Субградиентный спуск

Мальковский Н. В.

Санкт-Петербургский академический университет

Общая идея субградиента

Для произвольной дифференцируемой выпуклой функции f выполняется

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

Градиентный спуск

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

При правильном выборе α_k последовательность сходится.

На самом деле вместо $\nabla f(x_k)$ можно выбирать любой вектор g такой, что

$$f(y) \geq f(x_k) + g_k^T (y - x_k).$$

Выпуклая функция не обязана быть дифференцируемой, но обязана иметь хотя бы один такой вектор g в любой точке x.

Определение субградиента

Определение

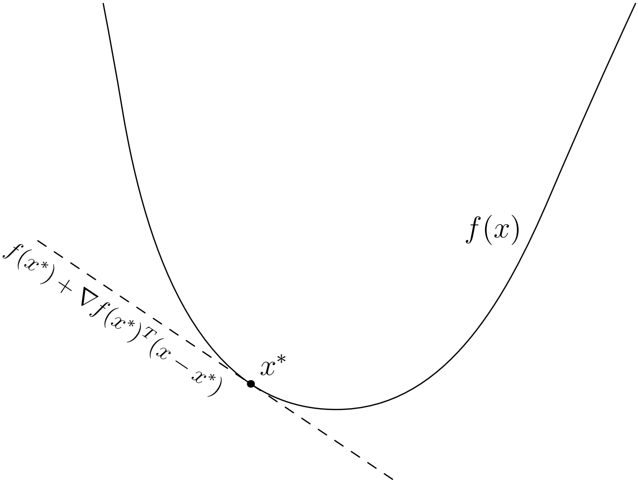
Пусть $f : \mathcal{D} \to \mathbb{R}$. Вектор g называется субградиентом функции f в точке $x \in \mathcal{D}$, если $\forall y \in \mathcal{D}$ выполняется

$$f(y) \geq f(x) + g^T(y - x).$$

Определение

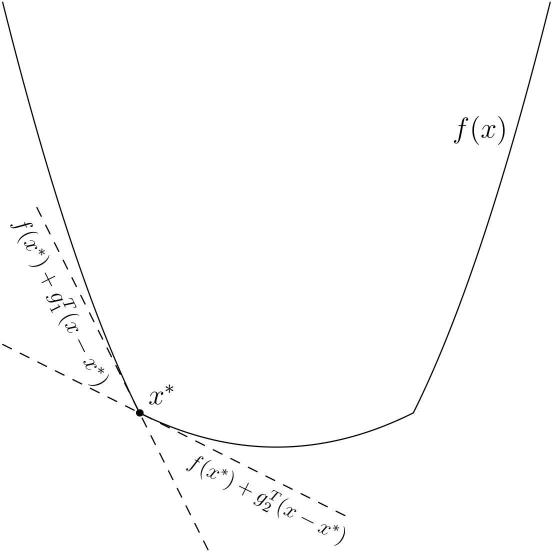
Субдифференциалом f в точке x называется множество всех субградиентов f в точке x и обозначается $\partial f(x)$.

Геометрические свойства субградиента



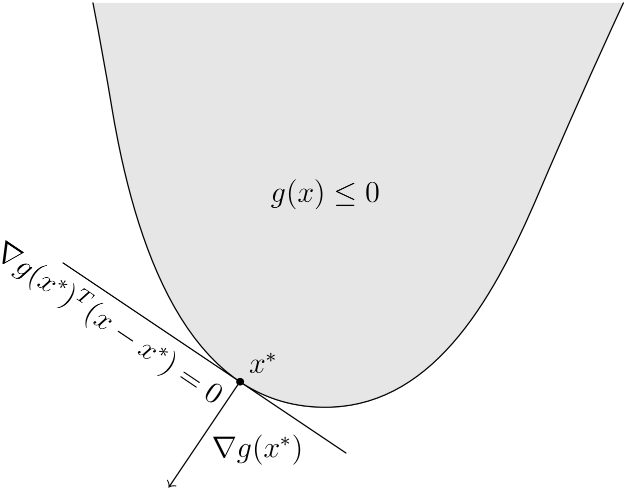
Опорные прямые для эпиграфа дифференцируемой выпуклой функции.

