Метод проекции градиента. Проксимальные методы.

Семинарист: Данилова М.

Метод проекции градиента

Проекция точки на выпуклое множество

Определение 1. Пусть $a \in \mathbb{R}^n$, $\mathrm{X} \subseteq \mathbb{R}^n$. Тогда $\pi_{\mathrm{X}}(a)$ – *проекция a на* X, если

$$\|\pi_{\mathrm{X}}(a) - a\| \leq \|x - a\| \quad \forall x \in \mathrm{X}$$

Замечания

- Если $a \in \mathrm{X}$, то $\pi_{\mathrm{X}}(a) = a$.
- Если X – открытое множество и $a \notin \mathrm{X}$, то проекции точки a на X **не существует**.
- Если множество X – выпуклое и замкнутое, то проекция существует и единственна.
- В случае произвольного множества проекция может быть не единственна.

Теорема 1. (Критерий для нормы l_2)
$\pi_X(a)$ – проекция точки a на X тогда и только тогда, когда $\langle \pi_X(a) - a,\ x - \pi_X(a) \rangle \geq 0 \quad \forall x \in X$

Метод проекции градиента

Будем решать задачу условной минимизации:

$$\min_{x \in \mathrm{Q}} \quad f(x),$$

где $\mathrm{Q} \subseteq \mathbb{R}^d$ – выпуклое и замкнутое, а $f(x)$ – дифференцируема.

- Если x_* – точка минимума, то x_* – решение вариационного неравенства:

$$\langle \nabla f(x_*), x - x_* \rangle \geq 0 \quad \forall x \in \mathrm{Q}.$$

- Если $f(x)$ – выпукла, то это условие является достаточным.

Метод проекции градиента

Algorithm 1 Метод проекции градиента

Require: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N
  for $k = 0, 1, \ldots, N-1$ **do**
    Вычислить $\nabla f(x^k)$
    $x^{k+1} = \pi_{\mathrm{Q}}\left(x^k - \gamma \nabla f(x^k)\right)$
  end for
Ensure: x^N

Замечания

1. Пусть f – L-гладкая, $x_k \in \mathrm{Q}$

$$
\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_{x \in \mathrm{Q}} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in \mathrm{Q}} \left\{ \frac{1}{2L}\|\nabla f(x^k)\|_2^2 + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in \mathrm{Q}} \left\{ \left\| \frac{1}{\sqrt{2L}} \nabla f(x^k) + \sqrt{\frac{L}{2}}(x - x^k) \right\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in \mathrm{Q}} \left\{ \frac{L}{2} \left\| \frac{1}{L} \nabla f(x^k) + x - x^k \right\|_2^2 \right\} \\
&= \operatorname*{argmin}_{x \in \mathrm{Q}} \left\{ \left\| x - (x^k - \frac{1}{L} \nabla f(x^k)) \right\|_2^2 \right\} \\
&= \pi_{\mathrm{Q}}(x^k - \frac{1}{L} \nabla f(x^k))
\end{aligned}
$$

   Таким образом, для L-гладкой функции на каждом шаге метода проекции градиента с длиной шага $\gamma = \frac{1}{L}$ мы минимизируем квадратичную аппроксимацию на множестве Q (аналогично градиентному методу, где $\mathrm{Q} = \mathbb{R}^n$).

2. Метод проекции градиента целесообразно использовать, когда множество Q – **простое множество** (легко искать проекцию точки на данное множество).

   Можно формализовать следующим образом: Q – простое множество, если решение следующей задачи

$$\min_{x \in \mathrm{Q}} \mathrm{c}^\top x$$

   находится быстрее(чаще всего аналитически), чем решение исходной.

   Пример:

$$
\begin{aligned}
&\min_{\|x - a\|_2 \le r} \mathrm{c}^\top x \\
x^* &= a - \frac{rc}{\|c\|_2}.
\end{aligned}
$$

Примеры простых множеств:

- $\mathrm{Q} = \mathbb{R}^n_+$ – неотрицательный ортант
- $\mathrm{Q} = \{x \in \mathbb{R}^n \mid x_i \in [a_i,\ b_i]\}$ – параллелипипед
- $\mathrm{Q} = \mathrm{B}_2(0,1) = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ - шар в 2-норме

Сходимость метода проекции градиента

Метод проекции градиента – релаксационный метод, то есть все точки принадлежат допустимому множеству Q, и функция убывает от итерации к итерации. ($f(x)$ – L–гладкая функция, длина шага $\gamma < \frac{2}{L}$.)

