$$\min_{x \in \mathbb{R}^d} f(x)$$

Пример: $f(x) = |x|$ выпукла? да

$-\delta \quad \delta$

$$|f'(\delta) - f'(-\delta)| \le L\,|\delta - (-\delta)|$$

$$2 \le 2L\delta$$

$\delta \to 0 \quad L \to \infty$

нет L-липш. градиента

Определение M-Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является M-Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \le M\|x - y\|_2.$$

+ выпуклость (сильная выпуклость не дружит с липшицевостью)

Субградиент и субдифференциал

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Вектор g будем называть субградиентом этой функции f в точке $x \in \mathbb{R}^d$, если для любого $y \in \mathbb{R}^d$ выполняется:

$$f(y) \ge f(x) + \langle g, y - x \rangle.$$

Множество $\partial f(x)$ всех субградиентов f в x будем называть субдифференциалом.

Теорема (условие оптимальности)

x^* – минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

$\partial f(0) = [-1; 1]$

$y \to x \quad x \to x^*$

Док-во:

$\Leftarrow \quad 0 \in \partial f(x^*)$

$$f(x) \ge f(x^*) + \langle \underset{=0}{\overset{\in \partial f(x^*)}{g}} ; x - x^* \rangle = f(x^*) \quad \forall x$$

минимум

$\Rightarrow$ x^* - минимум

$$f(x) \geq f(x^*) = f(x^*) + \langle 0; x - x^* \rangle$$

$0 \in \partial f(x^*)$ по опред. субдиф.

Лемма (свойство M-Липшицевой функции)

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда функция f является M-Липшицевой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$ имеем $\|g\|_2 \leq M$.

Док-во:

$\Rightarrow$ f - M- липшицева, выпукла

Δ $g \in \partial f(x)$

по выпуклости и опред. субград.

$$f(y) - f(x) \geq \langle g; y - x \rangle$$

f - M- липшицева $\quad f(y) - f(x) \leq M\|x - y\|_2$

$$\langle g; y - x \rangle \leq M\|y - x\|_2$$

$y = x + g$

$$\|g\|_2^2 \leq M\|g\|_2$$

$$\|g\|_2 \leq M$$

$\Leftarrow$ $\quad \|g\|_2 \leq M \quad \forall x \in \mathbb{R}^d \quad \forall g \in \partial f(x)$

по выпуклости и опред. субград.

$$f(y) - f(x) \geq \langle g; y - x \rangle \quad | \cdot (-1)$$

$$f(x) - f(y) \leq \langle g; x - y \rangle$$

КБШ

$$f(x) - f(y) \le \|g\|_2 \cdot \|x-y\|_2$$

$\|g\|_2 \le M$

$$f(x) - f(y) \le M \|x-y\|_2$$

аналогично

$$f(y) - f(x) \le M \|x-y\|_2$$

$$|f(x) - f(y)| \le M \cdot \|x-y\|_2 \qquad \blacksquare$$

Алгоритм 2 Субградиентный метод

Вход: размеры шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \dots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $x^{k+1} = x^k - \gamma g^k$
4: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Док-во:

• f — M-липшицева и выпуклая

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - \gamma g^k - x^*\|_2^2$$

$$= \|x^k - x^*\|_2^2 - 2\gamma \langle g^k ; x^k - x^* \rangle + \gamma^2 \|g^k\|_2^2$$

$\|g^k\|_2 \le M$

$$\leq \|x^k - x^*\|_2^2 - 2\gamma \langle \underline{g^k}; x^k - x^* \rangle + \gamma^2 M^2$$

Выпуклость: $\langle g^k; x^k - x^* \rangle \geq f(x^k) - f(x^*)$

$$\leq \|x^k - x^*\|_2^2 - 2\gamma\left(f(x^k) - f(x^*)\right) + \gamma^2 M^2$$

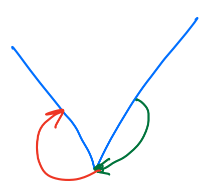
$$2\gamma\left(f(x^k) - f(x^*)\right) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

$$\sum_{k=0}^{K-1}$$

$$2\gamma \sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \|x^0 - x^*\|_2^2 - \cancel{\|x^K - x^*\|_2^2} + \gamma^2 M^2 K$$

$$\frac{1}{K}$$

$$2\gamma \cdot \left(\frac{1}{K}\sum f(x^k) - f(x^*)\right) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$