Геометрические свойства субградиента



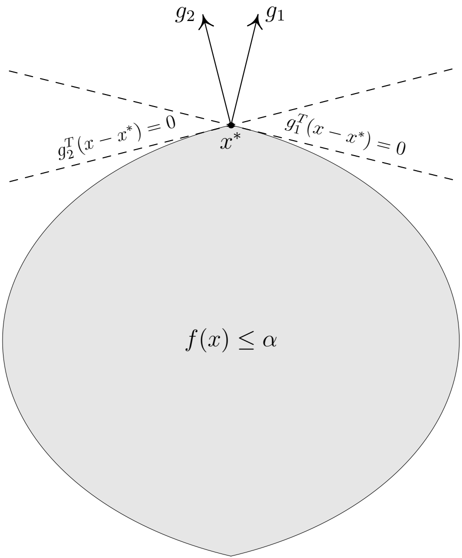
Опорные прямые для эпиграфа произвольной выпуклой функции.

Геометрические свойства субградиента



Опорные прямые для множества, ограниченного дифференцируемой выпуклой функцией.

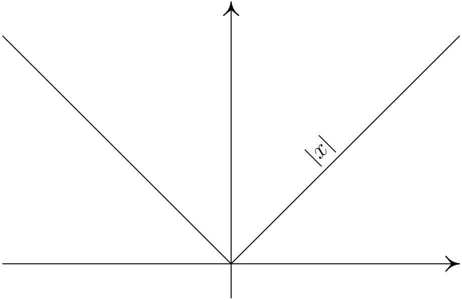
Геометрические свойства субградиента



Опорные прямые для множества, ограниченного произвольной выпуклой функцией.

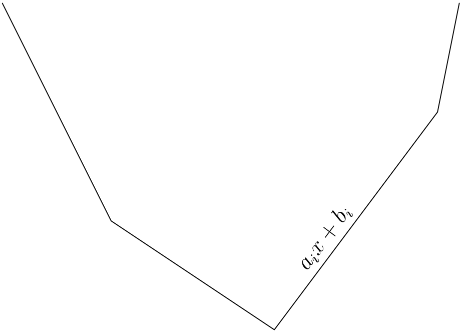
Пример: выпуклые кусочно-линейные функции

$$f(x) = |x| = \max\{x, -x\}$$



$$\partial f(x) = \begin{cases} \{1\}, & x > 0 \\ [-1, 1], & x = 0 \\ \{-1\}, & x < 0 \end{cases}$$

$$f(x) = \max_{1 \le i \le m} a_i x + b_i$$



$$\partial f(x) = \left[\min_{i \in I(x)} a_i ; \max_{i \in I(x)} a_i \right]$$

Свойства субградиента

1. Если f выпукла и замкнута на $\mathcal{D}$, то $\forall x \in \operatorname{Int}\mathcal{D}\ \partial f(x)$ – непустое замкнутое выпуклое множество.

Док-во. Если f выпукла на $\mathcal{D}$, то эпиграф f – выпуклое множество, тогда в точке $(x, f(x))$ существует опорная гиперплоскость с нормалью $(g, h), 0 \neq h \in \mathbb{R}, g \in \mathbb{R}^n$:

$$\forall y \in \mathcal{D}, \tau \geq f(y): \quad h(\tau - f(x)) + g^T(y - x) \geq 0.$$

Не умаляя общности можно считать, что $h^2 + ||g||^2 = 1$. Так как τ можно взять бесконечно большим, то для выполнения этого неравенства необходимо $h \geq 0$.
Так как f замкнута и выпукла, то в некоторой окрестности V_x точки x она удовлетворяет условию Липшица:

$$f(y) - f(x) \leq M||y - x||, \quad y \in V_x,$$

что дает при $y \in V_x$

$$-g^T(y - x) \leq h(f(y) - f(x)) \leq hM||y - x||.$$

Свойства субградиента

Взяв $y = x - \epsilon g$ получаем $||g||^2 \le Mh||g||$, что, учитывая $||g||^2 + h^2 = 1$, дает

$$h \ge \frac{1}{\sqrt{1+M^2}} > 0,$$

а значит $-\frac{1}{h}g$ – субградиент f в точе x.

