Задачи с регуляризацией, проксимальные методы

Оптимизация в машинном обучении

Эдуард Горбунов

Московский Физико-Технический Институт

5 ноября 2020

План лекции

- Сходимость (суб)градиентного спуска для выпуклых функций с ограниченным градиентом (субдифференциалом)
- Задачи с регуляризацией
- Проксимальный оператор
- Проксимальный градиентный спуск, сходимость в выпуклом и сильно выпуклом случаях
- Ускоренный проксимальный градиентный спуск (FISTA)

Постановка задачи

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

где

- $f(x)$ — выпуклая функция,
- $f(x)$ может быть недифференцируемой,
- метод имеет доступ к оракулу, который в запрашиваемой точке x возвращает произвольный субградиент $g(x)$ функции f в этой точке ($g(x) \in \partial f(x)$),
- существует такая константа $M > 0$, что для всех $x \in \mathbb{R}^n$ и для всех $g(x) \in \partial f(x)$ выполняется
$$\|g(x)\|_2 \leq M.$$

Напоминание

Субградиентом функции $f(x) : Q \to \mathbb{R}$ в точке x_0 называется такой вектор g, что для всех $x \in Q$ выполняется $f(x) - f(x_0) \geq \langle g, x - x_0 \rangle$.

Субградиентный спуск

Алгоритм 1 Субградиентный спуск

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

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

2: Вычислить произвольный субградиент $g(x^k)$ функции f в точке x^k

3: $x^{k+1} = x^k - \gamma g(x^k)$

4: **end for**

Выход: x^N

- Субградиентный спуск – это аналог градиентного спуска для негладких функций

Субградиентный спуск: скорость сходимости в выпуклом случае

Теорема 1

Предположим, что функция f удовлетворяет условиям с предыдущего слайда и имеет непустое множество решений. Пусть x^* — ближайшее решение к x^0, $R_0 = \|x^0 - x^*\|_2$ и $\gamma = \frac{R_0}{M\sqrt{N+1}}$. Тогда через N итераций субградиентного спуска имеем

$$f(\overline{x}^N) - f(x^*) \leq \frac{MR_0}{\sqrt{N+1}}, \tag{1}$$

где $\overline{x}^N = \frac{1}{N+1}\sum\limits_{k=0}^{N} x^k$. Кроме того, для достижения точности ε по функции ($f(\overline{x}^N) - f(x^*) \leq \varepsilon$) метод достаточно запустить на $N = O\left(\frac{M^2R_0^2}{\varepsilon^2}\right)$ итераций.

Субградиентный спуск: скорость сходимости в выпуклом случае

Доказательство Теоремы 1. Для удобства введём следующее обозначение: $R_k = \|\mathrm{x}^k - \mathrm{x}^*\|_2$. Рассмотрим R_{k+1}:

$$\begin{aligned} R_{k+1}^2 &= \|\mathrm{x}^k - \mathrm{x}^* - \gamma \mathrm{g}(\mathrm{x}^k)\|_2^2 \\ &= \|\mathrm{x}^k - \mathrm{x}^*\|_2^2 + 2\gamma\langle \mathrm{x}^* - \mathrm{x}^k, \mathrm{g}(\mathrm{x}^k)\rangle + \gamma^2\|\mathrm{g}(\mathrm{x}^k)\|_2^2 \\ &\leq R_k^2 + 2\gamma\left(f(\mathrm{x}^*) - f(\mathrm{x}^k)\right) + \gamma^2 M^2, \end{aligned}$$

где в последнем переходе мы воспользовались определением субградиента функции f в точке x^k и нашим предположением, что $\|\mathrm{g}(\mathrm{x}^k)\|_2 \leq M$. Перепишем полученное неравенство в другом виде:

$$f(\mathrm{x}^k) - f(\mathrm{x}^*) \leq \frac{R_k^2}{2\gamma} - \frac{R_{k+1}^2}{2\gamma} + \frac{\gamma M^2}{2}. \qquad (2)$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Доказательство Теоремы 1. Для удобства введём следующее обозначение: $R_k = \|x^k - x^*\|_2$. Рассмотрим R_{k+1}:

$$\begin{aligned} R_{k+1}^2 &= \|x^k - x^* - \gamma g(x^k)\|_2^2 \\ &= \|x^k - x^*\|_2^2 + 2\gamma\langle x^* - x^k, g(x^k)\rangle + \gamma^2\|g(x^k)\|_2^2 \\ &\leq R_k^2 + 2\gamma\left(f(x^*) - f(x^k)\right) + \gamma^2 M^2, \end{aligned}$$

где в последнем переходе мы воспользовались определением субградиента функции f в точке x^k и нашим предположением, что $\|g(x^k)\|_2 \leq M$. Перепишем полученное неравенство в другом виде:

$$f(x^k) - f(x^*) \leq \frac{R_k^2}{2\gamma} - \frac{R_{k+1}^2}{2\gamma} + \frac{\gamma M^2}{2}. \quad (2)$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Доказательство Теоремы 1. Для удобства введём следующее обозначение: $R_k = \|x^k - x^*\|_2$. Рассмотрим R_{k+1}:

$$\begin{aligned} R_{k+1}^2 &= \|x^k - x^* - \gamma g(x^k)\|_2^2 \\ &= \|x^k - x^*\|_2^2 + 2\gamma\langle x^* - x^k, g(x^k)\rangle + \gamma^2\|g(x^k)\|_2^2 \\ &\leq R_k^2 + 2\gamma\left(f(x^*) - f(x^k)\right) + \gamma^2 M^2, \end{aligned}$$

где в последнем переходе мы воспользовались определением субградиента функции f в точке x^k и нашим предположением, что $\|g(x^k)\|_2 \leq M$. Перепишем полученное неравенство в другом виде:

$$f(x^k) - f(x^*) \leq \frac{R_k^2}{2\gamma} - \frac{R_{k+1}^2}{2\gamma} + \frac{\gamma M^2}{2}. \qquad (2)$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Доказательство Теоремы 1. Для удобства введём следующее обозначение: $R_k = \|x^k - x^*\|_2$. Рассмотрим R_{k+1}:

$$\begin{aligned} R_{k+1}^2 &= \|x^k - x^* - \gamma g(x^k)\|_2^2 \\ &= \|x^k - x^*\|_2^2 + 2\gamma\langle x^* - x^k, g(x^k)\rangle + \gamma^2\|g(x^k)\|_2^2 \\ &\le R_k^2 + 2\gamma\left(f(x^*) - f(x^k)\right) + \gamma^2 M^2, \end{aligned}$$

где в последнем переходе мы воспользовались определением субградиента функции f в точке x^k и нашим предположением, что $\|g(x^k)\|_2 \le M$. Перепишем полученное неравенство в другом виде:

$$f(x^k) - f(x^*) \le \frac{R_k^2}{2\gamma} - \frac{R_{k+1}^2}{2\gamma} + \frac{\gamma M^2}{2}. \tag{2}$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Просуммируем неравенства (2) для $k = 0, 1, \ldots, N$ и поделим результат на $N+1$:

$$\begin{aligned}\frac{1}{N+1}\sum_{k=0}^{N}\left(f(\mathrm{x}^k) - f(\mathrm{x}^*)\right) &\leq \frac{1}{2\gamma(N+1)}\sum_{k=0}^{N}\left(R_k^2 - R_{k+1}^2\right) + \frac{\gamma M^2}{2}\\ &= \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}.\end{aligned}$$

Рассмотрим точку $\overline{\mathrm{x}}^N = \frac{1}{N+1}\sum\limits_{k=0}^{N}\mathrm{x}^k$. Поскольку функция $f(x)$ выпукла, из неравенства Йенсена следует: $f(\overline{\mathrm{x}}^N) \leq \frac{1}{N+1}\sum\limits_{k=0}^{N} f(\mathrm{x}^k)$.