но последний может

сходимость может быть хуже

Нер-во Йенсена

$$2\gamma\left(f\left(\frac{1}{K}\sum x^k\right) - f(x^*)\right) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2 \Big/ 2\gamma$$

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

$$\min_{\gamma} \qquad \gamma^* = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}} \qquad \frac{M^2}{2} - \frac{\|x^0 - x^*\|_2^2}{2\gamma^2 K} = 0$$

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \le \frac{\|x^0 - x\|_2 \cdot M}{\sqrt{K}}$$

Теорема сходимость субградиентного спуска для M-Липшицевых и выпуклых функций

Пусть задача безусловной оптимизации с M-Липшицевой, выпуклой целевой функцией f решается с помощью субградиентного спуска. Тогда справедлива следующая оценка сходимости

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

Более того, чтобы добиться точности ε по функции, необходимо

$$K = O\left(\frac{M^2\|x^0 - x^*\|_2^2}{\varepsilon^2}\right) \text{ итераций.}$$

- для гладких задач $\frac{1}{K}$ для град. спуска

  $\frac{1}{K^2}$ для Нестерова

негладкие задачи более сложные

- субград. спуск оптимален для негладких

$$\gamma^* = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$$

плохо

1) $\gamma_k = \dfrac{\|x^0 - x^*\|_2}{M\sqrt{k+1}}$ вместо $K \to k+1$

2)

$$\gamma^k = \frac{\|x^0 - x^*\|_2}{\sqrt{(k+1) M^2}} \longrightarrow \sum_{t=0}^{k} \|g^t\|_2^2$$

$$\gamma^k = \frac{\cancel{\|x^0 - x^*\|_2}}{\sqrt{\sum_{t=0}^{k} \|g^t\|_2^2}} \qquad \cancel{\|x^0 - x^*\|_2} \longrightarrow D$$

подбираем

Алгоритм 3 AdaGradNorm

Вход: $D > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм градиентов $G^0 = 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Вычислить $g^k \in \partial f(x^k)$

3: Вычислить $G^{k+1} = G^k + \|g^k\|_2^2$

4: $x^{k+1} = x^k - \frac{D}{\sqrt{G^{k+1}}} g^k$

5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K} x^k$

$$\min_{x_1, x_2} \; x_1^2 + 1000 x_2^2$$

$$\gamma^k = \frac{D}{\sqrt{\sum_{t=0}^{k} \|g^t\|_2^2}} \longrightarrow \gamma^{k,i} = \frac{D}{\sqrt{\sum_{t=0}^{k} (g_i^t)^2}}$$

для каждой из координат

Алгоритм 4 AdaGrad

Вход: $D > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов градиентов $G_i^0 = 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Вычислить $g^k \in \partial f(x^k)$

3: Для каждой координаты: $G_i^{k+1} = G_i^k + (g_i^k)^2$

4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D}{\sqrt{G_i^{k+1}}} g_i^k$

5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K} x^k$

$$G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2 \qquad \beta_2 \in (0;1)$$

идея, как в методе с моментумом

$$\gamma_{k,i} = \frac{D}{\sqrt{G_i^{k+1}}}$$

Алгоритм 5 RMSProp

Вход: $D_i > 0$, моментум $\beta_2 \in (0,1)$, стартовая точка $x^0 \in \mathbb{R}^d$, сглаженная сумма квадратов градиентов $G_i^0 = 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2$
4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}}} g_i^k$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

объед.