С другой стороны, если $g \in \partial f(x)$, то при $y = x - \epsilon g/||g|| \in V_x$

$$\epsilon||g|| = g^T(y-x) \le f(y) - f(x) \le M||y-x|| = M\epsilon,$$

что дает ограниченность $\partial f(x)$. Выпуклость и замкнутость легко проверяются по определению. ■

Замечание. Функция $f(t) = -\sqrt{t}$ задана на $\mathcal{D} = \{t \ge 0\}$, выпукла и замкнута, но при этом в точке $t = 0$ субдифференциал пуст.

Свойства субградиента

2. Если f выпукла и дифференцируема на $\mathcal{D}$, то при $x \in \mathcal{D}: \quad \partial f(x) = \{\nabla f(x)\}$.

Док-во. Очевидным образом $\nabla f(x) \in \partial f(x)$. С другой стороны, если $g \in \partial f(x)$, то

$$f(y) = f(x) + \nabla f(x)^T (y - x) + o(||y - x||) \geq f(x) + g^T (y - x),$$

$$(\nabla f(x) - g)^T (y - x) \geq o(||y - x||).$$

Последнее неравенство может быть выполнено только если $g = \nabla f(x)$. ∎

Свойства субградиента

3. Пусть $\mathcal{D} \in \mathbb{R}^n, \mathcal{B} \in \mathbb{R}^m$, $f : \mathcal{D} \times \mathcal{B} \to \mathbb{R}$ – выпуклая функция, $x \in \mathcal{D}, y \in \mathcal{B}$, тогда функция

$$\phi_x(y) = f(x, y)$$

выпукла и при этом если $(g, h) \in \partial f(x, y)$, то $h \in \partial \phi_x(y)$.

Док-во.

$$\begin{aligned} \phi_x(z) = f(x, z) &\geq f(x, y) + g^T(x - x) + h^T(z - y) \\ &= \phi_x(y) + h^T(z - y) \quad \blacksquare \end{aligned}$$

Далее будем обозначать $\partial_y f(x, y) = \partial \phi_x(y)$. Стоит отметить, что в отличии от дифференцирования, если $g \in \partial_x f(x, y)$, $h \in \partial_y f(x, y)$, то это еще не значит, что $(g, h) \in \partial f(x, y)$ (например $f(x) = ||x||_2$ при $x = 0_n$). Стоит однако отметить, что $\forall h \in \partial_y f(x, y)\ \exists g \in \partial_x f(x, y)$ такое, что $(g, h) \in \partial f(x, y)$.

Свойства субградиента

4. Обозначим за $f'(x;p)$ – производную f в точке x по направлению p, т.е.

$$f'(x;p) = \lim_{\alpha \to 0+} \frac{f(x+\alpha p) - f(x)}{\alpha}.$$

Если f выпукла на $\mathcal{D}$, то для $x \in \operatorname{Int} \mathcal{D}$ $f'(x;p)$ существует и

$$f'(x;p) = \sup_{g \in \partial f(x)} g^T p.$$

Док-во. Пусть $x \in \operatorname{Int} \mathcal{D}$, обозначим $\varphi(p) = f'(x;p)$. Если $g \in \partial f(x)$, то

$$f(x+\alpha p) \geq f(x) + \alpha g^T p.$$

Следовательно, учитывая $\varphi(0) = 0$, $\partial f(x) \subset \partial \varphi(0)$. Далее, если $g \in \partial \varphi(0)$, то

$$f(x+p) \geq f(x) + \varphi(p) \geq f(x) + g^T p.$$

Следовательно $\partial \varphi(0) \subset \partial f(x)$ (Первое неравенство и существование $f'(x,p)$ без доказательства).