Субградиентный спуск: скорость сходимости в выпуклом случае

Просуммируем неравенства (2) для $k = 0, 1, \ldots, N$ и поделим результат на $N+1$:

$$
\begin{aligned}
\frac{1}{N+1}\sum_{k=0}^{N}\left(f(x^k) - f(x^*)\right) &\leq \frac{1}{2\gamma(N+1)}\sum_{k=0}^{N}\left(R_k^2 - R_{k+1}^2\right) + \frac{\gamma M^2}{2} \\
&= \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}.
\end{aligned}
$$

Рассмотрим точку $\overline{x}^N = \frac{1}{N+1}\sum\limits_{k=0}^{N} x^k$. Поскольку функция $f(x)$ выпукла, из неравенства Йенсена следует: $f(\overline{x}^N) \leq \frac{1}{N+1}\sum\limits_{k=0}^{N} f(x^k)$.

Субградиентный спуск: скорость сходимости в выпуклом случае

Просуммируем неравенства (2) для $k = 0, 1, \ldots, N$ и поделим результат на $N + 1$:

$$
\begin{aligned}
\frac{1}{N+1}\sum_{k=0}^{N}\left(f(\mathrm{x}^k) - f(\mathrm{x}^*)\right) &\le \frac{1}{2\gamma(N+1)}\sum_{k=0}^{N}\left(R_k^2 - R_{k+1}^2\right) + \frac{\gamma M^2}{2} \\
&= \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}.
\end{aligned}
$$

Рассмотрим точку $\overline{\mathrm{x}}^N = \frac{1}{N+1}\sum\limits_{k=0}^{N}\mathrm{x}^k$. Поскольку функция $f(x)$ выпукла, из неравенства Йенсена следует: $f(\overline{\mathrm{x}}^N) \le \frac{1}{N+1}\sum\limits_{k=0}^{N} f(\mathrm{x}^k)$.

Субградиентный спуск: скорость сходимости в выпуклом случае

Из полученных неравенств и $R_{k+1}^2 \geq 0$ следует, что

$$\begin{aligned} f(\overline{x}^N) - f(x^*) &\leq \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2} \\ &\leq \frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}. \end{aligned} \tag{3}$$

Выражение в правой части является выпуклой дифференцируемой функцией от γ. Минимизируя это выражение по γ, получаем

$$\left(\frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}\right)'_\gamma = 0 \implies \gamma = \frac{R_0}{M\sqrt{N+1}}.$$

Подставляя это значение γ в неравенство выше, мы заключаем, что

$$f(\overline{x}^N) - f(x^*) \leq \frac{MR_0}{\sqrt{N+1}}.$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Из полученных неравенств и $R_{k+1}^2 \geq 0$ следует, что

$$\begin{aligned} f(\bar{x}^N) - f(x^*) &\leq \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2} \\ &\leq \frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}. \end{aligned} \quad (3)$$

Выражение в правой части является выпуклой дифференцируемой функцией от γ. Минимизируя это выражение по γ, получаем

$$\left(\frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}\right)'_\gamma = 0 \implies \gamma = \frac{R_0}{M\sqrt{N+1}}.$$

Подставляя это значение γ в неравенство выше, мы заключаем, что

$$f(\bar{x}^N) - f(x^*) \leq \frac{MR_0}{\sqrt{N+1}}.$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Из полученных неравенств и $R_{k+1}^2 \geq 0$ следует, что

$$f(\bar{x}^N) - f(x^*) \leq \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2} \leq \frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}. \tag{3}$$

Выражение в правой части является выпуклой дифференцируемой функцией от γ. Минимизируя это выражение по γ, получаем

$$\left(\frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}\right)'_{\gamma} = 0 \implies \gamma = \frac{R_0}{M\sqrt{N+1}}.$$

Подставляя это значение γ в неравенство выше, мы заключаем, что

$$f(\bar{x}^N) - f(x^*) \leq \frac{MR_0}{\sqrt{N+1}}.$$

Субградиентный спуск: скорость сходимости в выпуклом случае

Из полученных неравенств и $R_{k+1}^2 \geq 0$ следует, что

$$\begin{aligned} f(\overline{x}^N) - f(x^*) &\leq \frac{R_0^2}{2\gamma(N+1)} - \frac{R_{N+1}^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2} \\ &\leq \frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}. \end{aligned} \tag{3}$$

Выражение в правой части является выпуклой дифференцируемой функцией от γ. Минимизируя это выражение по γ, получаем

$$\left(\frac{R_0^2}{2\gamma(N+1)} + \frac{\gamma M^2}{2}\right)'_\gamma = 0 \implies \gamma = \frac{R_0}{M\sqrt{N+1}}.$$

Подставляя это значение γ в неравенство выше, мы заключаем, что

$$f(\overline{x}^N) - f(x^*) \leq \frac{MR_0}{\sqrt{N+1}}.$$

Субградиентный спуск: наблюдения

Условия теоремы можно немного ослабить. Во-первых, ограниченность субградиентов нам потребовалась только в точках x^k, где $k = 0, 1, \ldots, N$. Во-вторых, из неравенства (3) и $f(\overline{x}^N) - f(x^*) \geq 0$ мы получаем $R_{N+1}^2 \leq R_0^2 + (N+1)\gamma^2 M^2$. На самом деле, точно такими же рассуждениями мы можем получить аналогичное неравенство для всех $k = 0, 1, \ldots, N+1$:

$$R_k^2 \leq R_0^2 + k\gamma^2 M^2 = R_0^2 + \frac{k}{N+1} R_0^2 \leq 2R_0^2.$$

Следовательно, за $N+1$ итерацию субградиентного спуска с шагом $\gamma = \frac{R_0}{M\sqrt{N+1}}$ генерируемые точки x^k лежат в шаре с центром в x^* радиуса $\sqrt{2}R_0$ для всех $k = 0, 1, \ldots, N+1$, т.е. в некотором компакте. **Таким образом, условие ограниченности субградиентов достаточно потребовать только на этом шаре.**

Оказывается, что для рассмотренного класса задач полученная оценка **неулучшаема**.

Субградиентный спуск: наблюдения

Условия теоремы можно немного ослабить. Во-первых, ограниченность субградиентов нам потребовалась только в точках x^k, где $k = 0, 1, \ldots, N$. Во-вторых, из неравенства (3) и $f(\bar{x}^N) - f(x^*) \geq 0$ мы получаем $R_{N+1}^2 \leq R_0^2 + (N+1)\gamma^2 M^2$. На самом деле, точно такими же рассуждениями мы можем получить аналогичное неравенство для всех $k = 0, 1, \ldots, N+1$:

$$R_k^2 \leq R_0^2 + k\gamma^2 M^2 = R_0^2 + \frac{k}{N+1} R_0^2 \leq 2R_0^2.$$

Следовательно, за $N+1$ итерацию субградиентного спуска с шагом $\gamma = \frac{R_0}{M\sqrt{N+1}}$ генерируемые точки x^k лежат в шаре с центром в x^* радиуса $\sqrt{2}R_0$ для всех $k = 0, 1, \ldots, N+1$, т.е. в некотором компакте. **Таким образом, условие ограниченности субградиентов достаточно потребовать только на этом шаре.**

Оказывается, что для рассмотренного класса задач полученная оценка **неулучшаема**.

Субградиентный спуск: наблюдения

Условия теоремы можно немного ослабить. Во-первых, ограниченность субградиентов нам потребовалась только в точках x^k, где $k = 0, 1, \ldots, N$. Во-вторых, из неравенства (3) и $f(\bar{x}^N) - f(x^*) \geq 0$ мы получаем $R_{N+1}^2 \leq R_0^2 + (N+1)\gamma^2 M^2$. На самом деле, точно такими же рассуждениями мы можем получить аналогичное неравенство для всех $k = 0, 1, \ldots, N+1$:

$$R_k^2 \leq R_0^2 + k\gamma^2 M^2 = R_0^2 + \frac{k}{N+1} R_0^2 \leq 2R_0^2.$$

Следовательно, за $N+1$ итерацию субградиентного спуска с шагом $\gamma = \frac{R_0}{M\sqrt{N+1}}$ генерируемые точки x^k лежат в шаре с центром в x^* радиуса $\sqrt{2}R_0$ для всех $k = 0, 1, \ldots, N+1$, т.е. в некотором компакте. **Таким образом, условие ограниченности субградиентов достаточно потребовать только на этом шаре.**

Оказывается, что для рассмотренного класса задач полученная оценка **неулучшаема**.

