Метод сопряженных градиентов. 25 марта 2020 г.

Семинарист: Данилова М.

Введение

Градиентный метод: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

Главный недостаток - медленная скорость сходимости!

Ускорения градиентного метода

1. **Наискорейший спуск**

   на каждом шаге решаем задачу одномерной минимизации, идем до минимума по направлению антиградиента

   $$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$
   $$\alpha_k = \arg\min_{\alpha > 0} f(x_k - \alpha \nabla f(x_k))$$

2. **Многошаговые методы**

   в градиентном методе на каждом шаге никак не используется информация, полученная нп предыдущих итерациях. Естесственно попытаться учесть предысторию процесса для ускорения сходимости. Такого рода методы, в которых направление зависит от s предыдущих:

   $$x_{k+1} = \varphi_k(x_k, .., x_{k-s+1})$$

   называются s-шаговыми. Градиентный метод и метод Ньютона были одношаговыми, теперь рассмотрим многошаговые (s $>$ 1) для решения задачи безусловной минимизации

   $$\min_{x \in \mathbb{R}^n} f(x), \quad f(x) \in \mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$$

   (a) **Метод тяжелого шарика**

   $$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$
   $$\alpha > 0,\ \beta \geq 0 - \text{параметры}$$

   (b) **Метод Нестерова (быстрый градиентный метод)**

   $$\{x_k\}, \{y_k\} - \text{строим две последовательности}$$
   $$x_{k+1} = y_k - \frac{1}{L}\nabla f(x_k)$$
   $$y_{k+1} = x_{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x_{k+1} - x_k)$$

   (c) **Метод сопряженных градиентов**

   $$x_{k+1} = x_k + \alpha_k h_k$$
   $$h_k = -\nabla f(x_k) + \beta_k h_{k-1}$$
   $$\beta_0 = 0$$

   α_k, β_k - ?

   - $\alpha_k = \arg\min_{\alpha} f(x_k + \alpha h_k)$
   - β_k - разные способы (см.ниже)

Метод сопряженных градиентов

Метод сопряженных градиентов для квадратичных функций

Методы сопряженных градиентов (МСГ) были изначально предложены для минимизации квадратичных функций. Рассмотрим задачу

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

где $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,\ A = A^\top \succeq 0$, которая является самой характерной задачей выпуклой оптимизации. Изучая данный класс задач, можно пытаться понять локальную сходимость в выпуклых задачах. Так же такие задачи возникают в виде подзадач, например в методе Ньютона. Как известно, решение этой задачи есть $x^* = -A^{-1}b$. Поэтому нашу целевую функцию можно переписать в следующем виде:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c = \frac{1}{2}\langle Ax, x\rangle - \langle Ax^*, x\rangle + c =$$

$$= \frac{1}{2}\langle A(x - x^*), x - x^*\rangle - \frac{1}{2}\langle Ax^*, x^*\rangle + c.$$

То есть,

$$f^* = f(x^*) = c - \frac{1}{2}\langle Ax^*, x^*\rangle, \quad \nabla f(x) = A(x - x^*).$$

Предположим, что нам задана начальная точка x_0.

Рассмотрим **линейные подпространства Крылова**

$$\mathscr{L}_k = \text{Lin}\left\{A(x_0 - x^*), .., A^k(x_0 - x^*)\right\},\ k \geq 1,$$

где A^k - k-я степень матрицы A. Последовательность точек $\{x_k\}$, образованная **методом сопряженных градиентов**, определяется следующим образом:

$$x_k = \text{argmin}\{f(x) \mid x \in x_0 + \mathscr{L}_k\},\ k \geq 1.$$

Лемма 1. Для любого $k \geq 1$ имеет место равенство

$$\mathscr{L}_k = \text{Lin}\left\{\nabla f(x_0), .., \nabla f(x_{k-1})\right\}$$

Доказательство. Для $k = 1$ утверждение верно: $\nabla f(x_0) = A(x_0 - x)$. Предположим, что оно также выполнено для некоторого $k \geq 1$. Тогда

$$x_k = x_0 + \sum_{i=1}^{k} \lambda^{(i)} A^i (x_0 - x^*)$$

с некоторыми множителями $\lambda \in R^k$. Поэтому

$$\nabla f(x_k) = A(x_0 - x^*) + \sum_{i=1}^{k} \lambda^{(i)} A^{i+1}(x_0 - x^*) = y + \lambda^{(k)} A^{k+1}(x_0 - x^*)$$

