сходимость GD
* выпуклой гладк.
* сильновып.
* PL // ?

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

Introduction

Рассматривается классическая задача выпуклой оптимизации:

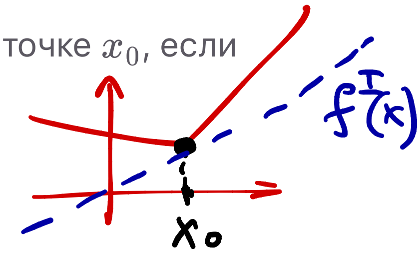
$f(x) = |x|$

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

$\nabla f(x^k) \to g^k$

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 $\forall x \in S$:



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

$$f(x) \geq \overbrace{f(x_0) + \langle g, x - x_0 \rangle}^{f^T(x)}$$

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 $f(x)$

GD

α = constant выпуклый

steepest descent

T итераций

$O\left(\frac{1}{T}\right)$

ускоренный GD → можно достичь $O\left(\frac{1}{T^2}\right)$

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

$\min\limits_{x \in \mathbb{R}^n} f(x)$

оптимальное решение

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:

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

$g_k^2 = g_k^T g_k = \|g_k\|^2$

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

субград: $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*)$

$f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

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

$\langle g^k, x^* - x^k \rangle \leq f(x^*) - f$

$\langle g^k, x^k - x^* \rangle \geq f^k - f^*$

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

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

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

$\boxed{g_k^2 \leq G^2}$

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

$\alpha = \text{const}$

$\alpha = \arg\min(\ldots)$

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

$\frac{R^2}{2} \frac{G}{R}\sqrt{T} + \frac{R}{G}\frac{GT}{2\sqrt{T}} = GR\sqrt{T}$

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

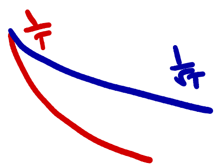
$\frac{1}{\alpha^*} = \frac{G}{R}\sqrt{T}$

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

$f(x_k) - f^* \leq \langle g_k, x_k - x^* \rangle$

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

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



Примечание: для негладких задач больше нет монотонного убывания $f(x)$ на кажд итерации!

вместо $f(x^T) - f^* \leq \varepsilon$ => $f(\overline{x}^T) - f^* \leq \varepsilon$

$\overline{x}^T = \frac{1}{T}\sum_{i=0}^{T-1} x_i$

Важные моменты:

$\boxed{f(\overline{x}^T) - f^* \leq \frac{GR}{\sqrt{T}}} = \varepsilon$

$O\left(\frac{1}{\sqrt{T}}\right)$

$\boxed{T \sim \frac{1}{\varepsilon^2}}$

- 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

$f(\alpha) : \mathbb{R} \to \mathbb{R}$

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

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

Минимизируя выпуклую правую часть по α_k, получаем:

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

Оценки изменятся следующим образом:

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

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

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

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

ЗАБИВАЕМ

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

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

Значит,

$$\boxed{\frac{1}{T}\left(\sum x_i\right)^2 \leq \sum_i x_i^2}$$

$(a+b)^2 \leq 2(a^2+b^2)$

$a^2+b^2+2ab \leq 2a^2+2b^2$

$2ab \leq a^2+b^2$

$(a-b)^2 \geq 0$

$(x_1 + \dots x_T)^2 \leq T(x_1^2 + \dots + x_T^2)$

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

$$\boxed{f(\overline{x}^T) - f^* \leq \frac{GR}{\sqrt{T}}}$$

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 $y = \arg\min_{x \in S} \sum_{k=0}^{T-1} f_k(x)$:

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

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

Пример выпуклой негладкой функции:

Least squares with l_1 regularization

$\nabla f(x) = ?$

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

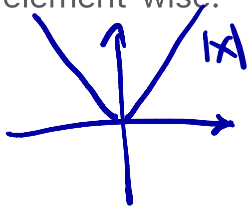
Algorithm will be written as:

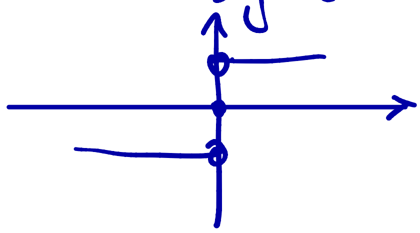
в Jax numpy

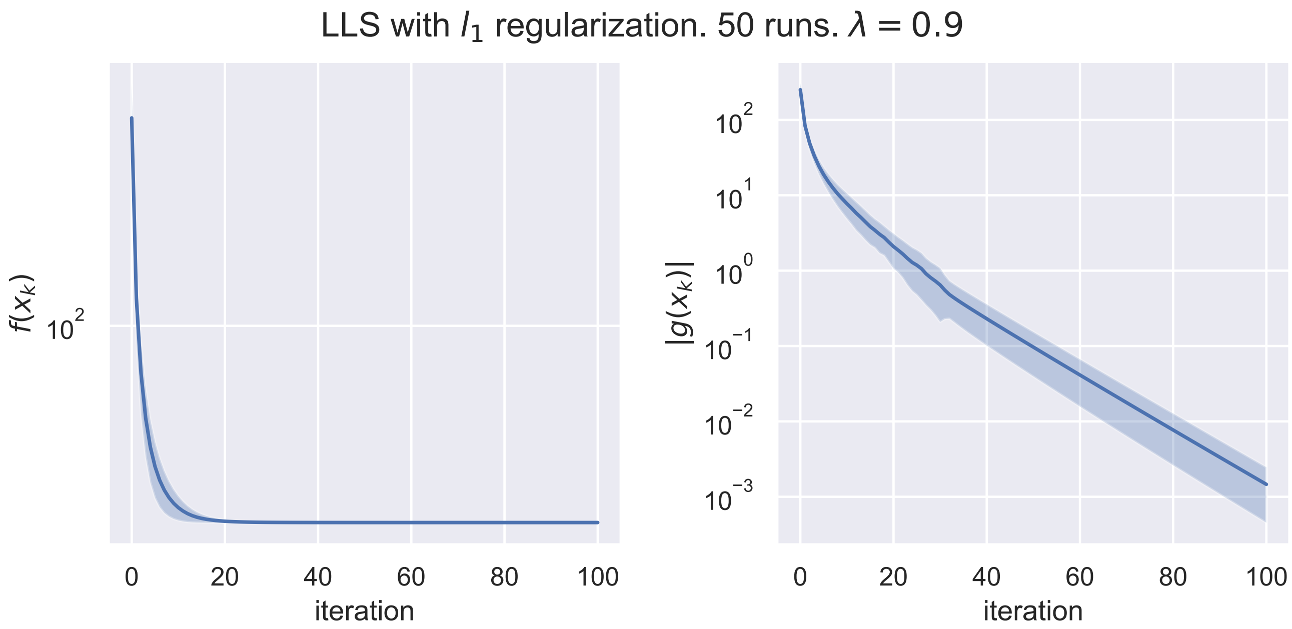
$$\boxed{x_{k+1} = x_k - \alpha_k \left(A^\top(Ax_k - b) + \lambda \text{sign}(x_k)\right)}$$

where signum function is taken element-wise.

$$\text{sign}(x) = \begin{cases} -1, & x<0 \\ 0, & x=0 \\ 1, & x>0 \end{cases}$$







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

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

Bounds

Conditions	$f(\bar{x}) - f(x^*) \leq$	Type of convergence	$\|x_k - x^*\| \leq$
Convex Lipschitz-continuous function(G)	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $\frac{GR}{\sqrt{k}}$	Sublinear	
Convex Lipschitz-continuous gradient (L)	$\mathcal{O}\left(\frac{1}{k}\right)$ $\frac{LR^2}{k}$	Sublinear	
μ-Strongly convex Lipschitz-continuous gradient(L)		Linear	$(1 - \eta\mu)^k R^2$

μ-Strongly convex Lipschitz-continuous hessian(M)		Locally linear $R < \overline{R}$	$\frac{\overline{R}R}{\overline{R}-R}\left(1 - \frac{2\mu}{L+3\mu}\right)$

- $R = \|x_0 - x^*\|$ - initial distance
- $\overline{R} = \dfrac{2\mu}{M}$
- $\overline{x} = \dfrac{1}{k}\sum_{i=1}^k x_i$
- $\|g_k\| \leq G$

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
- Illustration of l1 regularization