Субградиентный спуск: наблюдения

Условия теоремы можно немного ослабить. Во-первых, ограниченность субградиентов нам потребовалась только в точках x^k, где $k = 0, 1, \ldots, N$. Во-вторых, из неравенства (3) и $f(\bar{x}^N) - f(x^*) \geq 0$ мы получаем $R_{N+1}^2 \leq R_0^2 + (N+1)\gamma^2 M^2$. На самом деле, точно такими же рассуждениями мы можем получить аналогичное неравенство для всех $k = 0, 1, \ldots, N+1$:

$$R_k^2 \leq R_0^2 + k\gamma^2 M^2 = R_0^2 + \frac{k}{N+1} R_0^2 \leq 2R_0^2.$$

Следовательно, за $N+1$ итерацию субградиентного спуска с шагом $\gamma = \frac{R_0}{M\sqrt{N+1}}$ генерируемые точки x^k лежат в шаре с центром в x^* радиуса $\sqrt{2}R_0$ для всех $k = 0, 1, \ldots, N+1$, т.е. в некотором компакте. **Таким образом, условие ограниченности субградиентов достаточно потребовать только на этом шаре.**

Оказывается, что для рассмотренного класса задач полученная оценка **неулучшаема**.

Субградиентный спуск: наблюдения

Условия теоремы можно немного ослабить. Во-первых, ограниченность субградиентов нам потребовалась только в точках x^k, где $k = 0, 1, \ldots, N$. Во-вторых, из неравенства (3) и $f(\bar{x}^N) - f(x^*) \geq 0$ мы получаем $R_{N+1}^2 \leq R_0^2 + (N+1)\gamma^2 M^2$. На самом деле, точно такими же рассуждениями мы можем получить аналогичное неравенство для всех $k = 0, 1, \ldots, N+1$:

$$R_k^2 \leq R_0^2 + k\gamma^2 M^2 = R_0^2 + \frac{k}{N+1} R_0^2 \leq 2R_0^2.$$

Следовательно, за $N+1$ итерацию субградиентного спуска с шагом $\gamma = \frac{R_0}{M\sqrt{N+1}}$ генерируемые точки x^k лежат в шаре с центром в x^* радиуса $\sqrt{2}R_0$ для всех $k = 0, 1, \ldots, N+1$, т.е. в некотором компакте. **Таким образом, условие ограниченности субградиентов достаточно потребовать только на этом шаре.**

Оказывается, что для рассмотренного класса задач полученная оценка **неулучшаема**.

Краткий обзор того, что было в нашем курсе

До этого мы изучали задачи вида

$$f(x) \to \min_{x \in \mathbb{R}^n}. \tag{4}$$

$N(\varepsilon)$	выпуклость	μ-сильная выпуклость
L-гладкость	$\Omega\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right)$	$\Omega\left(\sqrt{\frac{L}{\mu}}\ln\frac{\mu R_0^2}{\varepsilon}\right)$
$\\|g(x)\\|_2 \leq M$	$\Omega\left(\frac{M^2R_0^2}{\varepsilon^2}\right)$	$\Omega\left(\frac{M^2}{\mu\varepsilon}\right)$

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

Про нижние оценки мы, к сожалению, не успеем поговорить в нашем курсе (очень хорошо про это написано в книге Ю.Е. Нестерова 2010 года (см. разделы 2.1.2, 2.1.4 и 3.2.1). Но формальные доказательства для оптимальных методов мы обсудим далее в курсе.

Краткий обзор того, что было в нашем курсе

До этого мы изучали задачи вида

$$f(x) \to \min_{x \in \mathbb{R}^n}. \tag{4}$$

$N(\varepsilon)$	выпуклость	μ-сильная выпуклость
L-гладкость	$\Omega\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right)$	$\Omega\left(\sqrt{\frac{L}{\mu}}\ln\frac{\mu R_0^2}{\varepsilon}\right)$
$\|g(x)\|_2 \leq M$	$\Omega\left(\frac{M^2R_0^2}{\varepsilon^2}\right)$	$\Omega\left(\frac{M^2}{\mu\varepsilon}\right)$

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

Про нижние оценки мы, к сожалению, не успеем поговорить в нашем курсе (очень хорошо про это написано в книге Ю.Е. Нестерова 2010 года (см. разделы 2.1.2, 2.1.4 и 3.2.1). Но формальные доказательства для оптимальных методов мы обсудим далее в курсе.

Краткий обзор того, что было в нашем курсе

Все приведённые на предыдущем слайде 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 уже знаем, а ничего лучше *в данной постановке задачи* получить нельзя.

Краткий обзор того, что было в нашем курсе

Все приведённые на предыдущем слайде 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 уже знаем, а ничего лучше *в данной постановке задачи* получить нельзя.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Рассмотрим задачу

$$F(x) = \underbrace{\frac{1}{2}x^\top Ax}_{f(x)} + \underbrace{\lambda\|x\|_1}_{R(x)} \to \min_{x\in\mathbb{R}^n}, \tag{5}$$

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

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

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Рассмотрим задачу

$$F(x) = \underbrace{\frac{1}{2}x^\top Ax}_{f(x)} + \underbrace{\lambda\|x\|_1}_{R(x)} \to \min_{x\in\mathbb{R}^n}, \qquad (5)$$

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

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

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Рассмотрим задачу

$$F(x) = \underbrace{\frac{1}{2}x^\top Ax}_{f(x)} + \underbrace{\lambda\|x\|_1}_{R(x)} \to \min_{x\in\mathbb{R}^n}, \tag{5}$$

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

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

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Единственный класс задач из четырёх классов, рассмотренных ранее, к которому мы можем отнести задачу (5) – это класс выпуклых функций с ограниченными субградиентами. Действительно, достаточно предполагать ограниченность субградиентов только на шаре $B_{\sqrt{2}R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq \sqrt{2}R_0\}$, где $R_0 = \|x^0 - x^*\|_2$. Поэтому можно ограничить $\nabla f(x)$ по норме некоторой константой на этом шаре. Пусть $\|\nabla F(x)\|_2 \leq M$, тогда градиентный спуск с правильно выбранным размером шага (порядка $\frac{\varepsilon}{M^2}$) будет сходиться для данной задачи со скоростью $O\left(\frac{M^2 R_0^2}{\varepsilon^2}\right)$.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Единственный класс задач из четырёх классов, рассмотренных ранее, к которому мы можем отнести задачу (5) – это класс выпуклых функций с ограниченными субградиентами. Действительно, достаточно предполагать ограниченность субградиентов только на шаре $B_{\sqrt{2}R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq \sqrt{2}R_0\}$, где $R_0 = \|x^0 - x^*\|_2$. Поэтому можно ограничить $\nabla f(x)$ по норме некоторой константой на этом шаре. Пусть $\|\nabla F(x)\|_2 \leq M$, тогда градиентный спуск с правильно выбранным размером шага (порядка $\frac{\varepsilon}{M^2}$) будет сходиться для данной задачи со скоростью $O\left(\frac{M^2 R_0^2}{\varepsilon^2}\right)$.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Единственный класс задач из четырёх классов, рассмотренных ранее, к которому мы можем отнести задачу (5) – это класс выпуклых функций с ограниченными субградиентами. Действительно, достаточно предполагать ограниченность субградиентов только на шаре $B_{\sqrt{2}R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq \sqrt{2}R_0\}$, где $R_0 = \|x^0 - x^*\|_2$. Поэтому можно ограничить $\nabla f(x)$ по норме некоторой константой на этом шаре. Пусть $\|\nabla F(x)\|_2 \leq M$, тогда градиентный спуск с правильно выбранным размером шага (порядка $\frac{\varepsilon}{M^2}$) будет сходиться для данной задачи со скоростью $O\left(\frac{M^2 R_0^2}{\varepsilon^2}\right)$.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

