Оптимальные градиентные методы

Мальковский Н. В.

Санкт-Петербургский академический университет

Недостаток градиентного спуска

Очевидным образом градиентный спуск обладает следующим свойством:

$$x_k \in x_0 + \mathrm{Span}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\} \tag{1}$$

так как

$$x_k = x_{k-1} - \alpha_k \nabla f(x_k) = \ldots = x_0 - \sum_{i=0}^{k-1} \alpha_i \nabla f(x_i)$$

Если ∇f липшицев с константой M, то градиентный спуск достигает оценки

$$f(x_k) - f(x^*) = \mathcal{O}\left(\frac{1}{k}\right)$$

Если при этом f сильно выпукла с константой m, то

$$f(x_k) - f(x^*) = \mathcal{O}\left(\left(\frac{M-m}{M+m}\right)^k\right)$$

Оказывается, что в обоих случаях градиентный спуск не является оптимальным алгоритмом среди удовлетворяющих (1).

Нижние оценки для градиентных методов

Теорема

$\forall k\ \exists n, x_0, f : \mathbb{R}^n \to \mathbb{R}$, *такие, что* $\nabla^2 f \preceq MI$, *любой метод, удовлетворяющий* (1), *примененный к* f *имеет нижнюю оценку*

$$f(x_k) - f(x^*) \geq \frac{\beta M ||x_k - x^*||}{(k+1)^2},$$

где $\beta > 0$ – *некоторая константа, не зависящая от* k, n, M, f, x_0.

Теорема

$\forall k\ \exists n, x_0, f : \mathbb{R}^n \to \mathbb{R}$, *такие, что* $mI \preceq \nabla^2 f \preceq MI$, *любой метод, удовлетворяющий* (1), *примененный к* f *имеет нижнюю оценку*

$$f(x_k) - f(x^*) \geq \frac{m}{2} \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^{2k} ||x_0 - x^*||^2.$$

Замечания по нижним оценкам

В случае выпуклой функции нижняя оценка дает $f(x_k) - f(x^*) = \mathcal{O}(1/k^2)$, из чего следует, что необходимое кол-во итераций градиентного метода для достижения ϵ-аппроксимации есть

$$k(\epsilon) = \mathcal{O}(1/\sqrt{\epsilon}),$$

в то время как градиентный спуск гарантирует только $k(\epsilon) = \mathcal{O}(1/\epsilon)$. В случае сильно выпуклой функции из нижней оценки следует, что необходимое кол-во итераций градиентного метода для достижения ϵ-аппроксимации есть

$$k(\epsilon) \geq \frac{\log \frac{m||x_0 - x^*||^2}{2\epsilon}}{-2\log\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)} \tag{2}$$

Как и нижняя оценка, так и сам градиентный спуск гарантируют, что $k(\epsilon) = \mathcal{O}(\log\frac{1}{\epsilon})$. Если же смотреть на k как функцию не только ϵ, но и m, M, то при больших M/m имеем

$$\log\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right) \sim -2\frac{\sqrt{m}}{\sqrt{M}+\sqrt{m}} \sim -2\frac{\sqrt{m}}{\sqrt{M}}$$

Замечания по нижним оценкам

Таким образом, для нижних оценок мы получаем, что

$$k(\epsilon, m, M) \geq \frac{\log \frac{m||x_0 - x^*||^2}{2\epsilon}}{-2\log\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)} \geq \frac{1}{4}\left(\frac{\sqrt{M}}{\sqrt{m}} - 1\right) \log \frac{m||x_0 - x^*||^2}{2\epsilon}$$

Второе неравенство получено из $\log(1+x) \leq x$. Так как $\log(1+x) \sim_0 x$, то неравенство является асимптотически оптимальным при $m/M \to 0$.