Доказательство. Воспользуемся определением проекции точки $a \in \mathbb{R}^n$ на множество Q:

$$\langle \pi_{\mathrm{Q}}(a) - a,\ x - \pi_{\mathrm{Q}}(a) \rangle \geq 0 \quad \forall x \in \mathrm{Q}$$

Пусть $\pi_{\mathrm{Q}}(a) = x^{k+1},\ a = x^k - \gamma \nabla f(x^k),\ x = x^k$:

$$\left\langle x^{k+1} - (x^k - \gamma \nabla f(x^k)),\ x^k - x^{k+1} \right\rangle \geq 0.$$

Обозначим $s_k = x^{k+1} - x^k$:

$$\left\langle s_k + \gamma \nabla f(x^k),\ s_k \right\rangle \leq 0.$$

Разложим функцию $f(x)$ в ряд Тейлора с остаточным членом в форме Лагранжа:

$$f(x^{k+1}) = f(x^k) + \left\langle \nabla f(\tilde{x}^k,\ s_k \right\rangle, \quad \tilde{x}^k \in [x^k,\ x^{k+1}].$$

Воспользуемся L-гладкостью функции $f(x)$ и разложением:

$$\begin{aligned} f(x^{k+1}) \leq f(x^k) + \left\langle \nabla f(x^k),\ x^{k+1} - x^k \right\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 = \\ f(x^k) + \left\langle \nabla f(x^k),\ s_k \right\rangle + \frac{L}{2}\|s_k\|^2 = \\ f(x^k) + \frac{1}{\gamma}\left[\langle \gamma \nabla f(x^k),\ s_k \rangle + \|s_k\|^2\right] + \left(\frac{L}{2} - \frac{1}{\gamma}\right)\|s_k\|^2 \end{aligned}$$

Следовательно $f(x^{k+1}) - f(x^k) \leq 0$ тогда и только тогда, когда $\frac{L}{2} - \frac{1}{\gamma} < 0 \ \rightarrow \ \gamma < \frac{2}{L}$. □

Теорема 2.

1. Пусть f – **выпуклая**, L–**гладкая**, $\mathrm{Q} \in \mathbb{R}^n$ – выпукло и замкнуто, $\overline{x}^N = \frac{1}{N}\sum\limits_{k=1}^{N} x^k$.

   Тогда

   $$f(\overline{x}_N) - f(x^*) \leq \frac{LR^2}{2N},$$

   где $R = \|x_0 - x^*\|_2,\ x^*$ – ближайшее к x_0 решение.

   Для получения точности $\varepsilon > 0$ необходимо сделать $N = O\left(\frac{LR^2}{\varepsilon}\right)$ итераций, сублинейная скорость сходимости.

2. Пусть f – μ – **сильно выпуклая**, L–**гладкая**, $Q \in \mathbb{R}^n$ – выпукло и замкнуто, тогда

$$\|x^N - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^N \|x_0 - x_*\|_2^2.$$

Для получения точности $\varepsilon > 0$ необходимо сделать $N = O\left(\frac{L}{\mu} \ln \frac{\|x_0 - x_*\|_2^2}{\varepsilon}\right)$ итераций, линейная скорость сходимости.

Проксимальные методы

Рассмотрим общий вид задачи условной оптимизации:

$$f(x) \to \min_{x \in Q}, \tag{1}$$

где $Q \subseteq \mathbb{R}^n$ — выпуклое замкнутое множество.

$N(\varepsilon)$	выпуклость	μ-сильная выпуклость
L-гладкость	$\Omega\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$\Omega\left(\sqrt{\frac{L}{\mu}}\ln\frac{\mu R^2}{\varepsilon}\right)$
$\|\nabla f(x)\|_2 \le M$	$\Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$\Omega\left(\frac{M^2}{\mu\varepsilon}\right)$

Таблица 1: Нижние оценки на число итераций $N = N(\varepsilon)$, необходимых методу первого порядка для нахождения такой точки x^N, что $f(x^N) - f(x^*) \le \varepsilon$.

Все приведённые нижние оценки точны в том смысле, что существуют методы оптимизации с верхними оценками на необходимое число итераций, соответствующими этим нижним оценкам. Казалось бы, на этом можно заканчивать изучение оптимизации: оптимальные методы мы изучили, а ничего лучше *в данной постановке задачи* получить нельзя.