Единственный класс задач из четырёх классов, рассмотренных ранее, к которому мы можем отнести задачу (5) – это класс выпуклых функций с ограниченными субградиентами. Действительно, достаточно предполагать ограниченность субградиентов только на шаре $B_{\sqrt{2}R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_2 \leq \sqrt{2}R_0\}$, где $R_0 = \|x^0 - x^*\|_2$. Поэтому можно ограничить $\nabla f(x)$ по норме некоторой константой на этом шаре. Пусть $\|\nabla F(x)\|_2 \leq M$, тогда градиентный спуск с правильно выбранным размером шага (порядка $\frac{\varepsilon}{M^2}$) будет сходиться для данной задачи со скоростью $O\left(\frac{M^2 R_0^2}{\varepsilon^2}\right)$.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

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 $O(\frac{LR_0^2}{\varepsilon})$. Более того, можно получить ускоренный метод, который будет работать ещё быстрее — со скоростью $O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right)$. Но для начала нам нужно формально определить, с каким новым классом задач мы имеем дело.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

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 $O(\frac{LR_0^2}{\varepsilon})$. Более того, можно получить ускоренный метод, который будет работать ещё быстрее — со скоростью $O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right)$. Но для начала нам нужно формально определить, с каким новым классом задач мы имеем дело.

Пример: минимизация суммы квадратичной функции и ℓ_1-нормы

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 $O(\frac{LR_0^2}{\varepsilon})$. Более того, можно получить ускоренный метод, который будет работать ещё быстрее — со скоростью $O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right)$. Но для начала нам нужно формально определить, с каким новым классом задач мы имеем дело.

Задачи с регуляризацией

Рассмотрим задачу

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}, \quad (6)$$

где

- $f(x) : \mathbb{R}^n \to \mathbb{R}$ — L-гладкая функция. В этой лекции мы, кроме того, будем всегда считать, что f — выпуклая.
- $R(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — 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.

Задачи с регуляризацией

Рассмотрим задачу

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}, \tag{6}$$

где

- $f(x) : \mathbb{R}^n \to \mathbb{R}$ — L-гладкая функция. В этой лекции мы, кроме того, будем всегда считать, что f — выпуклая.
- $R(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — 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.

Задачи с регуляризацией

Рассмотрим задачу

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}, \tag{6}$$

где

- $f(x) : \mathbb{R}^n \to \mathbb{R}$ — L-гладкая функция. В этой лекции мы, кроме того, будем всегда считать, что f — выпуклая.
- $R(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — 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 $\{x \in \mathbb{R}^n \mid R(x) \le \alpha\}$ замкнуто.

Задачи с регуляризацией

Рассмотрим задачу

$$F(x) = f(x) + R(x) \to \min_{x \in \mathbb{R}^n}, \tag{6}$$

где

- $f(x) : \mathbb{R}^n \to \mathbb{R}$ — L-гладкая функция. В этой лекции мы, кроме того, будем всегда считать, что f — выпуклая.
- $R(x) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — 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 $\{x \in \mathbb{R}^n \mid R(x) \le \alpha\}$ замкнуто.

Проксимальный оператор и его свойства

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

$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ R(y) + \frac{1}{2}\|y-x\|_2^2 \right\}. \tag{7}$$

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, функция $R(y) + \frac{1}{2}\|y-x\|_2^2$ является правильной, замкнутой и сильно выпуклой, а значит, имеет единственное решение (см. Теорему 5.25 из книги Амира Бека 2017 года).

Проксимальный оператор и его свойства

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

$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ R(y) + \frac{1}{2}\|y - x\|_2^2 \right\}. \tag{7}$$

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, функция $R(y) + \frac{1}{2}\|y - x\|_2^2$ является правильной, замкнутой и сильно выпуклой, а значит, имеет единственное решение (см. Теорему 5.25 из книги Амира Бека 2017 года).

Проксимальный оператор и его свойства

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

$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ R(y) + \frac{1}{2}\|y-x\|_2^2 \right\}. \tag{7}$$

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, функция $R(y) + \frac{1}{2}\|y-x\|_2^2$ является правильной, замкнутой и сильно выпуклой, а значит, имеет единственное решение (см. Теорему 5.25 из книги Амира Бека 2017 года).

Проксимальный оператор и его свойства

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

$$\mathrm{prox}_R(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ R(y) + \frac{1}{2}\|y-x\|_2^2 \right\}. \tag{7}$$

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, функция $R(y) + \frac{1}{2}\|y-x\|_2^2$ является правильной, замкнутой и сильно выпуклой, а значит, имеет единственное решение (см. Теорему 5.25 из книги Амира Бека 2017 года).

Проксимальный оператор и его свойства

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

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

а ② следует из определения субградиента R в точке u.

Проксимальный оператор и его свойства

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

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

а ② следует из определения субградиента R в точке u.

Проксимальный оператор и его свойства

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

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

   а ② следует из определения субградиента R в точке u.

Проксимальный оператор и его свойства

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

$$\langle x - y, \mathrm{prox}_R(x) - \mathrm{prox}_R(y)\rangle \geq \|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2^2, \quad (8)$$
$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2 \leq \|x - y\|_2. \quad (9)$$

Пусть $u = \mathrm{prox}_R(x)$, $v = \mathrm{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x - u, v - u\rangle \leq R(v) - R(u)$ и $\langle y - v, u - v\rangle \leq R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u - v, y - x + u - v\rangle \leq 0$, что эквивалентно (8). Теперь покажем (9). Если $u = v$, то неравенство (9) очевидно. Если же $u \neq v$, то из (8) и неравенства Коши-Буняковского-Шварца получаем: $\|u - v\|_2^2 \leq \langle x - y, u - v\rangle \leq \|x - y\|_2 \cdot \|u - v\|_2$, что после сокращения левой и правой частей на $\|u - v\|_2$ даёт (9).

Проксимальный оператор и его свойства

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

$$\langle x - y, \mathrm{prox}_R(x) - \mathrm{prox}_R(y)\rangle \geq \|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2^2, \tag{8}$$

$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2 \leq \|x - y\|_2. \tag{9}$$

Пусть $u = \mathrm{prox}_R(x)$, $v = \mathrm{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x - u, v - u\rangle \leq R(v) - R(u)$ и $\langle y - v, u - v\rangle \leq R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u - v, y - x + u - v\rangle \leq 0$, что эквивалентно (8). Теперь покажем (9). Если $u = v$, то неравенство (9) очевидно. Если же $u \neq v$, то из (8) и неравенства Коши-Буняковского-Шварца получаем: $\|u - v\|_2^2 \leq \langle x - y, u - v\rangle \leq \|x - y\|_2 \cdot \|u - v\|_2$, что после сокращения левой и правой частей на $\|u - v\|_2$ даёт (9).

Проксимальный оператор и его свойства

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

$$\langle x - y, \mathrm{prox}_R(x) - \mathrm{prox}_R(y)\rangle \geq \|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2^2, \quad (8)$$

$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2 \leq \|x - y\|_2. \quad (9)$$

Пусть $u = \mathrm{prox}_R(x)$, $v = \mathrm{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x - u, v - u\rangle \leq R(v) - R(u)$ и $\langle y - v, u - v\rangle \leq R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u - v, y - x + u - v\rangle \leq 0$, что эквивалентно (8). Теперь покажем (9). Если $u = v$, то неравенство (9) очевидно. Если же $u \neq v$, то из (8) и неравенства Коши-Буняковского-Шварца получаем: $\|u - v\|_2^2 \leq \langle x - y, u - v\rangle \leq \|x - y\|_2 \cdot \|u - v\|_2$, что после сокращения левой и правой частей на $\|u - v\|_2$ даёт (9).

