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

http://bicmr.pku.edu.cn/~wenzw/bigdata/lect-sto.pdf

Contents

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:

$$\begin{gathered}
\text{泰勒公式二阶展开} \\
f(x) \approx f(x_k) + < g_k, x - x_k > + \tfrac{1}{2\alpha_k} \|x - x_k\|_2^2 \\
\text{则} x_{k+1} = argmin_x f(x_k) + < g_k, x - x_k > + \tfrac{1}{2\alpha_k} \|x - x_k\|_2^2 \\
x_{k+1} = argmin_x < g_k, x - x_k > + \tfrac{1}{2\alpha_k} \|x - x_k\|_2^2 \\
x_{k+1} = argmin_x 2\alpha_k < g_k, x - x_k > + \|x - x_k\|_2^2 \\
\text{化简得到:} x_{k+1} = argmin_x < x, x > + < 2\alpha_k g_k - 2x_k, x > \\
\text{上述问题有显式解:} \\
x_{k+1} = x_k - \alpha_k g_k, \text{得证}
\end{gathered}$$

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 $K \geq 1$,

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

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:
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$\frac{\sum_{k=1}^{K}\alpha_k^2 M^2}{2\sum_{k=1}^{K}\alpha_k} \rightarrow 0$,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):

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

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:

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

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。

3.1 梯度法(GD)的收敛性

首先假设函数f是满足L-smooth和μ-strongly convex条件的。

Assumption

- f is L-smooth and μ-strongly convex.

lemma: Coercivity of gradients

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{L\mu}{L+\mu}\|x-y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 \quad (42)$$

Theorem: Convergence rates of GD

Let $\alpha_k = \frac{2}{L+\mu}$ and let $\kappa = \frac{L}{\mu}$. Define $\Delta_k = \|x_k - x^*\|$. Then we get,

$$f(x_{T+1}) - f(x^*) \leq \frac{L\Delta_1^2}{2}\exp(-\frac{4T}{\kappa+1}). \quad (43)$$

接下来证明收敛性，这里会用到coercivity of gradients的性质，大致的思路是先证明$f(x_{T+1}) - f(x^*)$与$f(x_{T+1}) - f(x^*)$之间的递推关系，然后利用$\alpha_k = \frac{2}{L+\mu}$的特殊性得到$f(x_{T+1}) - f(x^*) \leq \frac{L\nabla_1^2}{2}exp(-\frac{4T}{\kappa+1})$

Proof of Theorem

-

$$
\begin{aligned}
\Delta_{k+1}^2 &= ||x_{k+1} - x^*||_2^2 = ||x_k - \alpha_k \nabla f(x_k) - x^*||_2^2 \\
&= ||x_k - x^*||_2^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle + \alpha_k^2 ||\nabla f(x_k)||_2^2 \\
&= \Delta_k^2 - 2\alpha_k \boxed{\langle \nabla f(x_k), x_k - x^* \rangle} + \alpha_k^2 ||\nabla f(x_k)||_2^2
\end{aligned}
$$

- By the lemma

$$
\begin{aligned}
\Delta_{k+1}^2 &\leq \Delta_k^2 - 2\alpha_k \boxed{(\frac{L\mu}{L+\mu}\Delta_k^2 + \frac{1}{L+\mu}\|\nabla f(x)\|^2)} + \alpha_k^2 ||\nabla f(x_k)||_2^2 \\
&= (1 - 2\alpha_k \frac{L\mu}{L+\mu})\Delta_k^2 + (-\frac{2\alpha_k}{L+\mu} + \alpha_k^2)||\nabla f(x_k)||_2^2 \\
&\leq (1 - 2\alpha_k \frac{L\mu}{L+\mu})\Delta_k^2 + (-\frac{2\alpha_k}{L+\mu} + \alpha_k^2) L^2 \Delta_k^2 \qquad (44)
\end{aligned}
$$

Proof of Theorem

- $\alpha_k = \frac{2}{L+\mu}$

$$
\begin{aligned}
\Delta_{k+1}^2 &\leq (1 - \frac{4L\mu}{(L+\mu)^2})\Delta_k^2 \\
&= (\frac{L-\mu}{L+\mu})^2 \Delta_k^2 = (\frac{\kappa-1}{\kappa+1})^2 \Delta_k^2
\end{aligned}
$$

-

$$
\begin{aligned}
\Delta_{T+1}^2 &\leq (\frac{\kappa-1}{\kappa+1})^{2T} \Delta_1^2 \\
&= \Delta_1^2 \exp(2T \log(1 - \frac{2}{\kappa+1})) \\
&\leq \Delta_1^2 \exp(-\frac{4T}{\kappa+1})
\end{aligned}
$$