Пример

Рассмотрим задачу

$$F(x) = \underbrace{\frac{1}{2}x^\top \mathrm{A}x}_{f(x)} + \underbrace{\lambda\|x\|_1}_{R(x)} \to \min_{x \in \mathbb{R}^n}, \tag{2}$$

где $\mathrm{A} \in \mathbb{S}^n_+$ — симметричная положительно полуопределённая матрица, $\lambda > 0$.

- $f(x) = \frac{1}{2}x^\top \mathrm{A}x$ — выпуклая и L-гладкая функция с $L = \lambda_{\max}(\mathrm{A})$.
- $R(x) = \lambda\|x\|_1$ — выпуклая негладкая функция с ограниченными субградиентами: $\|\nabla R(x)\|_2 \le \lambda\sqrt{n}$, где $\nabla R(x)$ — произвольный субградиент функции $R(x)$ в точке x.

Единственный класс задач из четырёх классов, рассмотренных ранее, к которому мы можем отнести задачу (2) – это класс выпуклых функций с ограниченными субградиентами. Действительно, достаточно предполагать ограниченность субградиентов только на шаре $B_{2R}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \le 2R\}$, где $R = \|x^0 - x^*\|_2$. Поэтому можно ограничить $\nabla f(x)$ по норме некоторой константой на этом шаре. Пусть $\|\nabla F(x)\|_2 \le M$, тогда градиентный спуск с правильно выбранным размером шага (порядка $\frac{\varepsilon}{M^2}$) будет сходиться для данной задачи со скоростью $O\left(\frac{M^2R^2}{\varepsilon^2}\right)$.

НО полученная оценка не учитывает структуру задачи: мы полностью проигнорировали тот факт, что $f(x)$ имеет Липшицев градиент и что $R(x)$ достаточно «простая» функция. Оказывается для такого вида задач можно немного видоизменить градиентный спуск и получить

метод, который будет сходиться со скоростью $O(\frac{LR^2}{\varepsilon})$. Более того, можно получить ускоренный метод, который будет работать ещё быстрее — со скоростью $O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$. Но для начала нам нужно формально определить, с каким новым классом задач мы имеем дело.

Задачи выпуклой композитной оптимизации

(Задачи с регуляризацией)

Рассмотрим задачу

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}, \tag{3}$$

где

- $f(x) : \mathbb{R}^n \to \mathbb{R}$ — L-гладкая, выпуклая функция.
- $R(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{x \in \mathbb{R}^n \mid R(x) \leq \alpha\}$ замкнуто.

Пример 2.

1. $R(x) = 0 \Rightarrow \min\limits_{x \in \mathbb{R}^n} f(x)$
2. $R(x) = \delta_{\mathrm{Q}}(x) = \begin{cases} 0, & x \in \mathrm{Q} \\ +\infty, & x \notin \mathrm{Q} \end{cases} \quad \Rightarrow \quad \min\limits_{x \in \mathrm{Q}} f(x)$
3. $R(x) = \lambda \|x\|_1$

Проксимальный оператор и его свойства

Для функции $R(x)$, удовлетворяющей условиям (3), рассмотрим отображение из $\mathbb{R}^n$ в $\mathbb{R}^n$:

$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ R(y) + \frac{1}{2}\|y - x\|_2^2 \right\}. \tag{4}$$

Далее мы будем рассматривать только такие функции $R(x)$, относительно которых можно «быстро» вычислять проксимальный оператор.

1. $\mathrm{prox}_R(x)$ определяется однозначно для любого $x \in \mathbb{R}^n$, т.е. $\mathrm{prox}_R(x)$ — это отображение. Действительно, задача (4) является сильно выпуклой, а значит, имеет единственное решение.

2. $u = \operatorname{prox}_R(x) \overset{①}{\iff} x - u \in \partial R(u) \overset{②}{\iff} \langle x-u, y-u\rangle \le R(y) - R(u) \ \ \forall y \in \mathbb{R}^n$. Действительно, ① следует из необходимого и достаточного условия оптимальности первого порядка

$$u = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{R(y) + \frac{1}{2}\|y-x\|_2^2\right\} \iff 0 \in u - x + \partial R(u),$$

а ② следует из определения субградиента R в точке u.

