# 随机优化算法介绍

胡涛
北京大学信息科学技术学院
taohu@pku.edu.cn

## 1 概览

首先介绍Hoeffding Inequality:

martigale的定义可以参照`https://en.wikipedia.org/wiki/Martingale_(probability_theory)`

martingale difference sequence (MDS)的定义可以参照`https://en.wikipedia.org/wiki/Martingale_difference_sequence`

martigale和martingale difference sequence之间有一些联系。

总体需要优化的问题如下:

$$min_{x\in R^n} f(x) \text{ 其中} f_x \text{为需要优化的函数}$$

下面主要会介绍以下几种随机优化算法:

- 次梯度法
- 梯度法
- SVG方法及其变种
- 随机优化算法在深度学习中的应用

## 2 次梯度法

次梯度法的流程如下:

$$x_{k+1} = x_k - \alpha_k g_k, g_k \in \partial f(x_k) \tag{2.1}$$

上述的式子等价于下面的式子:

$$x_{k+1} = argmin_x f(x_k) + < g_k, x - x_k > + \frac{1}{2\alpha_k}\|x - x_k\|_2^2 \tag{2.2}$$

公式2.1的具体推导如下:

泰勒公式二阶展开

$$f(x) \approx f(x_k) + < g_k, x - x_k > + \frac{1}{2\alpha_k}||x - x_k||_2^2$$

则$$x_{k+1} = argmin_x f(x_k) + < g_k, x - x_k > + \frac{1}{2\alpha_k}||x - x_k||_2^2$$

$$x_{k+1} = argmin_x < g_k, x - x_k > + \frac{1}{2\alpha_k}||x - x_k||_2^2$$

$$x_{k+1} = argmin_x 2\alpha_k < g_k, x - x_k > + ||x - x_k||_2^2$$

化简得到:$$x_{k+1} = argmin_x < x, x > + < 2\alpha_k g_k - 2x_k, x >$$

上述问题有显式解:

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

次梯度法的公式很简单，那么次梯度法的收敛性如何呢？下面给予证明。
首先证明一个引理:

**Theorem 1: Convergence of subgradient**

Let $\alpha_k \geq 0$ be any non-negative sequence of stepsizes and the preceding assumptions hold. Let $x_k$ be generated by the subgradient iteration. Then for all $K \geq 1$,

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^*)] \leq \frac{1}{2}||x_1 - x^*||_2^2 + \frac{1}{2}\sum_{k=1}^{K} \alpha_k^2 M^2.$$

值得注意的是上面的引理有两个假设:

- 最优解至少是bounded, 即存在$x^* \in argmin_x f(x)$并且$f(x^*) > -\infty$
- 所有的次梯度都是bounded，即$||g||_2 \leq M \leq \infty$对所有的x和$g \in \partial f(x)$都成立

下面给出具体证明:
由于f(x)为凸函数，所以有:

$$< g_k, x^* - x_k > \leq f(x^*) - f(x_k) \tag{2.3}$$

$$\frac{1}{2}||x_{k+1} - x^*||_2^2 = \frac{1}{2}||x_k - \alpha_k g_k - x^*||_2^2$$

拼凑,$$= \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k < g_k, x^* - x_k > + \partial\alpha_k^2 2||g_k||_2^2$$

利用凸函数性质(2.3),$$\leq \frac{1}{2}||x_k - x^*||_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}M^2$$

利用归纳法,即得证.

引理证明完以后，下面接着证明次梯度法的收敛性.首先令$\bar{x}_k = \frac{\sum_{k=1}^K \alpha_k x_k}{\sum_{k=1}^K \alpha_k}$.结合上面的引理很显然可以推导出:

$$f(\bar{x}_k) - f(x^*) \leq \frac{\sum_{k=1}^K \alpha_k x_k + \sum_{k=1}^K \alpha_k^2 M^2}{2\sum_{k=1}^K \alpha_k}$$

可以得到以下几个结论:

- 根据实际应用中我们对步长的设置,$\sum_{k=1}^{\infty} \alpha_k = \infty$,并且$\frac{\sum_{k=1}^K \alpha_k^2 M^2}{2\sum_{k=1}^K \alpha_k} \to 0$,得知随着K增大,式子左边会趋近于0.

- 假设我们使用固定步长,$\alpha_k = \alpha, \|x_1 - x^*\| \leq R$,那么可以得到:

$$f(\bar{x_k}) - f(x^*) \leq \tfrac{R^2}{2K\alpha} + \tfrac{\alpha M^2}{2}$$

- 如果使用固定步长，上面的式子就不会趋近于0了,因为有$\frac{\alpha M^2}{2}$这一项。我们可以通过令步长$\alpha_k = \frac{R}{M\sqrt{k}}$,这样式子$\frac{\alpha M^2}{2}$就会趋近于0.

那么为什么$f(\bar{x_k}) - f(x^*)$趋近于0，次梯度法就收敛呢?

# 3 梯度法

梯度法的流程如下:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

上述的式子等价于下面的式子:

$$x_{k+1} = argmin_x f(x_k) + < \nabla f(x_k), x - x_k > + \frac{1}{2\alpha_k}\|x - x_k\|_2^2 \tag{2.4}$$

2.4式的具体推导可以参照次梯度法中的推导。

# 4 SVG方法及其变种

# 5 随机优化算法在深度学习中的应用

# 6 总结

# 7 附录(一些额外基础知识)

一些基本性质：

- convex function

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \forall \lambda \in [0,1], x, y$$

对于凸函数有以下性质:

$$f(y) \geq f(x) + \nabla f(x)(y - x)$$

将f(y)在x处二阶展开,可以得到如下结果:

$$f(y) = f(x) + \nabla f(x)(y - x) + \tfrac{\nabla f(x)^2}{2\beta^2}$$

很显然有:

$$f(y) \geq f(x) + \nabla f(x)(y - x)$$

- M-Lipschitz function

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

M-Lipschitz function有如下性质:

$$\|\nabla f(x)\|_2 \leq M$$

- L-smooth function

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|_2$$

L-smooth function有以下性质:
(1).$\frac{L}{2}x^T x - f(x)$为凸函数.
(2).$f(y) \leq f(x) + < \nabla f(x), y - x > + \frac{L}{2}\|x - y\|_2^2$
$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2$

- $\mu$-strongly convex function

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2}\lambda(1 - \lambda)\|x - y\|_2^2, \forall \lambda \in [0, 1], x, y$$

$\mu$-strongly convex function有以下性质:
(1).$f(x) - \frac{L}{2}x^T x$为凸函数.
(2).$f(y) \geq f(x) + < \nabla f(x), y - x > + \frac{\mu}{2}\|x - y\|_2^2$

Co-coercivity of gradient:

## Co-coercivity of gradient

if $f$ is convex with $\mathbf{dom}\, f = \mathbf{R}^n$ and $(L/2)x^\top x - f(x)$ is convex then

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y$$

*proof*: define convex functions $f_x, f_y$ with domain $\mathbf{R}^n$:

$$f_x(z) = f(z) - \nabla f(x)^\top z, \quad f_y(z) = f(z) - \nabla f(y)^\top z$$

the functions $(L/2)z^\top z - f_x(z)$ and $(L/2)z^\top z - f_y(z)$ are convex

- $z = x$ minimizes $f_x(z)$; from the left-hand inequality,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L}\|\nabla f_x(y)\|_2^2 = \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2 \end{aligned}$$

- similarly, $z = y$ minimizes $f_y(z)$; therefore

$$f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

  combining the two inequalities shows co-coercivity