k))$ ▷ вып-ся р-ми
    Отправить g_i^k мастеру ▷ выполняется рабочими
  end for
  Принять g_i^k от всех рабочих ▷ выполняется мастером
  Вычислить $g^k = \frac{1}{m}\sum_{i=1}^m g_i^k$ ▷ выполняется мастером
  $x^{k+1} = \text{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$ ▷ выполняется мастером
end for

Ensure: x^N

Покажите, что `prox-QGD` удовлетворяет Предположению 1 из лекции 7 с параметрами

$$A = L\left(1 + \frac{2\omega}{m}\right), \quad D_1 = \frac{2\omega}{m^2}\sum_{i=1}^m \|\nabla f_i(x^*)\|_2^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Используя Теорему 1, доказанную на лекции, оцените скорость сходимости `prox-QGD` в сильно выпуклом случае. *Подсказка:* не забудьте воспользоваться независимостью $g_1^k, g_2^k, \ldots, g_m^k$.