3. Для всех $x, y \in \mathbb{R}^n$ выполняются неравенства:

$$\begin{aligned} \langle x-y, \operatorname{prox}_R(x) - \operatorname{prox}_R(y)\rangle &\ge \|\operatorname{prox}_R(x) - \operatorname{prox}_R(y)\|_2^2, &(5)\\ \|\operatorname{prox}_R(x) - \operatorname{prox}_R(y)\|_2 &\le \|x-y\|_2. &(6)\end{aligned}$$

Пусть $u = \operatorname{prox}_R(x)$, $v = \operatorname{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x-u, v-u\rangle \le R(v) - R(u)$ и $\langle y-v, u-v\rangle \le R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u-v, y-x+u-v\rangle \le 0$, что эквивалентно (5). Теперь покажем (6). Если $u = v$, то неравенство (6) очевидно. Если же $u \ne v$, то из (5) и неравенства Коши-Буняковского-Шварца получаем: $\|u-v\|_2^2 \le \langle x-y, u-v\rangle \le \|x-y\|_2 \cdot \|u-v\|_2$, что после сокращения левой и правой частей на $\|u-v\|_2$ даёт (6).

Примеры вычисления проксимальных операторов

1. $R(x) = c$, где $c \in \mathbb{R}$. Тогда

$$\operatorname{prox}_R(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{c + \frac{1}{2}\|y-x\|_2^2\right\} = x.$$

2. $R(x) = \delta_{\mathrm{Q}}(x) = \begin{cases} 0, & x \in \mathrm{Q}, \\ +\infty, & x \notin \mathrm{Q}, \end{cases}$ где $\mathrm{Q} \subseteq \mathbb{R}^n$ — выпуклое замкнутое непустое множество. Заметим, что минимум в определении проксимального оператора не может достигаться вне множества Q, т.к. функция под минимумом вне этого множества равняется $+\infty$. Поэтому

$$\begin{aligned} \operatorname{prox}_R(x) &= \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\delta_{\mathrm{Q}}(y) + \frac{1}{2}\|y-x\|_2^2\right\} = \operatorname*{argmin}_{y\in\mathrm{Q}}\left\{\frac{1}{2}\|y-x\|_2^2\right\} \\ &= \pi_{\mathrm{Q}}(x) \quad \text{– метод проекции градиента.}\end{aligned}$$

3. $R(x) = \frac{1}{2}x^\top \mathrm{A}x + \mathrm{b}^\top x + \mathrm{c}$, где $\mathrm{A} \in \mathbb{S}_+^n$, $\mathrm{b} \in \mathbb{R}^n$ и $\mathrm{c} \in \mathbb{R}$. Пусть $u = \operatorname{prox}_R(x)$. Тогда

$$u = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\frac{1}{2}y^\top \mathrm{A}y + \mathrm{b}^\top y + \mathrm{c} + \frac{1}{2}\|y-x\|_2^2\right\},$$

что по свойству 2 эквивалентно тому, что

$$x - u = \mathrm{A}u + \mathrm{b} \iff u = (\mathrm{A}+\mathrm{I})^{-1}(x-\mathrm{b}) = \operatorname{prox}_R(x).$$

4. $R(x) = \lambda\|x\|_1$, где $\lambda > 0$. Чтобы найти прокс-оператор от данной функции, докажем вспомогательное утверждение.

   Утверждение 1 (Прокс-оператор сепарабельной функции)**.** Пусть $R(x) = R(x_1, \ldots, x_r) = \sum\limits_{i=1}^{r} R_i(x_i)$, где $x = (x_1^\top, \ldots, x_r^\top)^\top \in \mathbb{R}^n$ и $x_i \in \mathbb{R}^{n_i}$ для $i = 1, \ldots, r$. Тогда

$$\operatorname{prox}_R(x) = \begin{pmatrix} \operatorname{prox}_{R_1}(x_1) \\ \vdots \\ \operatorname{prox}_{R_r}(x_r) \end{pmatrix}.$$

   Доказательство. По определению мы имеем

$$\begin{aligned} \operatorname{prox}_R(x) &= \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^{r} R_i(y_i) + \frac{1}{2}\|y - x\|_2^2 \right\} \\ &= \operatorname*{argmin}_{y_i \in \mathbb{R}^{n_i}, i=\overline{1,n}} \left\{ \sum_{i=1}^{r} \left(R_i(y_i) + \frac{1}{2}\|y_i - x_i\|_2^2 \right) \right\}. \end{aligned}$$

   Отсюда следует, что задача распадается на r независимых подзадач. Используя определение $\operatorname{prox}_{R_i}(x_i)$, получаем требуемое. □