тяжелый шарик: $v^{k+1} = \beta_1 v^k + (1-\beta_1) g^k$

RMSProp: $G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2$

$$x_i^{k+1} = x_i^k - \frac{D}{\sqrt{G_i^{k+1}}} v_i^{k+1}$$

Алгоритм 6 Adam

Вход: $D_i > 0$, моментумы $\beta_1 \in (0,1)$ и $\beta_2 \in (0,1)$, стартовая точка $x^0 \in \mathbb{R}^d$, сглаженная сумма квадратов градиентов $G_i^0 = 0$, сглаженная сумма градиентов $v^0 = 0$, добавка $e > 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Вычислить $v^{k+1} = \beta_1 v^k + (1-\beta_1) g^k$
4: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2$
5: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1} + e}} v_i^{k+1}$
6: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

чтобы не делить на 0

$\beta_2 = 0.99, \; 0.999, \; 0.9999$ есть рецепты

$\beta_1 = 0.9 - 0.99$

D – все ещё трудно подобрать

$$D \approx \|x^0 - x^*\|_2$$ как подобрать?

знаем   близко? $x^k \to x^*$

$$d_k = \min\{\|x^k - x^0\|_2 ; d_{k-1}\}$$

Новая версия AdaGradNorm

$$\gamma_k = \frac{d_k}{\sqrt{\sum_{t=0}^{k} \|g^t\|_2^2}}$$

Можно сделать и с AdaGrad: $d_{k,i}$

Проксимальный оператор

- Поняли, что негладкие задачи «более сложные» по сравнению с гладкими задачами.
- Может быть получится «спрятать под ковер» отсутствие гладкости.
- Такую возможность дает проксимальный оператор:

Определение проксимального оператора

Для функции $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ проксимальный оператор определяется следующим образом:

$$\mathrm{prox}_r(x) = \arg\min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|^2 \right).$$

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r. Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

Доказательство: Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Вопрос: что можно сказать про эту задачу? Она сильно выпуклая, а значит имеет строго один уникальный минимум (существование $\hat{x}$ необходимо для того, чтобы $r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|^2$ где-то принимала конечное значение).

Примеры проксимального оператора

- $r(x) = \lambda \|x\|_1$, где $\lambda > 0$. Тогда

$$[\mathrm{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \mathrm{sign}(x_i)$$

Такой проксимальный оператор еще называют трешхолдом.

- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

$$\mathrm{prox}_r(x) = \frac{x}{1+\lambda}.$$

- $r(x) = \mathbb{I}_{\mathcal{X}}(x)$, где $\mathcal{X}$ – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

Вопрос: чему равен prox?

$$\mathrm{prox}_r(x) = \mathrm{proj}_{\mathcal{X}}(x).$$

- И еще множество других примеров и их комбинаций.

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r. Тогда для любых $x, y \in \mathbb{R}^d$ следующие три условия являются эквивалентными:

- $\mathrm{prox}_r(x) = y$,
- $x - y \in \partial r(y)$,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$ для любого $z \in \mathbb{R}^d$.

Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d}\left(r(\tilde{x}) + \frac{1}{2}\|x-\tilde{x}\|_2^2\right).$$
- Из условия оптимальности для выпуклой функции r это эквивалентно **вопрос:** чему?
$$0 \in \partial\left(r(\tilde{x}) + \frac{1}{2}\|x-\tilde{x}\|_2^2\right)\Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$
Получили эквивалентность первого и второго условий.
- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:
$$\langle g, z-y\rangle \le r(z) - r(y).$$
В частности справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$. Лемма доказана.

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено следующее:

- $\langle x - y, \text{prox}_r(x) - \text{prox}_r(y) \rangle \geq \|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2$,
- $\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$.

Доказательство

- Пусть $u = \text{prox}_r(x)$, $v = \text{prox}_r(y)$. Тогда из предыдущего свойства:

$$\langle x - u, z_1 - u\rangle \leq r(z_1) - r(u),$$
$$\langle y - v, z_2 - v\rangle \leq r(z_2) - r(v).$$

- Подставляем $z_1 = v$ и $z_2 = u$. Суммируем:

$$\langle x - u, v - u\rangle + \langle y - v, u - v\rangle \leq 0$$

Откуда

$$\langle x - y, v - u\rangle + \|v - u\|_2^2 \leq 0.$$

А это и требовалось доказать. **Вопрос:** как быстро доказать второе утверждение леммы? КБШ.

Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

- Такая задача называется композитной.
- Предположим, что f является L-гладкой выпуклой функцией, r выпуклой (необязательно гладкой, но) проксимально дружественной функцией.
- Получается целевая функция состоит из гладкой и в общем случае негладкой части. Если $r \equiv 0$, то получаем гладкую задачу, которую умеем решать. Если $f \equiv 0$, то получаем негладкую задачу.

Проксимальный градиентный метод

Алгоритм 9 Проксимальный градиентный метод

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

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(x^k)$
3: $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k))$
4: **end for**

Выход: x^K

- Если r непрерывно дифференцируема, то условие оптимальности для подзадачи подсчета проксимального оператора записывается, как:

$$0 = \gamma \nabla r(x^{k+1}) + x^{k+1} - \gamma \nabla f(x^k).$$

- Откуда получаем так называемую неявную запись метода:

$$x^{k+1} = x^k - \gamma(\nabla f(x^k) + \nabla r(x^{k+1}))$$

Сходимость

Лемма (свойство проксимального оператора)

Пусть $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклые функции. Дополнительно предположим, что f является непрерывно дифференцируемой и L-гладкой, а для r определен prox_r. Тогда x^* – решение композитной задачи оптимизации тогда и только тогда, когда для любого $\gamma > 0$ выполнено:

$$x^* = \text{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

Доказательство

- Условие оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*).$$

- Откуда

$$x^* - \gamma \nabla f(x^*) - x^* \in \gamma \partial r(x^*).$$

- Из свойств проксимального оператора

$$x^* = \mathrm{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

А это и требовалось.

Сходимость

- В итоге имеем следующие свойства:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$
$$x^* = \text{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

Вопрос: в доказательстве какого метода уже нам нужны были такие свойства? Градиентный спуск с проекцией. Вспомним, что проксимальный оператор включает в себя и оператор проекции.

- Поэтому доказательство будет один в один.

Доказательства сходимости

- Рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2$$

- Используем второй свойство с предыдущего слайда:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2 \\ &= \|\text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - \text{prox}_{\gamma r}(x^* - \gamma_k \nabla f(x^*))\|_2^2\end{aligned}$$

- Теперь первое свойство с предыдущего слайда:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq &\|x^k - \gamma_k \nabla f(x^k) - x^* + \gamma_k \nabla f(x^*)\|_2^2 \\ = &\|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Доказательства сходимости

- С предыдущего слайда:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

- Вспомним такой объект, как дивергенция Брэгмана, порожденную выпуклой функцией f:

$$V_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.$$

Доказательства сходимости

- Воспользуемся сильной выпуклостью и гладкостью:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 \\
& - 2\gamma_k \left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle + \frac{\mu}{2}\|x^k - x^*\|_2^2 \right) \\
& + 2\gamma_k^2 L \left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle \right) \\
= & (1 - \mu\gamma_k)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1) V_f(x^k, x^*)
\end{aligned}
$$

- Дальше как раньше подбирает γ_k, пользуемся неотрицательности дивергенции Брэгмана.

Проксимальный метод: итог

- Проксимальный градиентный спуск для композитной задачи с L-гладкой выпуклой функцией f и выпуклой проксимально дружественной функцией r имеет такую же сходимость, что и метод градиентного спуска для функции f. Свойства гладкости/негладкости r при этом не влияют.
- Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу. **Вопрос:** так ли это? если разрешить считать проксимальный оператор неточно (численно), то и правда можно решать любую задачу негладкой оптимизации. НО это с точки зрения теории не лучше, чем решать задачу субградиентным спуском, потому что при решении подзадачи проксимального используется какой-то вспомогательный метод (например, тот же субградиентный спуск).