Используя аналогичное рассуждение, получаем, что для градиентного спуска

$$k(\epsilon, m, M) \geq \frac{1}{4}\left(\frac{M}{m} - 1\right) \log \frac{M||x_0 - x^*||^2}{2\epsilon}$$

что хуже относительно параметров m, M

Оценивающие последовательности

Определение

Последовательности $\{\phi(\cdot)\}_{k=0}^{\infty}$ и $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$ называются *оценивающими последовательностями* *функции $f(\cdot)$, если для любого $x \in \mathbb{R}^n$ и всех $k \geq 0$ выполняется неравенство*

$$\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k \phi_0(x).$$

Если $\phi_k(\cdot), \lambda_k$ – оценивающие последовательности функции f и для последовательности x_k выполняется

$$f(x_k) \leq \phi_k^* = \min_{x \in \mathbb{R}^n} \phi_k(x),$$

то

$$\begin{aligned} f(x_k) &\leq \min_{x \in \mathbb{R}^n} \phi_k(x) \\ &\leq \min_{x \in \mathbb{R}} (1 - \lambda_k)f(x) + \lambda_k \phi_0(x) \\ &\leq (1 - \lambda_k)f(x^*) + \lambda_k \phi_0(x^*), \end{aligned}$$

из чего следует $f(x_k) - f(x^*) \leq \lambda_k(\phi_0(x^*) - f(x^*))$.

Оценивающие последовательности

Лемма

Пусть f – дифференцируемая функция, сильно выпуклая с константой m, ∇f липшицев с константой M, $\phi_0(\cdot)$ – произвольная функция, $\{y_k\}_{k=0}^{\infty}$ – произвольная последовательность в $\mathbb{R}^n$, $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\lambda_0 = 1$, то последовательности, заданные рекурентно по правилу

$$\begin{aligned} \lambda_{k+1} &= (1-\alpha_k)\lambda_k \\ \phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k \left(f(y_k) + \nabla f(y_k)(x-y_k) + \frac{m}{2}||x-y_k||^2 \right) \end{aligned} \tag{3}$$

образуют оценивающие последовательности.

Док-во. Очевидным образом $\phi_0(x) = (1-\lambda_0[=1])f(x) + \lambda_0[=1]\phi_0(x)$. Воспользуемся индукцией: пусть выполняется $\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x)$, тогда

$$\begin{aligned} \phi_{k+1}(x) &\leq (1-\alpha_k)\phi_k(x) + \alpha_k f(x) \\ &\leq (1-\alpha_k)((1-\lambda_k)f(x) + \lambda_k\phi_0(x)) + \alpha_k f(x) \\ &= (1-\lambda_{k+1})f(x) + \lambda_{k+1}\phi_0(x). \quad \blacksquare \end{aligned}$$

Оценивающие последовательности

Лемма

Если оценивающие последовательности генерируются по правилу (3) и при этом $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}||x - x_0||^2$ *то функции* $\phi_k(\cdot)$ *имеют вид* $\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}||x - v_k||^2$ *и выполняются соотношения*

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k m$$

$$v_{k+1} = \frac{1}{\gamma_{k+1}}((1-\alpha)\gamma_k v_k + \alpha_k m y_k - \alpha_k \nabla f(y_k))$$

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}||\nabla f(y_k)||^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{m}{2}||y_k - v_k||^2 + \nabla f(y_k)(v_k - y_k)\right). \qquad (4)$$

Оценивающие последовательности

Док-во. Во-первых, $\nabla^2\phi_0(x) = \gamma_0 I$, докажем по индукции, что $\nabla^2\phi_k(x) = \gamma_k I$:

$$\nabla^2\phi_{k+1}(x) = (1-\alpha_k)\nabla^2\phi_k(x) + \alpha_k m I = ((1-\alpha_k)\gamma_k + \alpha_k m) I = \gamma_{k+1} I$$

следовательно $\phi_k(x)$ действительно представляется в виде $\phi_k^* + \frac{\gamma_k}{2}||x - v_k||^2$.