-

$$
f(x_{T+1}) - f(x^*) \leq \frac{L}{2}\Delta_{T+1}^2 \leq \frac{L\Delta_1^2}{2} \exp(-\frac{4T}{\kappa+1})
$$

3.1.1 梯度法(GD)的另一种收敛性证明

首先考虑固定步长的情况:

附录中介绍了quadratic upper bound:

$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2$

在这里令$y = x - t\nabla f(x)$得到:

$$f(x - t\nabla f(x)) \le f(x) - t(1 - \frac{Lt}{2})\|\nabla f(x)\|_2^2$$

x^+表示下一次迭代的x，这里假设步长满足如下：

$$x^+ = x - t\nabla f(x), 0 < t < \frac{1}{L}$$

所以有

$f(x^+) \le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2$????
$\le f(x^*) + \nabla f(x)^T (x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2$(凸函数性质)
$= f(x^*) + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - x^* - t\nabla f(x)\|_2^2)$ (整理)
$= f(x^*) + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$

所以有:

$$f(x^+) - f(x^*) = \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$$

令$x = x^{(i=1)}, x^+ = x^{(i)}$,则
$\sum_{i=1}^{k}(f(x^{(i)}) - f(x^*)) \le \frac{1}{2t}\sum_{i=1}^{k}(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2) \le \frac{1}{2t}(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2) \le \frac{1}{2t}\|x^{(0)} - x^*\|_2^2$
因为$f(x^{(i)})$是递减的,那么

$$f(x^{(k)}) - f(x^*) \le \frac{1}{k}\sum_{i=1}^{k}(f(x^i) - f(x^*)) \le \frac{1}{2kt}\|x^{(0)} - x^*\|_2^2$$

其中k在分母中，因此梯度下降法达到$f(x^{(k)}) - f(x^*) \le \epsilon$的精度需要$o(\frac{1}{\epsilon})$的时间复杂度。

3.1.2 backtracking line search

line search方法有很多，这里只介绍backtracking line search:

initialize t_k at $\hat{t}$(for example $\hat{t} = 1$),take $t_k = \beta t_k$ until :
$f(x - t_k\nabla f(x)) \le f(x) - \alpha t_k\|\nabla f(x)\|_2^2$

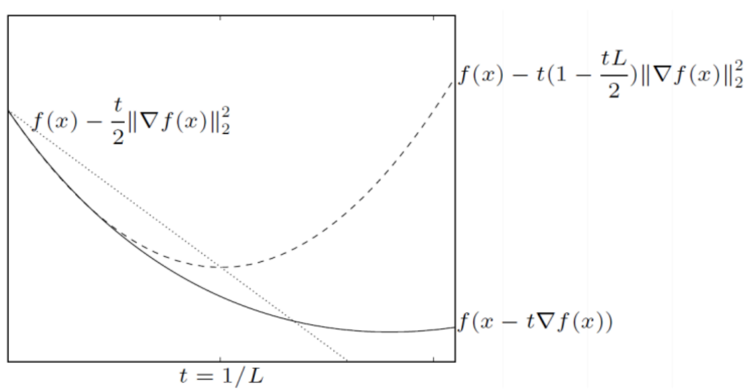
$0 < \beta < 1, \alpha = \frac{1}{2}$ mostly for simplify proofs

前面说到quadratic upper bound有:

$$f(x-t\nabla f(x)) \le f(x) - t(1-\frac{Lt}{2})\|\nabla f(x)\|_2^2$$

原函数,quadratic upper bound,line search对应的三条曲线如下所示(其中步长t为自变量,纵轴表示函数值，其中$t=\frac{1}{L}$是quadratic upper bound的极小值点)。因此最好的步长t是line search方法和原函数的交点处的t值(记为t_{opt})，t值太大也不可取，因为我们无法确定t_{opt}右侧的情况，但是能够保证在t_{opt}左侧:line search直线始终是原函数的linear upper bound。

line search with $\alpha = 1/2$ if f has a Lipschitz continuous gradient



3.1.3 Gradient method for strongly convex function

前面介绍的是L-Lipschitz Continuous Gradient的情况，利用了函数的quadratic upper bound来证明收敛性。这里主要介绍在μ-strongly convex情况下GD收敛性的证明。

首先回顾co-coercivity性质之一：

> co-coercvity of ∇g gives
> $(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{\mu L}{\mu+L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2$