   Возвращаясь к исходной задаче, замечаем, что $R(x) = \sum\limits_{i=1}^{n} \lambda|x_i|$, то есть достаточно найти прокс-оператор функции одного аргумента $g(x) = \lambda|x|$, $x \in \mathbb{R}$:

$$u = \operatorname{prox}_g(x) = \operatorname*{argmin}_{y \in \mathbb{R}} \left\{ \lambda|y| + \frac{1}{2}(y - x)^2 \right\}.$$

   - Минимум достигается при $y > 0$, тогда и только тогда, когда $\lambda + u - x = 0 \iff u = x - \lambda$. Это означает, что если $x > \lambda$, то $\operatorname{prox}_g(x) = x - \lambda$.
   - Аналогичными рассуждениями получаем, что если $x < -\lambda$, то $\operatorname{prox}_g(x) = x + \lambda$.
   - Во всех остальных случаях, т.е. при $x \in [-\lambda, \lambda]$, $\operatorname{prox}_g(x) = 0$.

Полученный результат можно записать в следующем виде:

$$\operatorname{prox}_g(x) = [|x| - \lambda]_+ \cdot \operatorname{sign}(x), \text{ где } \operatorname{sign}(x) \stackrel{\text{def}}{=} \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0 \end{cases}$$

и $[y]_+ \stackrel{\text{def}}{=} \max\{y, 0\}$. Отсюда следует, что для $R(x) = \lambda\|x\|_1$

$$\operatorname{prox}_R(x) = [|x| - \lambda\mathbf{1}]_+ \odot \operatorname{sign}(x),$$

где $\mathbf{1} \stackrel{\text{def}}{=} (1, \ldots, 1)^\top \in \mathbb{R}^n$, модуль $|x|$, срезка $[|x| - \lambda\mathbf{1}]_+$ и сигнум (знак) $\operatorname{sign}(x)$ применяются к векторам покомпонентно и $y \odot z \stackrel{\text{def}}{=} (y_1 z_1, \ldots, y_n z_n)^\top$ обозначает произведение Адамара двух векторов (покомпонентное произведение).

Проксимальный градиентный спуск

Для задачи (3) рассмотрим следующий метод.

Algorithm 2 Проксимальный градиентный спуск

Require: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N
1: **for** $k = 0, 1, \ldots, N-1$ **do**
2: Вычислить $\nabla f(x^k)$
3: $x^{k+1} = \operatorname{prox}_{\gamma R}\left(x^k - \gamma \nabla f(x^k)\right)$
4: **end for**
Ensure: x^N

- Внешне метод очень похож на градиентный спуск: по-прежнему требуется вычислять градиентный шаг, но теперь дополнительно от получаемой градиентным шагом точки вычисляется прокс.
- Пусть $x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} F(x)$. Тогда $x^* = \operatorname{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$.
- Пусть $f(x)$ — L-гладкая, тогда шаг проксимального метода можно выразить следующим образом:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|_2^2 + R(x) \right\} \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{L}{2} \left\| x - x_k + \frac{1}{L}\nabla f(x_k) \right\|_2^2 + R(x) \right\} \\ &= \operatorname{prox}_{\frac{1}{L}R(x)} \left(x_k - \frac{1}{L}\nabla f(x_k) \right) \end{aligned}$$

Сильно выпуклый случай

Теорема 3. Пусть $f(x)$ — μ-сильно выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция и $\gamma \leq \frac{1}{L}$. Тогда для любого $N > 0$ выход Алгоритма 2 удовлетворяет неравенству:

$$\|x^N - x^*\|_2^2 \leq (1 - \gamma\mu)^N \|x^0 - x^*\|_2^2. \tag{7}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ итераций, где $R = \|x^0 - x^*\|_2$, выполняется $\|x^N - x^*\|_2^2 \leq \varepsilon$.