Далее, v_k – единственная (в силу $\gamma_k > 0$) точка минимума $\phi_k(x)$, а значит удовлетворяет условиям оптимальности первого порядка:

$$\begin{aligned}
0_n &= \nabla\phi_{k+1}(v_{k+1}) \\
&= \nabla_x \left[(1-\alpha_k)\left(\phi_k^* + \frac{\gamma_k}{2}||x - v_k||^2\right) \right. \\
&\quad \left. +\alpha_k \left(f(y_k) + \nabla f(y_k)^T (x - y_k) + \frac{m}{2}||x - y_k||^2 \right) \right] |_{x=v_{k+1}} \\
&= (1-\alpha_k)\gamma_k(v_{k+1} - v_k) + \alpha_k \nabla f(y_k) + \alpha_k m(v_{k+1} - y_k),
\end{aligned}$$

что дает рекуррентное соотношение для v_k.

В силу соотношений для для v_k, γ_k

$$v_{k+1} - y_k = \frac{1}{\gamma_{k+1}}[(1 - \alpha_k)\gamma_k(v_k - y_k) - \alpha_k \nabla f(y_k)].$$

Подставив y_k в ϕ_k получаем

$$\phi^*_{k+1} + \frac{\gamma_{k+1}}{2}||y_k - v_{k+1}||^2 = (1 - \alpha_k)\left(\phi^*_k + \frac{\gamma_k}{2}||y_k - v_k||^2\right).$$

Осталось подставить указанное выше соотношение для $v_{k+1} - y_k$ для получения соотношения на v_k. ■

Вывод общей схемы

Таким образом на данный момент, используя оценивающие последовательности, сгенерированные по правилу (3) с $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}||x - x_0||^2$ имеем

- Свободу выбора γ_0, α_k, y_k
- Величины γ_k $(k > 0)$, λ_k, v_k, ϕ_k^* однозначно определены этим выбором
- Для получения рабочей схемы нужно научиться делать выбор так, чтобы гарантировать

$$f(x_k) \leq \phi_k^*.$$

Предположим, что у нас есть $x_k : \ f(x_k) \leq \phi_k^*$. Из (4) получаем

$$\phi_{k+1}^* \geq (1 - \alpha_k)f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}||\nabla f(y_k)||^2$$
$$+ \frac{\alpha_k(1 - \alpha_k)\gamma_k}{\gamma_{k+1}} \left(\frac{m}{2}||y_k - v_k||^2 + \nabla f(y_k)^T(v_k - y_k)\right).$$

Вывод общей схемы

Упростим это неравенство воспользовавшись $||y_k - v_k||^2 \geq 0$ и $f(x_k) \geq f(y_k) + \nabla f(y_k)(x_k - y_k)$:

$$\phi^*_{k+1} \geq f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}||\nabla f(y_k)||^2 + (1-\alpha_k)\nabla f(y_k)\left(\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\right)$$

- y_k можно выбрать таким образом, чтобы обнулить $\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k$
- α_k можно выбрать так, чтобы выполнялось $\frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2M}$
- Наконец, если при этом выбрать $x_k = y_k - \frac{1}{M}\nabla f(y_k)$, то $f(x_{k+1}) \leq f(y_k) - \frac{1}{2M}||\nabla f(y_k)||^2$, что влечет $\phi^*_{k+1} \geq f(x_{k+1})$

Вывод общей схемы

При $M\alpha_k^2 = \gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k m$, α_k задается квадратным уравнением

$$Mx^2 + (\gamma_k - m)x - \gamma_k = 0.$$

Отметим, что значение левой части в 0 и 1 есть $-\gamma_k$ и $M - m$ соответственно, что гарантирует наличие корня на интервале $(0, 1)$ при $\gamma_k > 0$.

Для y_k получаем

$$y_k \left(\frac{\alpha_k \gamma_k}{\gamma_{k+1}} + 1 \right) = \frac{\alpha_k \gamma_k}{\gamma_{k+1}} v_k + x_k.$$

Таким образом

$$y_k = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\alpha_k \gamma_k + \gamma_{k+1}} = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + m\alpha_k}$$

Общая схема оптимальных методов

Инициализация.

Выбрать начальное приближение x_0, константу $\gamma_0 > 0$, взять $v_0 = x_0$.