для некоторой точки y из $\mathscr{L}_k$. Таким образом,

$$\mathscr{L}_{k+1} \equiv \text{Lin}\left\{\mathscr{L}_k, A^{k+1}(x_0 - x^*)\right\} = \text{Lin}\left\{\mathscr{L}_k, \nabla f(x_k)\right\} = \text{Lin}\left\{\nabla f(x_0), .., \nabla f(x_k)\right\}.$$

□

Следующая лемма помогает понять поведение последовательности $\{x_k\}$, а именно, что **градиенты на точках последовательности $\{x_k\}$, генерируемой МСГ, ортогональны.**

Лемма 2. Для любых $k, i \geq 0, k \neq i$ имеет место равенство $\langle \nabla f(x_k), \nabla f(x_i)\rangle = 0$.

Доказательство. Пусть $k > i$. Рассмотрим функцию

$$\varphi(\lambda) = f\left(x_0 + \sum_{j=1}^{k} \lambda^{(j)} \nabla f(x_{j-1})\right),\ \lambda \in \mathbb{R}^k.$$

В силу леммы 1 для некоторого λ_* имеем $x_k = x_0 + \sum\limits_{j=1}^{k} \lambda_*^{(j)} \nabla f(x_{j-1})$. Однако по определению x_k есть точка минимума функции $f(x)$ на $\mathscr{L}_k$. Поэтому $\nabla\varphi(\lambda_*) = 0$. Остается вычислить компоненты этого вектора:

$$0 = \frac{\partial\varphi(\lambda_*)}{\partial\lambda^{(i)}} = \langle \nabla f(x_k), \nabla f(x_i)\rangle .$$

□

Следствие 1. Последовательность, образованная методом сопряженных градиентов для задачи (1) конечна.

Следствие 2. Для любого $p \in \mathscr{L}_k$ верно равенство $\langle \nabla f(x_k), p\rangle = 0$.

Итак, последний вспомогательный результат объясняет название метода.

Обозначим $h_i = x_{i+1} - x_i$. Очевидно, что $\mathscr{L}_k = \mathrm{Lin}\{h_0, .., h_{k-1}\}$.

Лемма 3. Для любого $k \neq i$ верно равенство $\langle Ah_k, h_i\rangle = 0$. Такие направления называются **сопряженными относительно матрицы A**.

Перепишем метод сопряженных градиентов в алгоритмической форме. Так как $\mathscr{L}_k = \mathrm{Lin}\{h_0, .., h_{k-1}\}$, можно представить x_{k+1} в виде

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \sum_{j=0}^{k-1} \lambda^{(j)} h_j.$$

В наших обозначениях получаем

$$h_k = -\alpha_k \nabla f(x_k) + \sum_{j=0}^{k-1} \lambda^{(j)} h_j. \tag{2}$$

Вычисляя коэффициенты, умножая (2) на A и h_i, где $0 \le i \le k-1$, и используя леммы 3 и 2 имеем

$$x_{k+1} = x_k - \alpha_k h_k,$$

где

$$h_k = \nabla f(x_k) - \frac{\|\nabla f(x_k)\|^2}{\langle \nabla f(x_k) - \nabla f(x_{k-1}), h_{k-1}\rangle} h_{k-1}$$

Подробности см. в книге Ю. Е. Нестерова “Введение в выпуклую оптимизацию”

В итоге для МСГ мы получаем следующий алгоритм согласно которому проводятся вычисления:

Algorithm 1 Метод сопряженных градиентов для квадратичных функций

1: Пусть $x_0 \in R^n$. Вычислим $f(x_0), \nabla f(x_0)$. Положим $h_0 = \nabla f(x_0)$
2: k-я итерация ($k \ge 0$)

- Найдем $x_{k+1} = x_k + \alpha_k h_k$ с помощью точного одномерного поиска:

$$\alpha_k = \underset{\alpha}{\mathrm{argmin}}\, f(x_k + \alpha h_k) = -\frac{\langle \nabla f(x_k), h_k\rangle}{\langle h_k, Ah_k\rangle}$$

- Вычислим $f(x_{k+1})$ и $\nabla f(x_{k+1})$.
- Вычислим коэффициент β_k:

$$\beta_k = \frac{\langle \nabla f(x_{k+1}), Ah_k\rangle}{\langle h_k, Ah_k\rangle}$$

- Положим $h_{k+1} = \nabla f(x_{k+1}) - \beta_k h_k$.