Свойства субградиента

Рассмотрим $g_p \in \partial\varphi(p)$, $\alpha > 0$, тогда

$$\alpha\varphi(y) = \varphi(\alpha y) \geq \varphi(p) + g_p^T(\alpha y - p).$$

Устремляя $\alpha \to \infty$ получаем

$$\varphi(y) \geq g_p^T y = \varphi(0) + g_p^T y.$$

Следовательно, $g_p \in \partial\varphi(0)$. Устремляя $\alpha \to 0$ получаем

$$\varphi(p) - g_p^T p \leq 0.$$

Но раз $g_p \in \partial f(x)$, то

$$\varphi(p) = \lim_{\alpha \to 0+} \frac{f(x + \alpha p) - f(x)}{\alpha} \geq g_p^T p.$$

Значит,

$$g_p^T p = \sup_{g \in \partial f(x)} g^T p = \max_{g \in \partial f(x)} g^T p = f'(x, p). \quad \blacksquare$$

Свойства субградиента

5. Если f_1, f_2 выпуклы на $\mathcal{D}$, $f = \alpha f_1 + \beta f_2$, то $\partial f(x) = \alpha \partial f_1(x) + \beta \partial f_2(x)$.

Док-во. В силу линейности производной по направлению

$$\begin{aligned} f'(x; p) &= \max_{g \in \partial f(x)} g^T p = \alpha \max_{g \in \partial f_1(x)} g^T p + \beta \max_{g \in \partial f_2(x)} g^T p \\ &= \max_{g \in \alpha \partial f_1(x) + \beta \partial f_2(x)} g^T p. \end{aligned}$$

Таким образом опорные функции $\partial f(x)$ и $\alpha \partial f_1(x) + \beta \partial f_2(x)$ совпадают. Следовательно, совпадают и сами множества. ■

Свойства субградиента

6. Если $f_1, \dots, f_m$ – выпуклые функции, то для функции $f(x) = \max_{1 \le i \le m} f_i(x)$ выполняется

$$\partial f(x) = \mathrm{Conv} \cup_{i \in I(x)} \partial f_i(x),$$

где $I(x) = \{i \mid f_i(x) = f(x)\}$, $\mathrm{Conv}\, X$ – выпуклая оболочка множества X.

Док-во. Для простоты полагаем, что $I(x) = \{1, \dots, k\}$.

$$\begin{aligned}
f'(x; p) &= \max_{i \in I(x)} f_i'(x; p) \\
&= \max_{1 \le i \le k} \max_{g_i \in \partial f_i(x)} g_i^T p \\
&= \max_{\alpha \in \Delta_k} \left\{ \sum_{i=1}^{k} \alpha_i \max_{g_i \in \partial f_i(x)} g_i^T p \right\} \\
&= \max_{\alpha \in \Delta_k,\ g_i \in \partial f_i(x)} \left\{ \sum_{i=1}^{k} \alpha_i g_i^T p \right\} \\
&= \max_{\alpha \in \Delta_k,\ g \in \sum_{i=1}^{k} \alpha_i \partial f_i(x)} \{g^T p\} = \max_{g \in \mathrm{Conv} \cup_{i \in I(x)} \partial f_i(x)} \{g^T p\}. \quad \blacksquare
\end{aligned}$$

7. Если $x^* \in \operatorname{Int} \mathcal{D}$, то x^* является точкой минимума f на $\mathcal{D}$ тогда и только тогда, когда $0_n \in \partial f(x^*)$.

Док-во. Эквивалентность полностью описывается следующим неравенством

$$f(x) \geq f(x^*) + 0_n^T(x - x^*). \quad \blacksquare$$

Свойства субградиента

8. f – непрерывна по Липшицу с константой M тогда и только тогда, когда $\forall x \in \mathcal{D},\ \forall g \in \partial f(x) :\ ||g|| \leq M$.

Док-во. (Необходимость) Очевидным образом, если для некоторого x существует $g \in \partial f(x),\ ||g|| > M$, то для $y = x + \alpha g$ имеем

$$f(y) - f(x) \geq g^T(y - x) = |\alpha| ||g||^2 > M||y - x||.$$

(Достаточность) Если $g \in \partial f(x)$, то

$$f(x) - f(y) \leq g^T(x - y) \leq ||g|| \cdot ||y - x|| \leq M||y - x||. \quad \blacksquare$$