Итерация $k \geq 0$.

1. Вычислить $\alpha_k \in (0, 1)$ из уравнения
$$Mx^2 + (\gamma_k - m)x - \gamma_k = 0,$$
взять $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k m$.
2. Взять
$$y_k = \frac{\alpha_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + m\alpha_k}$$
и вычислить $f(y_k)$, $\nabla f(y_k)$.
3. Выбрать x_{k+1} так, чтобы выполнялось
$$f(x_{k+1}) \leq f(y_k) - \frac{1}{2M} ||\nabla f(y_k)||^2.$$
4. Взять
$$v_{k+1} = \frac{1}{\gamma_{k+1}} [(1 - \alpha)\gamma_k v_k + \alpha_k m y_k - \alpha_k \nabla f(y_k)]$$

Скорость сходимости для общей схемы

Теорема

Общая схема генерирует последовательность оценок x_k, для которой выполняется

$$f(x_k) - f(x^*) \leq \lambda_k \left(f(x_0) - f(x^*) + \frac{\gamma_0}{2} ||x_0 - x^*||^2 \right)$$

Док-во. По построению при $\phi_0(x^*) = f(x_0) + \frac{\gamma_0}{2} ||x_0 - x^*||^2$.

Теорема

Если в общей схеме $\gamma_0 = M$, то для последовательности x_k выполняется

$$f(x_k) - f(x^*) \leq M \min \left\{ \left(1 - \sqrt{\frac{m}{M}}\right)^k, \frac{4}{(2+k)^2} \right\} ||x_0 - x^*||^2$$

Док-во. Далее (после леммы).

Скорость сходимости общей схемы

Лемма

Если в общей схеме взять $\gamma_0 \geq m$, то

$$\lambda_k = \prod_{i \leq k-1} (1 - \alpha_k) \leq \min \left\{ \left(1 - \sqrt{\frac{m}{M}}\right)^k, \frac{4M}{(2\sqrt{M} + k\sqrt{\gamma_0})^2} \right\}$$

Док-во. Так как $\gamma_0 \geq m$, то по индукции

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k m \geq (1 - \alpha_k)m + \alpha_k m = m.$$

Следовательно

$$\alpha_k = \sqrt{\frac{\gamma_{k+1}}{M}} \geq \sqrt{\frac{m}{M}}.$$

Покажем по индукции, что $\gamma_k \geq \gamma_0 \lambda_k$:

$$\gamma_{k+1} \geq (1 - \alpha_k)\gamma_k \geq (1 - \alpha_k)\gamma_0\lambda_k = \gamma_0\lambda_{k+1}.$$

Скорость сходимости для общей схемы

Далее, обозначим $\omega_k = 1/\sqrt{\lambda_k}$. По построению λ_k убывает, а значит

$$\omega_{k+1} - \omega_k = \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k \lambda_{k+1}}} = \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k \lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})} \geq$$

$$\geq \frac{\alpha_k \lambda_k}{2\lambda_k \sqrt{\lambda_{k+1}}} \geq \frac{1}{2}\sqrt{\frac{\gamma_0}{M}}.$$

Суммируя по $1 \leq i \leq k$ получаем

$$\omega_k = \frac{1}{\sqrt{\lambda_k}} \geq 1 + \frac{k}{2}\sqrt{\frac{\gamma_0}{M}} \quad \blacksquare$$

Док-во. теоремы применить лемму с $\gamma_0 = M$ и воспользоваться $f(x_0) - f(x^*) \leq \frac{M}{2}||x_0 - x^*||^2$.

