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

备注：此笔记由北京大学文再文老师《大数据分析中的算法》课堂讲义整理而成

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
- Variance Reduction
- 随机优化算法在深度学习中的应用

2 次梯度法

2.1 次梯度法(Subgradient methods)

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:

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

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

- 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?

2.2 投影次梯度法(Projected Subgradient methods)

考虑如下问题:

$$min_{x \in C} f(x)$$

投影次梯度法步骤如下:

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

其中的投影操作为:

$$\pi_C = argmin_{y \in C} \|y - x\|_2^2$$

投影次梯度法也可以写成:

$$x_{k+1} = argmin_{y \in C} f(x_k) + < g_k, x - x_k > + \frac{1}{2\alpha}\|x - x_k\|_2^2$$

投影次梯度法的证明如下所示：
同样这里先证明一个引理:

Theorem 2: Convergence of projected subgradient method

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

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^*)] \leq \frac{1}{2}||x_1 - x^*||_2^2 + \frac{1}{2}\sum_{k=1}^{K} \alpha_k^2 M^2. \tag{20}$$

值得注意的是上面的引理同样有两个假设:

- 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

首先利用到了convex set上projection 的non-expansiveness(non-expansiveness的证明见附录):

$$if \pi_C = argmin_{y \in C} \|y - x\|_2^2$$
$$then, \|y_1 - y_2\| \geq \|x_1 - x_2\|$$

利用π_C的non-expansiveness，可以得到:

$$\|x_{k+1} - x^*\|_2^2 = \| \pi_C(x_k - \alpha g_k) - x^*\|_2^2 = \| \pi_C(x_k - \alpha g_k) - \pi_C(x^*)\|_2^2 \leq \|x_k - \alpha g_k - x^*\|_2^2$$

接下来的证明就和次梯度法类似了:

$$\frac{1}{2}\|x_{k+1} - x^*\|_2^2 \leq \frac{1}{2}\|x_k - \alpha_k g_k - x^*\|_2^2$$
$$\text{拼凑}, = \frac{1}{2}\|x_k - x^*\|_2^2 - \alpha_k < g_k, x^* - x_k > + \partial \alpha_k^2 2\|g_k\|_2^2$$
$$\text{利用凸函数性质(2.3)}, \leq \frac{1}{2}\|x_k - x^*\|_2^2 - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2}M^2$$
$$\text{利用归纳法,即得证.}$$

接下来的收敛性证明和次梯度法类似，这里就不详细介绍了。

2.3 随机次梯度法(Stochasticc Subgradient Methods)

2.4 Azuma-Hoeffding Inequality

2.5 Adaptive stepsizes and metrics

选择一个合适的度量方式(metrics)，或者一个更好的步长方案，往往能够实现更快的收敛。因此在梯度下降方法上一般都从以上两个方面来改进。

一个简单的方案是:

$$h(x) = \frac{1}{x} x^T A x$$

其中A和数据项有关。

2.5.1 Adaptive stepsize

回顾上面我们介绍的随机次梯度法的bound:

$$E[f(\bar{x_k}) - f(x^*)] \leq E[\frac{R^2}{K\alpha_k} + \frac{1}{2K}\sum_{k=1}^K \alpha_k \|g_k\|_*^2]$$

很显然当k趋近于无穷大时,$\frac{1}{2K}\sum_{k=1}^K \alpha_k \|g_k\|_*^2$的收敛性无法保证，这里我们构造了一种步长规则，使得$\frac{1}{2K}\sum_{k=1}^K \alpha_k \|g_k\|_*^2$收敛。

令$\alpha_k = \frac{R}{\sqrt{\sum_{k=1}^K \|g_k\|_*^2}}$，则可以得到:

$$E[f(\bar{x_k}) - f(x^*)] \leq \frac{2R}{K} E[(\sum_{k=1}^K \|g_k\|_*^2)^{\frac{1}{2}}]$$

假设$E[\|g_k^2\|_*] \leq M^2, \forall k$,上式可以进一步化简:

$$E[(\sum_{k=1}^K \|g_k\|_*^2)^{\frac{1}{2}}] \leq \sqrt{M^2 K} \leq M\sqrt{K}$$

可以看到，随着K增大,整个式子是收敛的.

2.5.2 Variable metric methods

Variable metric methods本质上就是选择不同的metric，即H_k矩阵的构造。

再来回顾一下原问题:

$$x_{k+1} = argmin_{x\in C} x_k + < g_k, x - x_k > + \frac{1}{2} < x - x_k, H_k(x - x_k) >$$

去掉无关变量:

$$x_{k+1} = argmin_{x\in C} < g_k, x > + \frac{1}{2} < x - x_k, H_k(x - x_k) >$$

这就是梯度法的基础模型,下面举出几种常见的H_k的取法。

- 投影次梯度法.