Проксимальный оператор и его свойства

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

$$\langle x - y, \mathrm{prox}_R(x) - \mathrm{prox}_R(y)\rangle \geq \|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2^2, \quad (8)$$
$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2 \leq \|x - y\|_2. \quad (9)$$

Пусть $u = \mathrm{prox}_R(x)$, $v = \mathrm{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x - u, v - u\rangle \leq R(v) - R(u)$ и $\langle y - v, u - v\rangle \leq R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u - v, y - x + u - v\rangle \leq 0$, что эквивалентно (8). Теперь покажем (9). Если $u = v$, то неравенство (9) очевидно. Если же $u \neq v$, то из (8) и неравенства Коши-Буняковского-Шварца получаем: $\|u - v\|_2^2 \leq \langle x - y, u - v\rangle \leq \|x - y\|_2 \cdot \|u - v\|_2$, что после сокращения левой и правой частей на $\|u - v\|_2$ даёт (9).

Проксимальный оператор и его свойства

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

$$\langle x - y, \mathrm{prox}_R(x) - \mathrm{prox}_R(y)\rangle \geq \|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2^2, \quad (8)$$
$$\|\mathrm{prox}_R(x) - \mathrm{prox}_R(y)\|_2 \leq \|x - y\|_2. \quad (9)$$

Пусть $u = \mathrm{prox}_R(x)$, $v = \mathrm{prox}_R(y)$, тогда по свойству 2 имеем: $\langle x - u, v - u\rangle \leq R(v) - R(u)$ и $\langle y - v, u - v\rangle \leq R(u) - R(v)$. Складывая эти неравенства, получаем: $\langle u - v, y - x + u - v\rangle \leq 0$, что эквивалентно (8). Теперь покажем (9). Если $u = v$, то неравенство (9) очевидно. Если же $u \neq v$, то из (8) и неравенства Коши-Буняковского-Шварца получаем: $\|u - v\|_2^2 \leq \langle x - y, u - v\rangle \leq \|x - y\|_2 \cdot \|u - v\|_2$, что после сокращения левой и правой частей на $\|u - v\|_2$ даёт (9).

Примеры вычисления проксимальных операторов

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

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

2. $R(x) = \delta_Q(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q, \end{cases}$ где $Q \subseteq \mathbb{R}^n$ — 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

$$\begin{aligned} \mathrm{prox}_R(x) &= \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \delta_Q(y) + \frac{1}{2}\|y - x\|_2^2 \right\} = \underset{y \in Q}{\operatorname{argmin}} \left\{ \frac{1}{2}\|y - x\|_2^2 \right\} \\ &= \pi_Q(x). \end{aligned}$$

Примеры вычисления проксимальных операторов

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

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

2. $R(x) = \delta_Q(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q, \end{cases}$ где $Q \subseteq \mathbb{R}^n$ — 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

$$\begin{aligned} \mathrm{prox}_R(x) &= \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \delta_Q(y) + \frac{1}{2}\|y - x\|_2^2 \right\} = \operatorname*{argmin}_{y \in Q} \left\{ \frac{1}{2}\|y - x\|_2^2 \right\} \\ &= \pi_Q(x). \end{aligned}$$

Примеры вычисления проксимальных операторов

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

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

2. $R(x) = \delta_Q(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q, \end{cases}$ где $Q \subseteq \mathbb{R}^n$ — 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

$$\begin{aligned} \mathrm{prox}_R(x) &= \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \delta_Q(y) + \frac{1}{2}\|y - x\|_2^2 \right\} = \operatorname*{argmin}_{y \in Q} \left\{ \frac{1}{2}\|y - x\|_2^2 \right\} \\ &= \pi_Q(x). \end{aligned}$$

Примеры вычисления проксимальных операторов

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

2. $R(x) = \delta_Q(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q, \end{cases}$ где $Q \subseteq \mathbb{R}^n$ — 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
$$\begin{aligned}\mathrm{prox}_R(x) &= \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\delta_Q(y) + \frac{1}{2}\|y - x\|_2^2\right\} = \operatorname*{argmin}_{y\in Q}\left\{\frac{1}{2}\|y - x\|_2^2\right\} \\ &= \pi_Q(x).\end{aligned}$$

Примеры вычисления проксимальных операторов

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

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

2. $R(x) = \delta_Q(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q, \end{cases}$ где $Q \subseteq \mathbb{R}^n$ — 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

$$\begin{aligned}\mathrm{prox}_R(x) &= \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\delta_Q(y) + \frac{1}{2}\|y - x\|_2^2\right\} = \operatorname*{argmin}_{y\in Q}\left\{\frac{1}{2}\|y - x\|_2^2\right\} \\ &= \pi_Q(x).\end{aligned}$$

Примеры вычисления проксимальных операторов

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

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

что по свойству 2 эквивалентно тому, что

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

Примеры вычисления проксимальных операторов

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

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

что по свойству 2 эквивалентно тому, что

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

Примеры вычисления проксимальных операторов

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

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

что по свойству 2 эквивалентно тому, что

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

Примеры вычисления проксимальных операторов

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

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

что по свойству 2 эквивалентно тому, что

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

Примеры вычисления проксимальных операторов

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

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

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

Примеры вычисления проксимальных операторов

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

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

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

Примеры вычисления проксимальных операторов

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

Прокс-оператор сепарабельной функции

Пусть $R(\mathrm{x}) = R(\mathrm{x}_1, \ldots, \mathrm{x}_r) = \sum\limits_{i=1}^{r} R_i(\mathrm{x}_i)$, где $\mathrm{x} = (\mathrm{x}_1^\top, \ldots, \mathrm{x}_r^\top)^\top \in \mathbb{R}^n$ и $\mathrm{x}_i \in \mathbb{R}^{n_i}$ для $i = 1, \ldots, r$. Тогда

$$\mathrm{prox}_R(\mathrm{x}) = \begin{pmatrix} \mathrm{prox}_{R_1}(\mathrm{x}_1) \\ \vdots \\ \mathrm{prox}_{R_r}(\mathrm{x}_r) \end{pmatrix}.$$

Примеры вычисления проксимальных операторов

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

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

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

Примеры вычисления проксимальных операторов

Прокс-оператор сепарабельной функции

Доказательство. По определению мы имеем

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

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

Примеры вычисления проксимальных операторов

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

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

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

Примеры вычисления проксимальных операторов

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 $g(x) = \lambda|x|$, $x \in \mathbb{R}$:

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

- 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 $\mathrm{prox}_g(x) = x - \lambda$.
- Аналогичными рассуждениями получаем, что если $x < -\lambda$, то $\mathrm{prox}_g(x) = x + \lambda$.
- Во всех остальных случаях, т.е. при $x \in [-\lambda, \lambda]$, $\mathrm{prox}_g(x) = 0$.

Примеры вычисления проксимальных операторов

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 $g(x) = \lambda|x|$, $x \in \mathbb{R}$:

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

- 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 $\mathrm{prox}_g(x) = x - \lambda$.
- Аналогичными рассуждениями получаем, что если $x < -\lambda$, то $\mathrm{prox}_g(x) = x + \lambda$.
- Во всех остальных случаях, т.е. при $x \in [-\lambda, \lambda]$, $\mathrm{prox}_g(x) = 0$.

Примеры вычисления проксимальных операторов

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

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

- 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 $\mathrm{prox}_g(x) = x - \lambda$.
- Аналогичными рассуждениями получаем, что если $x < -\lambda$, то $\mathrm{prox}_g(x) = x + \lambda$.
- Во всех остальных случаях, т.е. при $x \in [-\lambda, \lambda]$, $\mathrm{prox}_g(x) = 0$.