Упрощение схемы

*Далее считаем, что x_{k+1} генерируется как шаг градиентного метода из y_k, т. е. $x_{k+1} = y_k - \frac{1}{M}\nabla f(y_k)$

Использовав выражение для $y_k = \frac{\alpha_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + \alpha_k m}$ можно получить $\gamma_k v_k = \frac{1}{\alpha_k}[y_k(\gamma_k + \alpha_k m) - \gamma_{k+1}x_k]$, таким образом

$$
\begin{aligned}
v_{k+1} &= \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k m y_k - \alpha_k\nabla f(y_k)] \\
&= \frac{1}{\gamma_{k+1}}\left\{\frac{1-\alpha_k}{\alpha_k}[y_k(\gamma_k + \alpha_k m) - \gamma_{k+1}x_k] + \alpha_k m y_k - \alpha_k\nabla f(y_k)\right\} \\
&= \frac{1}{\gamma_{k+1}}\left[\frac{(1-\alpha_k)\gamma_k}{\alpha_k}y_k + m y_k\right] - \frac{1-\alpha_k}{\alpha_k}x_k - \frac{\alpha_k}{\gamma_{k+1}}\nabla f(y_k) \\
&= \frac{1}{\alpha_k}y_k - \frac{1}{\alpha_k}x_k + x_k - \frac{\alpha_k}{\gamma_{k+1}}\nabla f(y_k) \quad (\gamma_{k+1} = (1-\alpha_k)\gamma_k + m\alpha_k) \\
&= x_k + \frac{1}{\alpha_k}(y_k - x_k) - \frac{1}{\alpha_k M}\nabla f(y_k) \quad (\gamma_{k+1} = M\alpha_k^2) \\
&= x_k + \frac{1}{\alpha_k}(x_{k+1} - x_k) \quad (x_{k+1} = y_k - \frac{1}{M}\nabla f(y_k))
\end{aligned}
$$

Упрощение схемы

Далее

$$y_{k+1} = \frac{1}{\gamma_{k+1} + \alpha_{k+1} m}(\alpha_{k+1}\gamma_{k+1} v_{k+1} + \gamma_{k+2} x_{k+1})$$
$$= x_{k+1} \frac{\alpha_{k+1}\gamma_{k+1}(v_{k+1} - x_{k+1})}{\gamma_{k+1} + \alpha_{k+1} m} = x_{k+1} + \beta_k (x_{k+1} - x_k),$$

где

$$\beta_k = \frac{\alpha_{k+1}\gamma_{k+1}(1 - \alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1} m)}.$$

Таким образом мы избавились от v_k. Теперь избавимся от γ_k:

$$\beta_k = \frac{\alpha_{k+1}\gamma_{k+1}(1 - \alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1} m)} = \frac{\alpha_{k+1}\gamma_{k+1}(1 - \alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1}^2 M - (1 - \alpha_{k+1})\gamma_{k+1})}$$
$$= \frac{\gamma_{k+1}(1 - \alpha_k)}{\alpha_k(\gamma_{k+1} + \alpha_{k+1} M)} = \frac{\alpha_k(1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$$

Для вычисления α_{k+1} можно использовать уравнение $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \frac{m}{M}\alpha_{k+1}$.

Упрощенная схема

Инициализация Выбрать начальное приближение x_0, $\alpha_0 \in (0, 1)$. Взять $y_0 = x_0$.

Итерация $k \leq 0$

1. Вычислить $\nabla f(y_k)$, взять $x_{k+1} = y_k - \frac{1}{M}\nabla f(y_k)$.
2. Вычислить α_{k+1} из уравнения

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \frac{m}{M}\alpha_{k+1}$$

3. Взять

$$\beta_k = \frac{\alpha_k(1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}, \quad y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k).$$

Итоговые замечания

Замечание 1. Условие $\gamma_0 \geq m$ эквивалентно $\alpha_0 \geq \sqrt{\frac{m}{M}}$

Замечание 2. Если взять $\alpha_0 = \sqrt{\frac{m}{M}}$, то

$$\alpha_k = \sqrt{\frac{m}{M}}, \quad \beta_k = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}},$$

что делает схему еще проще:

$$x_{k+1} = y_k - \frac{1}{M}\nabla f(y_k), \quad y_{k+1} = x_{k+1} + \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}(x_{k+1} - x_k)$$

Замечание 3. Выбор $\alpha_0 = \sqrt{m/M}$ не годится при $m = 0$.