Пример: l_1 и l_2 нормы

$$f(x) = ||x||_1 = \max_{s \in \{-1,1\}^n} s^T x$$

Очевидным образом, если $s^T x = p^T x$ при $s, p \in \{-1, 1\}^n$, то $x_i \neq 0 \Rightarrow p_i = s_i$. Таким образом $\partial f(x) = J_1(x) \times \ldots \times J_m(x)$, где

$$J_i(x) = \begin{cases} \{1\}, & x_i > 0 \\ [-1, 1], & x_i = 0 \\ \{-1\}, & x_i < 0 \end{cases}$$

$$f(x) = ||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

f дифференцируема во всех точках кроме 0_n, следовательно, если $x \neq 0_n$, то $\partial f(x) = \{\nabla f(x)\} = \left\{\frac{x}{||x||_2}\right\}$. Для 0_n имеем

$$f(0_n; p) = \lim_{\alpha \to 0+} \frac{||\alpha p||_2}{\alpha} = ||p||_2,$$

что является опорной функцией единичного шара. Следовательно, $\partial f(0_n) = \{x \mid ||x||_2 \leq 1\}$.

Субградиент математического ожидания

Пусть $F : \mathcal{D} \times \Omega \to \mathbb{R}$, F выпукла по первому аргументу, ω – некоторая случайная величина на Ω. Рассмотрим функцию

$$f(x) = E_\omega F(x, \omega).$$

- f – выпукла.
- $E_\omega \partial_x F(x, \omega) \subset \partial f(x)$

Если $g : \Omega \to \partial_x F(x, \omega)$, то $E_\omega g(\omega) \in \partial f(x)$:

$$F(y, \omega) \geq F(x, \omega) + g^T(\omega)(y - x)$$

Взяв математическое ожидание от обоих частей неравенства получаем вложенность субградиента и его непустоту (а следовательно и выпуклость).

Субградиент поточечного супремума

Пусть $F : \mathcal{D} \times Y \to \mathbb{R}$ выпукла по первому аргументу. Рассмотрим

$$f(x) = \sup_{y \in Y} F(x, y)$$

Ранее рассматривался случай конечного Y, для произвольного же выполняется

$$\mathrm{Conv} \bigcup_{F(x,y)=f(x)} \partial_x F(x, y) \subset \partial f(x)$$

В частности, если $F(x, y) = f(x)$, то $\partial_x F(x, y) \subset \partial f(x)$.

Субградиент поточечного инфиума

Пусть $F : \mathcal{D} \times Y \to \mathbb{R}$ выпукла (по (x, y)). Рассмотрим

$$f(x) = \inf_{y \in Y} F(x, y)$$

Покажем, как найти хотя бы один субградиент: пусть $f(x^*) = F(x^*, y^*)$ для некоторых x^*, y^*. Так как $F(x^*, y^*) = \inf_{y \in Y} F(x^*, y)$, то $0_Y \in \partial_y F(x^*, y^*)$, следовательно существует $\underline{(g, 0_Y) \in \partial F(x^*, y^*)}$, таким образом

$$F(x, y) \geq F(x^*, y^*) + g^T(x - x^*) + 0_Y^T(y - y^*).$$

Минимизируя по y левую часть (правая не зависит от y) и учитывая $f(x^*) = F(x^*, y^*)$ получаем

$$f(x) \geq f(x^*) + g^T(x - x^*)$$

Замечание. Нахождение субградиента предполагает достижимость инфиума, т.е. фактически $f(x) = \min_{y \in Y} F(x, y)$.

Условия ККТ и субградиент

Пусть $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$, $g : \mathcal{D} \to \mathbb{R}^m$ $1 \le i \le m$ – непрерывно дифференцируемые функции на $\mathcal{D}$. Рассмотрим задачу

$$\begin{array}{ll} \text{минимизировать} & f(x) \\ \text{при условии} & g_i(x) \le 0_m. \end{array}$$

Введем две вспомогательные функции

$$F(t, x) = \max\{f(x) - t; g_i(x), 1 \le i \le m\}$$

$$f^*(t) = \min_x F(t, x)$$

Лемма

Если $t^* = \min_{g(x) \le 0_m} f(x)$, *то*

$$\begin{cases} f^*(t) \le 0, & t \ge t^*, \\ f^*(t) > 0, & t < t^*. \end{cases}$$