Примеры вычисления проксимальных операторов

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 $g(x) = \lambda|x|$, $x \in \mathbb{R}$:

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

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

Примеры вычисления проксимальных операторов

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 $g(x) = \lambda|x|$, $x \in \mathbb{R}$:

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

- 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 $\mathrm{prox}_g(x) = x - \lambda$.
- Аналогичными рассуждениями получаем, что если $x < -\lambda$, то $\mathrm{prox}_g(x) = x + \lambda$.
- Во всех остальных случаях, т.е. при $x \in [-\lambda, \lambda]$, $\mathrm{prox}_g(x) = 0$.

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

где $1 \stackrel{\text{def}}{=} (1, \ldots, 1)^\top \in \mathbb{R}^n$,модуль $|x|$, срезка $[|x| - \lambda 1]_+$ и сигнум (знак) $\mathrm{sign}(x)$ 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).

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

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

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

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

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

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

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

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

Примеры вычисления проксимальных операторов

Полученный результат можно записать в следующем виде:

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

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

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

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

Проксимальный градиентный спуск

Для задачи (6) рассмотрим следующий метод.

Алгоритм 1 Проксимальный градиентный спуск

Вход: 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: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma \nabla f(x^k)\right)$
4: **end for**

Выход: 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 $x^* = \mathrm{argmin}_{x \in \mathbb{R}^n} F(x)$. Тогда $x^* = \mathrm{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$.

Проксимальный градиентный спуск

Для задачи (6) рассмотрим следующий метод.

Алгоритм 1 Проксимальный градиентный спуск

Вход: 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: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma \nabla f(x^k)\right)$
4: **end for**

Выход: 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 $x^* = \mathrm{argmin}_{x \in \mathbb{R}^n} F(x)$. Тогда $x^* = \mathrm{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$.

Проксимальный градиентный спуск

Для задачи (6) рассмотрим следующий метод.

Алгоритм 1 Проксимальный градиентный спуск

Вход: 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: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma \nabla f(x^k)\right)$

4: **end for**

Выход: 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 $x^* = \mathrm{argmin}_{x \in \mathbb{R}^n} F(x)$. Тогда $x^* = \mathrm{prox}_{\gamma R}(x^* - \gamma \nabla f(x^*))$.

Проксимальный градиентный спуск: сильно выпуклый случай

Теорема 2

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 1 удовлетворяет неравенству:

$$\|x^N - x^*\|_2^2 \leq (1 - \gamma\mu)^N \|x^0 - x^*\|_2^2. \tag{10}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{L}{\mu} \ln \frac{R_0^2}{\varepsilon}\right)$ итераций, где $R_0 = \|x^0 - x^*\|_2$, выполняется $\|x^N - x^*\|_2^2 \leq \varepsilon$.

Проксимальный градиентный спуск: сильно выпуклый случай

Теорема 2

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 1 удовлетворяет неравенству:

$$\|x^N - x^*\|_2^2 \leq (1-\gamma\mu)^N \|x^0 - x^*\|_2^2. \tag{10}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{L}{\mu}\ln\frac{R_0^2}{\varepsilon}\right)$ итераций, где $R_0 = \|x^0 - x^*\|_2$, выполняется $\|x^N - x^*\|_2^2 \leq \varepsilon$.

Проксимальный градиентный спуск: сильно выпуклый случай

Доказательство Теоремы 2. Пользуясь тем, что прокс-оператор является нерастягивающим (см. (9)), мы получаем

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|\mathrm{prox}_{\gamma R}(x^k - \gamma\nabla f(x^k)) - \mathrm{prox}_{\gamma R}(x^* - \gamma\nabla f(x^*))\|_2^2 \\
&\overset{(9)}{\leq} \|x^k - x^* - \gamma\left(\nabla f(x^k) - \nabla f(x^*)\right)\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\
&\quad +\gamma^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2.
\end{aligned}
$$

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

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

Проксимальный градиентный спуск: сильно выпуклый случай

Доказательство Теоремы 2. Пользуясь тем, что прокс-оператор является нерастягивающим (см. (9)), мы получаем

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|\mathrm{prox}_{\gamma R}(x^k - \gamma\nabla f(x^k)) - \mathrm{prox}_{\gamma R}(x^* - \gamma\nabla f(x^*))\|_2^2 \\
&\overset{(9)}{\leq} \|x^k - x^* - \gamma\left(\nabla f(x^k) - \nabla f(x^*)\right)\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\
&\quad +\gamma^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2.
\end{aligned}
$$

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

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

Проксимальный градиентный спуск: сильно выпуклый случай

Доказательство Теоремы 2. Пользуясь тем, что прокс-оператор является нерастягивающим (см. (9)), мы получаем

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(x^k - \gamma\nabla f(x^k)) - \operatorname{prox}_{\gamma R}(x^* - \gamma\nabla f(x^*))\|_2^2 \\ &\overset{(9)}{\leq} \|x^k - x^* - \gamma\left(\nabla f(x^k) - \nabla f(x^*)\right)\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\ &\quad +\gamma^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2.\end{aligned}$$

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

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

Проксимальный градиентный спуск: сильно выпуклый случай

Доказательство Теоремы 2. Пользуясь тем, что прокс-оператор является нерастягивающим (см. (9)), мы получаем

$$\begin{aligned}
\|x^{k+1}-x^*\|_2^2 &= \|\mathrm{prox}_{\gamma R}(x^k-\gamma\nabla f(x^k)) - \mathrm{prox}_{\gamma R}(x^*-\gamma\nabla f(x^*))\|_2^2 \\
&\overset{(9)}{\leq} \|x^k - x^* - \gamma\left(\nabla f(x^k) - \nabla f(x^*)\right)\|_2^2 \\
&= \|x^k-x^*\|_2^2 - 2\gamma\langle x^k-x^*, \nabla f(x^k)-\nabla f(x^*)\rangle \\
&\quad +\gamma^2\|\nabla f(x^k)-\nabla f(x^*)\|_2^2.
\end{aligned}$$

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

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

Проксимальный градиентный спуск: сильно выпуклый случай

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 $\|x^{k+1} - x^*\|_2^2$:

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

Проксимальный градиентный спуск: сильно выпуклый случай

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 $\|x^{k+1} - x^*\|_2^2$:

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

Проксимальный градиентный спуск: сильно выпуклый случай

Из полученных неравенств имеем:

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

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 $x, y \in \mathbb{R}^n$

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

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

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

Проксимальный градиентный спуск: сильно выпуклый случай

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 $2\gamma(1 - \gamma L) V_f(x^k, x^*) \geq 0$, 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.

Проксимальный градиентный спуск: сильно выпуклый случай

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 $2\gamma(1 - \gamma L) V_f(x^k, x^*) \geq 0$, 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.

Проксимальный градиентный спуск: сильно выпуклый случай

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 $2\gamma(1 - \gamma L) V_f(x^k, x^*) \geq 0$, 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.

Проксимальный градиентный спуск: сильно выпуклый случай

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 $2\gamma(1 - \gamma L)\, V_f(x^k, x^*) \geq 0$, 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.

Проксимальный градиентный спуск: выпуклый случай

Теорема 3

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 1 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \tag{11}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{LR_0^2}{\varepsilon}\right)$ итераций, где $R_0 = \|x^0 - x^*\|_2$, выполняется $F(x^N) - F(x^*) \leq \varepsilon$.

Проксимальный градиентный спуск: выпуклый случай

Теорема 3

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 1 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \tag{11}$$

Иными словами, для проксимального градиентного спуска с $\gamma = \frac{1}{L}$ через $N = O\left(\frac{LR_0^2}{\varepsilon}\right)$ итераций, где $R_0 = \|x^0 - x^*\|_2$, выполняется $F(x^N) - F(x^*) \leq \varepsilon$.

