Subgradient descent

Introduction

Рассматривается классическая задача выпуклой оптимизации:

$$\min_{x \in S} f(x),$$

Подразумевается, что $f(x)$ - выпуклая функция на выпуклом множестве S. Для начала будем рассматривать задачу безусловной минимизации (БМ), $S = \mathbb{R}^n$

Вектор g называется **субградиентом** функции $f(x) : S \to \mathbb{R}$ в точке x_0, если $\forall x \in S$:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

Градиентный спуск предполагает, что функция $f(x)$ является дифференцируемой в каждой точке задачи. Теперь же, мы будем предполагать лишь выпуклость.

Итак, мы имеем оракул первого порядка:

Вход: $x \in \mathbb{R}^n$

Выход: $\partial f(x)$ и $f(x)$

Algorithm

$$x_{k+1} = x_k - \alpha_k g_k, \tag{SD}$$

где g_k - произвольный субградиент функции $f(x)$ в т. x_k, $g_k \in \partial f(x_k)$

Bounds

Vanilla version

Запишем как близко мы подошли к оптимуму $x^* = \arg\min\limits_{x \in \mathbb{R}^n} f(x) = \arg f^*$ на последней итерации:

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 g_k^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \end{aligned}$$

Для субградиента: $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$. Из написанного выше:

$$2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_k - x^*\|^2 + \alpha_k^2 g_k^2 - \|x_{k+1} - x^*\|^2$$

Просуммируем полученное неравенство для $k = 0, \ldots, T-1$

$$
\begin{aligned}
\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle &= \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 g_k^2 \\
&\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 g_k^2 \\
&\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2
\end{aligned}
$$

Здесь мы предположили $R^2 = \|x_0 - x^*\|^2, \qquad \|g_k\| \leq G$.
Предполагая $\alpha_k = \alpha$ (постоянный шаг), имеем:

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T
$$

Минимизация правой части по α дает $\alpha^* = \dfrac{R}{G}\sqrt{\dfrac{1}{T}}$

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T} \tag{Subgradient Bound}
$$

Тогда (используя неравенство Йенсена и свойство субградиента $f(x^*) \geq f(x_k) + \langle g_k, x^* - x_k \rangle$) запишем оценку на т.н. `Regret`, а именно:

$$
\begin{aligned}
f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k \right) - f^* &\leq \frac{1}{T} \left(\sum_{k=0}^{T-1} f(x_k) - f^* \right) \\
&\leq \frac{1}{T} \left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\
&\leq GR \frac{1}{\sqrt{T}}
\end{aligned}
$$

Важные моменты:

- Получение оценок не для x_T, а для среднего арифметического по итерациям $\overline{x}$ - типичный трюк при получении оценок для методов, где есть выпуклость, но нет удобного убывания на каждой итерации. Нет гарантий успеха на каждой итерации, но есть гарантия успеха в среднем
- Для выбора оптимального шага необходимо знать (предположить) число итераций заранее. Возможный выход: инициализировать T небольшим значением, после достижения этого количества итераций удваивать T и рестартовать алгоритм. Более интеллектуальный способ: адаптивный выбор длины шага.

Steepest subgradient descent

Попробуем выбирать на каждой итерации длину шага более оптимально. Тогда:

$$
\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 + \alpha_k^2 g_k^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle
$$

Минимизируя выпуклую правую часть по α_k, получаем:

$$
\alpha_k = \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2}
$$

Оценки изменятся следующим образом:

$$
\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}
$$

$$
\langle g_k, x_k - x^* \rangle^2 = \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)\|g_k\|^2
$$

$$
\langle g_k, x_k - x^* \rangle^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)G^2
$$

$$
\sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle^2 \leq \sum_{k=0}^{T-1}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)G^2
$$

$$
\sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle^2 \leq \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right)G^2
$$

$$
\frac{1}{T}\left(\sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle\right)^2 \leq \sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle^2 \leq R^2G^2
$$

Значит,

$$
\sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}
$$

Что приводит к абсолютно такой же оценке $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ на невязку по значению функции. На самом деле, для такого класса функций нельзя получить результат лучше, чем $\frac{1}{\sqrt{T}}$ или $\frac{1}{\varepsilon^2}$ по итерациям

Online learning

Рассматривается следующая игра: есть игрок и природа. На каждом из $k = 0, \ldots, T-1$ шагов:

- *Игрок* выбирает действие x_k
- *Природа* (возможно, враждебно) выбирает выпуклую функцию f_k, сообщает игроку значение $f(x_k), g_k \in \partial f(x_k)$
- Игрок вычисляет следующее действие, чтобы минимизировать регрет:

$$
R_{T-1} = \sum_{k=0}^{T-1} f_k(x_k) - \min_x \sum_{k=0}^{T-1} f_k(x) \tag{Regret}
$$

В такой постановке цель игрока состоит в том, чтобы выбрать стратегию, которая минимизирует разницу его действия с наилучгим выбором на каждом шаге.