Условия ККТ и субградиент

Док-во. Пусть $f(x^*) = t^*$, тогда при $t \geq t^*$

$$f^*(t) \leq F(t, x^*) = \max\{t^* - t, g_i(x^*)\} \leq 0.$$

С другой стороны, если для некоторого $t < t^*$ выполняется $f^*(t) \leq 0$, то для $y = \operatorname{argmin}_x F(t, x)$ имеем

$$f^*(t) = \max\{f(y) - t, g_i(y)\} \leq 0,$$

т.е. $g(y) \leq 0_m$ и $f(y) - t < f(y) - t^* \leq 0$, а значит x^* – не точка минимума исходной задачи. ■

Следствие. $x^* = \operatorname{argmin}_{g(x) \leq 0_m} f(x) \Leftrightarrow x^* = \operatorname{argmin}_x F(t^*, x)$.

Если $x^* = \operatorname{argmin}_{g(x) \leq 0_m} f(x)$, то $F(t^*, x^*) = 0$. Из леммы следует, что $0 = F(t^*, x^*) = f^*(t^*)$, т.е. x^* минимизирует $F(t^*, \cdot)$.

С другой стороны, если $F(t^*, x^*)$ достигает минимума на x^*, то $F(t^*, x^*) = 0$. Следовательно, $g(x^*) \leq 0_m$, $f(x^*) = t^*$. ■

Условия ККТ и субградиент

Наконец, из свойств субградиента, если x^* минимизирует $F(t^*, \cdot)$, то

$$0_n \in \partial_x F(t^*, x^*) = \text{Conv} \bigcup_{i \in I(x^*)} \{\nabla f(x^*); \nabla g_i(x^*)\}$$

В соответствующую выпуклую оболочку всегда выходит $\nabla f(x^*)$, а так же активные ограничения $g_i(x^*) = 0$. Из характеристики выпуклой оболочки получаем, что существуют такие неотрицателные коэффициенты $\lambda_0, \ldots, \lambda_m$, что

$$0_n = \lambda_0 \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*),$$

При этом $\lambda_i \neq 0 \Rightarrow g_i(x^*) = 0$. Добавляя условия регулярности (векторы $\nabla g_i(x^*)$ линейно независимы) получаем, что $\lambda_0 > 0$.

Условия ККТ и субградиент

Итого имеем: $x^* = \text{argmin}_{g(x) \le 0_m} f(x)$, тогда существуют $\lambda_1, \ldots, \lambda_m$ такие, что

1. $$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) = 0_n$$
2. $$g(x^*) \le 0_n$$
3. $$\lambda_i \ge 0$$
4. $$\lambda_i g_i(x^*) = 0$$

Субградиентный спуск

Итак, пусть $f : \mathcal{D} \to \mathbb{R}$ – выпуклая функция. Заменяя в градиентном методе градиент на субградиент получаем

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k) \tag{1}$$

Основное преимущество: применим для любой выпуклой функции

Основной недостаток: экспоненциальную сходимость можно получить в довольно экзотических случаях. В подавляющем большинстве сходимость медленная.

Основная проблема: в отличии от градиентного спуска нельзя гарантировать, что $g_k \to 0_n$ даже если $x_k \to x^*$.

Основные предположения

В дальнейшем будет предполагаться следующее

- f – выпуклая на $\mathcal{D}$ функция.
- f непрерывна по Липшицу с константой L, иначего говоря все субградиенты f равномерно ограничены на $\mathcal{D}$ константой L.
- Расстояние от начального приближения до ближайшей точки минимума ограничено R. Иначе говоря,

$$||x_0 - x^*|| \leq R$$

Способы выбора шага

- Постоянный

$$\alpha_k = \alpha$$

- Расходящийся ряд

$$\alpha_k \to 0, \ \sum_{i=0}^{\infty} \alpha_k = \infty$$

- Расходящийся ряд со сходящимся рядом квадратов

$$\sum_{i=0}^{\infty} \alpha_k = \infty, \ \sum_{i=0}^{\infty} \alpha_k^2 < \infty$$

- Нормированный

$$\alpha_k = \frac{\gamma_k}{||g_k||}$$

γ_k – одна из вышеуказанных последовательностей.