Проксимальный градиентный спуск: выпуклый случай

Доказательство Теоремы 3. Рассмотрим функцию

$$\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + R(x) + \frac{L}{2}\|x - x^k\|_2^2.$$

1. Функция $\varphi_k(x)$ является сильно выпуклой с константой L (первые три слагаемых – выпуклые функции, последнее слагаемое – L-сильно выпуклая функция).
2. $x^{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \varphi_k(x)$. Действительно, по определению

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

Проксимальный градиентный спуск: выпуклый случай

Доказательство Теоремы 3. Рассмотрим функцию

$$\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + R(x) + \frac{L}{2}\|x - x^k\|_2^2.$$

1. Функция $\varphi_k(x)$ является сильно выпуклой с константой L (первые три слагаемых – выпуклые функции, последнее слагаемое – L-сильно выпуклая функция).
2. $x^{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n} \varphi_k(x)$. Действительно, по определению

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

Проксимальный градиентный спуск: выпуклый случай

Доказательство Теоремы 3. Рассмотрим функцию

$$\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + R(x) + \frac{L}{2}\|x - x^k\|_2^2.$$

1. Функция $\varphi_k(x)$ является сильно выпуклой с константой L (первые три слагаемых – выпуклые функции, последнее слагаемое – L-сильно выпуклая функция).
2. $x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \varphi_k(x)$. Действительно, по определению

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

Проксимальный градиентный спуск: выпуклый случай

Доказательство Теоремы 3. Рассмотрим функцию

$$\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + R(x) + \frac{L}{2}\|x - x^k\|_2^2.$$

1. Функция $\varphi_k(x)$ является сильно выпуклой с константой L (первые три слагаемых – выпуклые функции, последнее слагаемое – L-сильно выпуклая функция).
2. $x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \varphi_k(x)$. Действительно, по определению

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

Проксимальный градиентный спуск: выпуклый случай

3. Из первых двух пунктов следует $\varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2$.

Небольшое наблюдение/пояснение

Покажем, что для функции $\varphi(x) = \psi(x) + R(x)$, где $\psi(x)$ — μ-сильно выпуклая дифференцируемая, $R(x)$ — выпуклая, но не обязательно гладкая функция, $x^* = \operatorname*{argmin}_{x\in\mathbb{R}^n} \varphi(x)$, выполняется $\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$ для всех $x \in \mathbb{R}^n$. Из условий оптимальности, имеем $-\nabla\psi(x^*) \in \partial R(x^*)$, а значит, $R(x) - R(x^*) \geq \langle -\nabla\psi(x^*), x - x^*\rangle$. Отсюда и из сильной выпуклости ψ получаем

$$
\begin{aligned}
\psi(x) - \psi(x^*) &\geq \langle \nabla\psi(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \\
&\geq R(x^*) - R(x) + \frac{\mu}{2}\|x - x^*\|_2^2,
\end{aligned}
$$

что и требовалось доказать.

Проксимальный градиентный спуск: выпуклый случай

3. Из первых двух пунктов следует $\varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2$.

Небольшое наблюдение/пояснение

Покажем, что для функции $\varphi(x) = \psi(x) + R(x)$, где $\psi(x)$ — μ-сильно выпуклая дифференцируемая, $R(x)$ — выпуклая, но не обязательно гладкая функция, $x^* = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\,\varphi(x)$, выполняется $\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$ для всех $x \in \mathbb{R}^n$. Из условий оптимальности, имеем $-\nabla\psi(x^*) \in \partial R(x^*)$, а значит, $R(x) - R(x^*) \geq \langle -\nabla\psi(x^*), x - x^*\rangle$. Отсюда и из сильной выпуклости ψ получаем

$$\begin{aligned}\psi(x) - \psi(x^*) &\geq \langle\nabla\psi(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|_2^2\\ &\geq R(x^*) - R(x) + \frac{\mu}{2}\|x - x^*\|_2^2,\end{aligned}$$

что и требовалось доказать.

Проксимальный градиентный спуск: выпуклый случай

3. Из первых двух пунктов следует $\varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2$.

Небольшое наблюдение/пояснение

Покажем, что для функции $\varphi(x) = \psi(x) + R(x)$, где $\psi(x)$ — μ-сильно выпуклая дифференцируемая, $R(x)$ — выпуклая, но не обязательно гладкая функция, $x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \varphi(x)$, выполняется $\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$ для всех $x \in \mathbb{R}^n$. Из условий оптимальности, имеем $-\nabla\psi(x^*) \in \partial R(x^*)$, а значит, $R(x) - R(x^*) \geq \langle -\nabla\psi(x^*), x - x^* \rangle$. Отсюда и из сильной выпуклости ψ получаем

$$\begin{aligned}\psi(x) - \psi(x^*) &\geq \langle \nabla\psi(x^*), x - x^* \rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \\ &\geq R(x^*) - R(x) + \frac{\mu}{2}\|x - x^*\|_2^2,\end{aligned}$$

что и требовалось доказать.

Проксимальный градиентный спуск: выпуклый случай

3. Из первых двух пунктов следует $\varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2$.

Небольшое наблюдение/пояснение

Покажем, что для функции $\varphi(x) = \psi(x) + R(x)$, где $\psi(x)$ — μ-сильно выпуклая дифференцируемая, $R(x)$ — выпуклая, но не обязательно гладкая функция, $x^* = \operatorname*{argmin}_{x \in \mathbb{R}^n} \varphi(x)$, выполняется $\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$ для всех $x \in \mathbb{R}^n$. Из условий оптимальности, имеем $-\nabla\psi(x^*) \in \partial R(x^*)$, а значит, $R(x) - R(x^*) \geq \langle -\nabla\psi(x^*), x - x^*\rangle$. Отсюда и из сильной выпуклости ψ получаем

$$
\begin{aligned}
\psi(x) - \psi(x^*) &\geq \langle \nabla\psi(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \\
&\geq R(x^*) - R(x) + \frac{\mu}{2}\|x - x^*\|_2^2,
\end{aligned}
$$

что и требовалось доказать.

Проксимальный градиентный спуск: выпуклый случай

3. Из первых двух пунктов следует $\varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2$.

Небольшое наблюдение/пояснение

Покажем, что для функции $\varphi(x) = \psi(x) + R(x)$, где $\psi(x)$ — μ-сильно выпуклая дифференцируемая, $R(x)$ — выпуклая, но не обязательно гладкая функция, $x^* = \operatorname*{argmin}_{x\in\mathbb{R}^n} \varphi(x)$, выполняется $\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$ для всех $x \in \mathbb{R}^n$. Из условий оптимальности, имеем $-\nabla\psi(x^*) \in \partial R(x^*)$, а значит, $R(x) - R(x^*) \geq \langle -\nabla\psi(x^*), x - x^*\rangle$. Отсюда и из сильной выпуклости ψ получаем

$$
\begin{aligned}
\psi(x) - \psi(x^*) &\geq \langle \nabla\psi(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \\
&\geq R(x^*) - R(x) + \frac{\mu}{2}\|x - x^*\|_2^2,
\end{aligned}
$$

что и требовалось доказать.