Несмотря на весьма сложную (на первый взгляд) постановку задачи, существует стратегия, при которой регрет растет как $\sqrt{T}$, что означает, что усредненный регрет $\frac{1}{T}R_{T-1}$ падает, как $\frac{1}{\sqrt{T}}$

Если мы возьмем оценку (Subgradient Bound) для субградиентного метода, полученную выше, мы имеем:

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq G \|x_0 - x^*\| \sqrt{T}
$$

Однако, в её выводе мы нигде не использовали тот факт, что $x^* = \arg\min_{x \in S} f(x)$. Более того, мы вообще не использовали никакой специфичности точки x^*. Тогда можно записать это для произвольной точки y:

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - y \rangle \leq G \|x_0 - y\| \sqrt{T}
$$

Запишем тогда оценки для регрета, взяв $y = \arg\min_{x \in S} \sum\limits_{k=0}^{T-1} f_k(x)$:

$$
\begin{aligned}
R_{T-1} = \sum_{k=0}^{T-1} f_k(x_k) - \min_x \sum_{k=0}^{T-1} f_k(x) = \sum_{k=0}^{T-1} f_k(x_k) - \sum_{k=0}^{T-1} f_k(y) = \\
= \sum_{k=0}^{T-1} \left(f_k(x_k) - f_k(y) \right) \leq \sum_{k=0}^{T-1} \langle g_k, x_k - y \rangle \leq \\
\leq G \|x_0 - y\| \sqrt{T}
\end{aligned}
$$

Итого мы имеем для нашей стратегии с постоянным шагом:

$$
\overline{R_{T-1}} = \frac{1}{T} R_{T-1} \leq G \|x_0 - x^*\| \frac{1}{\sqrt{T}}, \qquad \alpha_k = \alpha = \frac{\|x_0 - x^*\|}{G} \sqrt{\frac{1}{T}}
$$

Examples

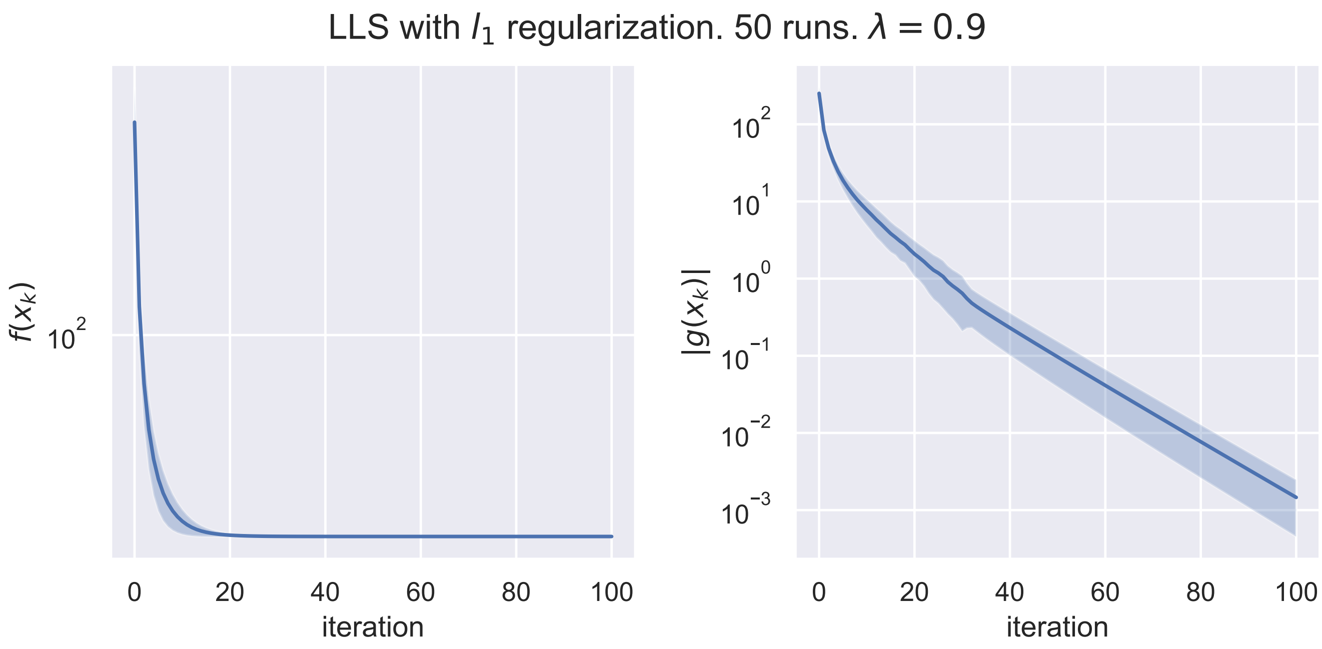
Least squares with l_1 regularization

$$
\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1
$$

Algorithm will be written as:

$$
x_{k+1} = x_k - \alpha_k \left(A^\top (Ax_k - b) + \lambda \text{sign}(x_k) \right)
$$

where signum function is taken element-wise.



Support vector machines

Let $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$

We need to find $\omega \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$
\min_{\omega \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^m max[0, 1 - y_i(\omega^\top x_i + b)]
$$

Code

- Open in Colab - Wolfe's example and why we usually have oscillations in non-smooth optimization.
- Open in Colab - Linear least squares with l_1- regularization.

References

- Great cheatsheet by Sebastian Pokutta
- Lecture on subgradient methods @ Berkley