Основное неравенство субградиентного спуска

Пусть $\phi_k = \min_{1 \le i \le k} f(x_k)$, тогда

$$\begin{aligned}
||x_{k+1} - x^*||^2 &= ||x_k - x^*||^2 - 2\alpha_k g_k^T(x_k - x^*) + \alpha_k^2 ||g_k||^2 \\
&\le ||x_k - x^*||^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 L^2 \\
&\le ||x_0 - x^*||^2 - 2\sum_{i=0}^{k} \alpha_i (f(x_i) - f(x^*)) + L^2 \sum_{i=0}^{k} \alpha_i^2 \\
&= ||x_0 - x^*||^2 - 2(\phi_k - f(x^*)) \sum_{i=0}^{k} \alpha_i + L^2 \sum_{i=0}^{k} \alpha_i^2
\end{aligned}$$

Таким образом

$$\phi_k - f(x^*) \le \frac{R + L^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} \tag{2}$$

Сходимость субградиентного спуска

Теорема (О сходимости субградиентного спуска)

Если f – выпуклая на $\mathcal{D}$ функция, x^* – точка минимума f на $\mathcal{D}$, f непрерывна по Липшицу с константой L, $||x_0 - x^*|| \leq R$, то для наилучшего приближения, генерируемого по правилу (1) выполняется

- *При* $\alpha_k = \alpha$

$$\overline{\lim_{k\to+\infty}} \phi_k - f(x^*) \leq \frac{\alpha L^2}{2}$$

- *При* $\alpha_k \to 0,\ \sum_{k=1}^{\infty} \alpha_k = \infty$

$$\phi_k - f(x^*) \to 0.$$

Сходимость субградиентного спуска

Док-во. Первое утверждение выводится непосредственно подстановкой $\alpha_k = \alpha$ и предельным переходом в (1)

Для второго утверждения в силу $\alpha_i \to 0$ существует $N_1 : \ \forall k > N_1 \ \alpha_i < \frac{\epsilon}{L^2}$, а в силу $\sum_{i=0}^{\infty} \alpha_i = \infty$ существует $N_2 : \ \forall n > N_2 \ \sum_{i=0}^{n} \alpha_i > \frac{R}{\epsilon}$ таким образом для $k \geq \max\{N_1, N_2\}$ получаем

$$\phi_k - f(x^*) \leq \frac{R + L^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} \leq \frac{R}{2 \sum_{i=0}^{k} \alpha_i} + \frac{L^2 \epsilon \sum_{i=0}^{k} \alpha_i}{2 \sum_{i=0}^{k} \alpha_i}$$

$$< \frac{R}{2 \frac{R}{\epsilon}} + \frac{\epsilon \sum_{i=0}^{k} \alpha_i}{2 \sum_{i=0}^{k} \alpha_i} = \epsilon$$

Оптимальный выбор шага относительно (2)

Оценка (2) представляет собой функцию переменных $\alpha_1, \ldots, \alpha_k$, минимизация которой будет давать гарантию лучшей сходимости.

Обозначим

$$\Phi(\alpha_1, \ldots, \alpha_k) = \frac{R + L^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}.$$

Дифференцируя по α_i получаем

$$\frac{\partial \Phi}{\partial \alpha_i} = \frac{4L^2 \alpha_i \sum_{i=1}^k \alpha_i - 2(R + L^2 \sum_{i=1}^k \alpha_i^2)}{(2 \sum_{i=1}^k \alpha_i)^2} = 0.$$

Уравнение идентично для различных i, что дает равенство всех α_i. Используя это получаем упрощенное уравнение на α_i:

$$4L^2 k \alpha_i^2 = 2R + 2L^2 k \alpha_i^2,$$

что дает

$$\alpha_i = \frac{\sqrt{R}}{L\sqrt{k}}$$

Ссылки на литературу

Нестеров Ю. Е. Методы выпуклой оптимизации // параграфы 3.1.5–3.2.3

Vandenberghe L. Subgradients (slides)

Boyd S. et al. Subgradient Descent (course notes)