Доказательство. Пользуясь тем, что прокс-оператор является нерастягивающим (см. (6)), мы получаем

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(x^k - \gamma \nabla f(x^k)) - \operatorname{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))\|_2^2 \\ &\overset{(6)}{\leq} \|x^k - x^* - \gamma\left(\nabla f(x^k) - \nabla f(x^*)\right)\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\ &\quad + \gamma^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2. \end{aligned}$$

Из сильной выпуклости функции f имеем

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\mu}{2}\|x^* - x^k\|_2^2,$$

откуда следует, что

$$-\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*) \rangle \leq -\frac{\mu}{2}\|x^k - x^*\|_2^2 - \underbrace{\left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle\right)}_{V_f(x^k, x^*)}.$$

Из полученных неравенств имеем:

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma\mu)\|x^k - x^*\|_2^2 - 2\gamma V_f(x^k, x^*) + \gamma^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2.$$

Кроме того, из L-гладкости и выпуклости функции f следует (см. Теорему 2.1.5 из книги Ю.Е. Нестерова, 2010 года), что для любых $x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right) = 2LV_f(x, y).$$

Используя это неравенство с $x = x^k$ и $y = x^*$, мы продолжаем наши цепочку неравенств для $\|x^{k+1} - x^*\|_2^2$:

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma\mu)\|x^k - x^*\|_2^2 - 2\gamma\left(1 - \gamma L\right)V_f(x^k, x^*)$$

Из выпуклости f имеем:

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

Кроме того, т.к. $\gamma > 0$ и $\gamma \leq \frac{1}{L}$, то $2\gamma\left(1 - \gamma L\right)V_f(x^k, x^*) \geq 0$, откуда следует, что

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma\mu)\|x^k - x^*\|_2^2.$$

Поскольку формула выше выполнена для всех целых $k \geq 0$, то отсюда следует, что

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma\mu)^{k+1}\|x^0 - x^*\|_2^2.$$

Наконец, подставляя $\gamma = \frac{1}{L}$ и пользуясь неравенством $(1 - t)^k \leq e^{-tk}$, получаем, что для достижения $\|x^N - x^*\|_2^2 \leq \varepsilon$ достаточно $N = \frac{L}{\mu}\ln\frac{\|x^0 - x^*\|_2^2}{\varepsilon}$ итераций данного метода. □

Выпуклый случай

Теорема 4. Пусть $f(x)$ — выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция и $\gamma = \frac{1}{L}$. Тогда для любого $N > 0$ выход Алгоритма 2 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \tag{8}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{LR^2}{\varepsilon}\right)$ итераций, где $R = \|x^0 - x^*\|_2$, выполняется $F(x^N) - F(x^*) \leq \varepsilon$.

Проксимальный ускоренный градиентный метод (FISTA)

FISTA: выпуклый случай

Пусть функция f выпукла. Тогда проксимальный градиентный метод можно ускорить следующим способом.

Algorithm 3 Проксимальный ускоренный градиентный метод (FISTA) для выпуклых функций

Require: стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N
1: $y^0 = x^0$, $t_0 = 1$
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: Вычислить $\nabla f(y^k)$
4: $x^{k+1} = \operatorname{prox}_{\frac{1}{L}R}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)$
5: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
6: $y^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}}(x^{k+1} - x^k)$
7: **end for**
Ensure: x^N

Теорема 5. Пусть $f(x)$ — выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция. Тогда для любого $N > 0$ выход Алгоритма 3 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{2L\|x^0 - x^*\|_2^2}{(N+1)^2}. \tag{9}$$

Иными словами, для FISTA через $N = O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций, где $R = \|x^0 - x^*\|_2$, выполняется $F(x^N) - F(x^*) \leq \varepsilon$.

FISTA: сильно выпуклый случай

Пусть теперь функция f μ-сильно выпукла.

Algorithm 4 Проксимальный ускоренный градиентный метод (FISTA) для сильно выпуклых функций

Require: стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N
1: $y^0 = x^0$, $\varkappa = \frac{L}{\mu}$
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: Вычислить $\nabla f(y^k)$
4: $x^{k+1} = \operatorname{prox}_{\frac{1}{L}R}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)$
5: $y^{k+1} = x^{k+1} + \frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}(x^{k+1} - x^k)$
6: **end for**
Ensure: x^N

Теорема 6. Пусть $f(x)$ — μ-сильно выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция. Тогда для любого $N > 0$ выход Алгоритма 4 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^N \left(F(x^0) - F(x^*) + \frac{\mu}{2}\|x^0 - x^*\|_2^2\right). \tag{10}$$

Иными словами, для FISTA через $N = O\left(\sqrt{\frac{L}{\mu}} \ln \frac{F(x^0)-F(x^*)+\mu R^2}{\varepsilon}\right)$ итераций, где $R = \|x^0 - x^*\|_2$, выполняется $F(x^N) - F(x^*) \leq \varepsilon$.