Проксимальный градиентный спуск: выпуклый случай

4. Кроме того, L-гладкость функции f влечёт

$$\begin{aligned}\varphi_k(x^{k+1}) &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + R(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\ &\geq F(x^{k+1}).\end{aligned}$$

5. Объединяя результаты пунктов 3 и 4, получим

$$\varphi_k(x) - F(x^{k+1}) \geq \varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2.$$

Проксимальный градиентный спуск: выпуклый случай

4. Кроме того, L-гладкость функции f влечёт

$$\begin{aligned}\varphi_k(x^{k+1}) &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + R(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\ &\geq F(x^{k+1}).\end{aligned}$$

5. Объединяя результаты пунктов 3 и 4, получим

$$\varphi_k(x) - F(x^{k+1}) \quad \geq \quad \varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2.$$

Проксимальный градиентный спуск: выпуклый случай

4. Кроме того, L-гладкость функции f влечёт

$$\begin{aligned}\varphi_k(x^{k+1}) &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + R(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\ &\geq F(x^{k+1}).\end{aligned}$$

5. Объединяя результаты пунктов 3 и 4, получим

$$\varphi_k(x) - F(x^{k+1}) \geq \varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2.$$

Проксимальный градиентный спуск: выпуклый случай

4. Кроме того, L-гладкость функции f влечёт

$$
\begin{aligned}
\varphi_k(x^{k+1}) &= f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + R(x^{k+1}) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\
&\geq F(x^{k+1}).
\end{aligned}
$$

5. Объединяя результаты пунктов 3 и 4, получим

$$
\varphi_k(x) - F(x^{k+1}) \quad \geq \quad \varphi_k(x) - \varphi_k(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2.
$$

Проксимальный градиентный спуск: выпуклый случай

6. Используя эквивалентную запись $\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + F(x) - f(x) + \frac{L}{2}\|x - x^k\|_2^2$, мы приходим к неравенству

$$F(x) - F(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2 - \frac{L}{2}\|x - x^k\|_2^2 + \underbrace{f(x) - f(x^k) - \langle \nabla f(x^k), x - x^k\rangle}_{V_f(x,x^k)\geq 0}$$

7. Если подставить $x = x^k$, то получим, что $F(x^k) - F(x^{k+1}) \geq \frac{L}{2}\|x^k - x^{k+1}\|_2^2 \geq 0$, т.е. метод монотонный (значение функции в генерируемых точках не увеличивается).

Проксимальный градиентный спуск: выпуклый случай

6. Используя эквивалентную запись $\varphi_k(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + F(x) - f(x) + \frac{L}{2}\|x - x^k\|_2^2$, мы приходим к неравенству

$$F(x) - F(x^{k+1}) \geq \frac{L}{2}\|x - x^{k+1}\|_2^2 - \frac{L}{2}\|x - x^k\|_2^2 + \underbrace{f(x) - f(x^k) - \langle \nabla f(x^k), x - x^k \rangle}_{V_f(x, x^k) \geq 0}$$

7. Если подставить $x = x^k$, то получим, что $F(x^k) - F(x^{k+1}) \geq \frac{L}{2}\|x^k - x^{k+1}\|_2^2 \geq 0$, т.е. метод монотонный (значение функции в генерируемых точках не увеличивается).

Проксимальный градиентный спуск: выпуклый случай

8. Если подставить $x = x^*$, то получим

$$F(x^*) - F(x^{k+1}) \geq \frac{L}{2}\|x^* - x^{k+1}\|_2^2 - \frac{L}{2}\|x^* - x^k\|_2^2.$$

Суммируя полученное неравенство для $k = 0, 1, \ldots, N-1$ и деля левую и правую части на N, получим

$$\begin{aligned} F(x^N) - F(x^*) &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(F(x^{k+1}) - F(x^*)\right) \\ &\leq \frac{L}{2N}\sum_{k=0}^{N-1}\left(\|x^* - x^k\|_2^2 - \|x^* - x^{k+1}\|_2^2\right) \\ &\leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \end{aligned}$$

Проксимальный градиентный спуск: выпуклый случай

8. Если подставить $x = x^*$, то получим

$$F(x^*) - F(x^{k+1}) \geq \frac{L}{2}\|x^* - x^{k+1}\|_2^2 - \frac{L}{2}\|x^* - x^k\|_2^2.$$

Суммируя полученное неравенство для $k = 0, 1, \ldots, N-1$ и деля левую и правую части на N, получим

$$\begin{aligned} F(x^N) - F(x^*) &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(F(x^{k+1}) - F(x^*)\right) \\ &\leq \frac{L}{2N}\sum_{k=0}^{N-1}\left(\|x^* - x^k\|_2^2 - \|x^* - x^{k+1}\|_2^2\right) \\ &\leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \end{aligned}$$

Проксимальный градиентный спуск: выпуклый случай

8. Если подставить $x = x^*$, то получим

$$F(x^*) - F(x^{k+1}) \geq \frac{L}{2}\|x^* - x^{k+1}\|_2^2 - \frac{L}{2}\|x^* - x^k\|_2^2.$$

Суммируя полученное неравенство для $k = 0, 1, \ldots, N-1$ и деля левую и правую части на N, получим

$$\begin{aligned} F(x^N) - F(x^*) &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(F(x^{k+1}) - F(x^*)\right) \\ &\leq \frac{L}{2N}\sum_{k=0}^{N-1}\left(\|x^* - x^k\|_2^2 - \|x^* - x^{k+1}\|_2^2\right) \\ &\leq \frac{L\|x^0 - x^*\|_2^2}{2N}. \end{aligned}$$

Проксимальный градиентный спуск: выпуклый случай

8. Если подставить $x = x^*$, то получим

$$F(x^*) - F(x^{k+1}) \geq \frac{L}{2}\|x^* - x^{k+1}\|_2^2 - \frac{L}{2}\|x^* - x^k\|_2^2.$$

Суммируя полученное неравенство для $k = 0, 1, \ldots, N-1$ и деля левую и правую части на N, получим

$$\begin{aligned}
F(x^N) - F(x^*) &\leq \frac{1}{N}\sum_{k=0}^{N-1}\left(F(x^{k+1}) - F(x^*)\right) \\
&\leq \frac{L}{2N}\sum_{k=0}^{N-1}\left(\|x^* - x^k\|_2^2 - \|x^* - x^{k+1}\|_2^2\right) \\
&\leq \frac{L\|x^0 - x^*\|_2^2}{2N}.
\end{aligned}$$

Проксимальный ускоренный градиентный метод (FISTA)

Пусть функция f выпукла (но не сильно выпукла). Тогда проксимальный градиентный метод можно ускорить следующим способом.

Алгоритм 2 Проксимальный ускоренный градиентный метод (FISTA) для выпуклых функций

Вход: 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: $x^{k+1} = \mathrm{prox}_{\frac{1}{L}R}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)$
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**

Выход: x^N

FISTA: выпуклый случай

Теорема 4 (без доказательства)

Пусть $f(x)$ — выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция. Тогда для любого $N > 0$ выход Алгоритма 2 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{2L\|x^0 - x^*\|_2^2}{(N+1)^2}. \tag{12}$$

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: выпуклый случай

Теорема 4 (без доказательства)

Пусть $f(x)$ — выпуклая L-гладкая функция, $R(x)$ — правильная выпуклая замкнутая функция. Тогда для любого $N > 0$ выход Алгоритма 2 удовлетворяет неравенству:

$$F(x^N) - F(x^*) \leq \frac{2L\|x^0 - x^*\|_2^2}{(N+1)^2}. \tag{12}$$

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.

Алгоритм 3 Проксимальный ускоренный градиентный метод (FISTA) для сильно выпуклых функций

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций N

1: $y^0 = x^0$, $\kappa = \frac{L}{\mu}$
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: Вычислить $\nabla f(y^k)$
4: $x^{k+1} = \mathrm{prox}_{\frac{1}{L}R}\left(y^k - \frac{1}{L}\nabla f(y^k)\right)$
5: $y^{k+1} = x^{k+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x^{k+1} - x^k)$
6: **end for**

Выход: x^N

FISTA: сильно выпуклый случай

Теорема 5 (без доказательства)

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

$$F(x^N) - F(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^N \left(F(x^0) - F(x^*) + \frac{\mu}{2}\|x^0 - x^*\|_2^2\right). \tag{13}$$

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 $F(x^N) - F(x^*) \leq \varepsilon$.

FISTA: сильно выпуклый случай

Теорема 5 (без доказательства)

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

$$F(x^N) - F(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^N \left(F(x^0) - F(x^*) + \frac{\mu}{2}\|x^0 - x^*\|_2^2\right). \tag{13}$$

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 $F(x^N) - F(x^*) \leq \varepsilon$.

Литература

- Нестеров Ю. Е. Методы выпуклой оптимизации. — М.: Издательство МЦНМО, 2010. — 281 с.
- Beck, Amir. First-order methods in optimization. Society for Industrial and Applied Mathematics, 2017.