$$H_k = \alpha_k I$$

- 牛顿法

$$H_k = \nabla^2 f(x_k)$$

- AdaGrad

$$H_k = \frac{1}{\alpha} diag(\sum_{k=1}^{K} g_k . * g_k)^{\frac{1}{2}}$$

其中,点乘表示elementwise mulitplication。diag表示将矢量按照对角线扩充为矩阵。

2.5.3 Variable metric methods收敛性

Theorem 9: Convergence of Variable metric methods

Let $H_k > 0$ be a sequence of positive define matrices, where H_k is a function of $g_1, \cdots, g_k$. Let g_k be stochastic subgradient with $\mathbb{E}[g_k|x_k] \in \partial f(x_k)$. Then

$$\mathbb{E}[\sum_{k=1}^{K}(f(x_k) - f(x^*))] \leq \frac{1}{2}\mathbb{E}[\sum_{k=2}^{K}(\|x_k - x^*\|_{H_k}^2 - \|x_k - x^*\|_{H_{k-1}}^2)] + \frac{1}{2}\mathbb{E}[\|x_1 - x^*\|_{H_1}^2 + \sum_{k=1}^{K}\|g_k\|_{H_k^{-1}}^2].$$

2.5.4 Optimality Guarantees

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 Variance Reduction

首先指出几个assumption,只有在这些假设下，下面的结论才会成立，当然这些假设都是很合理的:

- f(x) is L-smooth
- f(x) is μ-strongly convex
- $E_s[\nabla f_s(x)] = \nabla f(x)$
  其中s代表随机采样,本条件的含义是随机采样的梯度的期望和不采用随机采样的梯度一样。
- $E_s \|\nabla f_s(x)\|^2 \leq M^2$
- GD有线性(linear)收敛速度$o(n.log(\frac{1}{\epsilon}))$
- GD有次线性(sublinear)收敛速度$o(\frac{1}{\epsilon})$

4.1 回顾GD,SGD

Gradient Descend:

$$\begin{aligned} \triangle_{k+1}^2 &= \|x_{k+1} - x^*\|_2^2 = \|x_k - \alpha \nabla f(x_k) - x^*\|_2^2 \\ &= \triangle_k^2 - 2\alpha < \nabla f(x_k), x_k - x^* > + \alpha^2 \|\nabla f(x_k)\|_2^2 \\ &\leq (1 - 2\alpha\mu)\ \triangle_k^2 + \alpha^2 \|\nabla f(x_k)\|_2^2 (\mu - strongly convex) \\ &\quad TODO \\ &\leq (1 - 2\alpha\mu + \alpha^2 L^2)\ \triangle_k^2\ (L - smooth) \end{aligned}$$

Stochastic Gradient Descend:

$$\begin{aligned} E\ \triangle_{k+1}^2 &= E[\|x_k - \alpha \nabla f_{s_k}(x_k) - x^*\|_2^2] \\ &= E\ \triangle_k^2 - 2\alpha E < \nabla f_{s_k}(x_k), x_k - x^* > + \alpha^2 E \|\nabla f_{s_k}(x_k)\|_2^2 \\ &= E\ \triangle_k^2 - 2\alpha E < \nabla f(x_k), x_k - x^* > + \alpha^2 E \|\nabla f_{s_k}(x_k)\|_2^2 \end{aligned}$$

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:

$$f(y) = f(x) + \nabla f(x)(y-x) + \frac{\nabla f(x)^2}{2\rho^2}$$

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).$f(y) \leq f(x) + < \nabla f(x), y-x > + \frac{L}{2}\|x-y\|_2^2$
$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x-x^*\|_2^2$

- μ-strongly convex function

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

μ-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

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,

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

- 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

7.1 Co-coercivity of gradient

7.2 Projection Operator is non-expansive

$$if \pi_C = argmin_{y \in C}\|y - x\|_2^2,$$
$$then, \|y_1 - y_2\| \geq \|x_1 - x_2\|$$

这里的集合C必须是非空凸集。

可以参考: https://math.stackexchange.com/questions/1426343/prove-that-projection-operator-is-non-expansive

下面给出具体证明:

首先需要知道projection operator的variational characterization(又叫做Bourbaki-Cheney-Goldstein inequality,具体可以参考Proposition 1.1.9 in the book Convex Optimization Theory by Dimitri Bertsekas) :

$$if \pi_C = argmin_{y \in C}\|y - x\|_2^2,$$
$$then, < x_1 - y_1, x - y_1 > \leq 0$$

下面利用projection operator的variational characterization来证明non-expansive:

由于

$$< x_1 - y_1, x - y_1 > \leq 0, \forall x$$

所以可得:

$$< x_1 - y_1, y_2 - y_1 >\leq 0$$

同理可得:

$$< x_2 - y_2, y_1 - y_2 >\leq 0$$

将上面两个不等式相加，整理，并且运用柯西不等式:

$$\begin{gathered} < x_1 - x_2, y_2 - y_1 > + < y_2 - y_1, y_2 - y_1 >\leq 0 \\ then, < y_2 - y_1, y_2 - y_1 >\leq < x_2 - x_1, y_2 - y_1 > \\ then, < y_2 - y_1, y_2 - y_1 >\leq < x_2 - x_1, y_2 - y_1 >\leq \|x_2 - x_1\|\|y_2 - y_1\| \\ then, \|y_2 - y_1\|^2 \leq \|x_2 - x_1\|\|y_2 - y_1\| \end{gathered}$$

所以有:

$$\|y_2 - y_1\| \leq \|x_2 - x_1\|$$