Замечания:

1. направление h_k - это линейная комбинация антиградиента и предыдущего направления
2. "соседние"направления h_k и h_{k-1} A-сопряженны, т.е. $\langle h_k, Ah_{k-1}\rangle = 0$ (из этого условия выбирается число β_{k-1})
3. $\forall k, i \geq 0,\ k \neq i$ градиенты $\nabla f(x_k)$ и $\nabla f(x_i)$ ортогональный друг другу, т.е. $\langle \nabla f(x_k), \nabla f(x_i)\rangle = 0$
4. на k-ом шаге мы находимся в точке минимума нашей функции на подпространстве, порожденном предыдущими градиентами
5. МСГ - метод первого порядка
6. при минимизации сильно выпуклых функций обладает сверхлинейной и даже квадратичной скоростью сходимости

Теорема 1. Если выпуклая квадратичная функция достигает своего минимального значения на $\mathbb{R}^n$, то метод сопряженных градиентов находит точный минимум не более чем за n шагов.

Метод сопряженных градиентов для произвольных функций

Выбор α_k:

$$\alpha_k = \underset{\alpha}{\operatorname{argmin}} f(x_k + \alpha h_k) \text{ - точный одномерный поиск.}$$

Выбор β_k:

1. $$\beta_k = \frac{\|\nabla f(x_{k+1})\|^2}{\langle \nabla f(x_{k+1}) - \nabla f(x_k), h_k\rangle}$$
2. формула Флетчера-Ривса:
$$\beta_k = -\frac{||\nabla f(x_{k+1})||^2}{||\nabla f(x_k)||^2}$$
3. формула Полака-Рибьера:
$$\beta_k = -\frac{\langle \nabla f(x_{k+1}), \nabla f(x_{k+1}) - \nabla f(x_k)\rangle}{\|\nabla f(x_k)\|^2}$$

Все они дают одинаковый результат на квадратичных функциях, но в общем случае образуют разные последовательности.

2ой и 3ий варианты наиболее часто используются на практике, так же используется стратегия обновления, которая в определенный момент устанавливает $\beta_k = 0$ (обычно после каждой n-ой итерации).

Отмечается преимущество МСГ в эффективности решения задач БМ. МСГ – наиболее часто применяемый методы решения задач БМ на классе дважды непрерывно дифференцируемых ограниченных снизу функций. Скорость сходимости МСГ зависит от свойств целовой функции. В окрестности точки строгого минимума схемы МСГ имеют локальную квадратичную сходимость. В общем случае установлен только факт глобальной сходимости для задач гладкой выпуклой оптимизации, скорость которого не лучше, чем у градиентного.

Оценки сходимости

	μ-сильно выпуклая и L-гладкая f	выпуклая и L-гладкая f	μ-сильно выпуклая f и $\|\nabla f(x)\|_2 \leq M$	выпуклая f и $\|\nabla f(x)\|_2 \leq M$
Нижние оценки	$\Omega\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)$	$\Omega\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$\Omega\left(\frac{M^2}{\mu\varepsilon}\right)$	$\Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$
Градиентный метод	$O\left(\frac{L}{\mu}\ln\left(\frac{LR^2}{\varepsilon}\right)\right)$	$O\left(\frac{LR^2}{\varepsilon}\right)$	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$
Метод Нестерова	$O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{LR^2}{\varepsilon}\right)\right)$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Таблица 1: Оценки на число подсчётов градиента (число итераций) для детерминированных методов, гарантирующие $f(\hat{x}) - f(x^*) \leq \varepsilon$ через указанное число подсчётов градиентов (итераций), где $\hat{x}$ — точка, которую возвращает метод, $R = \|x^0 - x^*\|_2$, где x^0 — стартовая точка. В последних двух столбцах достаточно потребовать ограниченность градиентов только в шаре с центром в x^* и радиусом $2R$.