令$x^+ = x - t\nabla f(x), 0 < t \le \frac{2}{\mu+L}$
则有:

$$\|x^+ - x^*\| = \|x^+ - t\nabla f(x) - x^*\|_2^2(\text{拆开})$$

$$= \|x^+ - x^*\|_2^2 - 2t\nabla f(x)^T(x-x^*) + t^2\|\nabla f(x)\|_2^2$$

$$\le \|x^+ - x^*\| - \frac{2t\mu L}{\mu+L}\|x-x^*\|_2^2 + (t^2 - \frac{2t}{t+\mu})\|\nabla f(x)\|_2^2(\text{此处有co-coercivity of gradient性质得到})$$

$$\leq (1 - t\frac{2\mu L}{\mu+L})\|x - x^*\|_2^2 + t(t - \frac{2}{\mu+L})\|\nabla f(x)\|_2^2$$

$$\leq (1 - t\frac{2\mu L}{\mu+L})\|x - x^*\|_2^2$$

综上，使用归纳法可以得到：

$$\|x^{(k)} - x^*\|_2^2 \leq c^k \|x^{(0)} - x^*\|_2^2, c = 1 - t\frac{2\mu L}{\mu+L}$$

所以有：

$$f(x^{(k)}) - f(x^*) \leq \frac{L}{2}\|x^{(k)} - x^*\| \leq \frac{c^k L}{2}\|x^0 - x^*\|_2^2$$

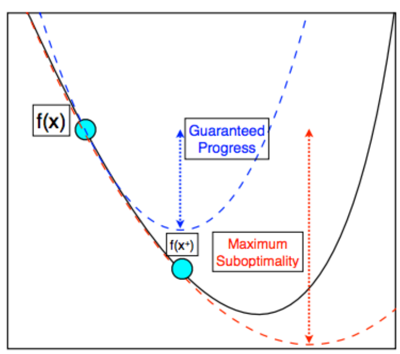
因此梯度下降法达到$f(x^{(k)}) - f(x^*) \leq \epsilon$的精度需要$o(log(\frac{1}{\epsilon}))$的时间复杂度。//TODOTODO

3.1.4 利用二阶上界和二阶下界来证明GD收敛性

这里提供另外一种证明方法，比较直观理解GD。`https://www.cs.ubc.ca/~schmidtm/Documents/2013_Notes_ConvexOptim.pdf`

- We have bounds on x^+ and x^*:

$$f(x^+) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2, \quad f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$



其中

$f(x^+) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$是有L-Lipschitz Continuous gradient的性质得到。(令$y = x^+ = x - \alpha\nabla f(x), \alpha = \frac{1}{L}$，带入即可。)

$f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$是由$\mu$-strongly convex性质得到。

- We have bounds on x^+ and x^*:

$$f(x^+) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|^2, \quad f(x^*) \ge f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

combine them to get

$$f(x^+) \le f(x) - \frac{\mu}{L}[f(x) - f(x^*)]$$

$$f(x^+) - f(x^*) \le \left(1 - \frac{\mu}{L}\right)[f(x) - f(x^*)]$$

- This gives a linear convergence rate:

$$f(x^t) - f(x^*) \le \left(1 - \frac{\mu}{L}\right)^t [f(x^0) - f(x^*)]$$

- Each iteration multiplies the error by a fixed amount.

(very fast if μ/L is not too close to one)

3.1.5 BB step

Consider the problem

$$\min f(x)$$

- Steepest gradient descent method: $x^{k+1} := x^k - \alpha^k g^k$:

$$\alpha^k := \arg\min_{\alpha} f(x^k - \alpha g^k)$$

- Let $s^{k-1} := x^k - x^{k-1}$ and $y^{k-1} := g^k - g^{k-1}$.
- BB: choose α^k so that $D^k := \alpha^k I$ satisfies $D^k y^{k-1} = s^{k-1}$:

$$\alpha^k := \frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}} \text{ or } \alpha^k := \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}}$$

3.2 随机梯度法(SGD)的收敛性

本节主要证明SGD的收敛性，先回顾一下基本公式:

- ERM(Empirical Risk Minimization) problem
  $min_{x \in R^n} f(x) = \frac{1}{n}\sum_1^n f_i(x)$
- 前面介绍的Gradient Descent如下:
  $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$
- 本节的SGD如下:
  $x_{k+1} = x_k - \alpha_k \nabla f_{s_k}(x_k)$
  其中的s_k表示从1,...n中采样。

和前面一样，在证明收敛性之前，给出四个基本假设方便后面证明。

- f(x) is L-smooth.
  $||\nabla f(x) - \nabla f(y)||_2^2 \le L||x-y||_2^2$
- f(x) is μ-strongly convex .
  $< \nabla f(x) - \nabla f(y), x-y > \ge \mu||x-y||_2^2$
- $E_s[\nabla f_s(x)] = \nabla f(x)$
  随机取的过程长远来看相当于按顺序取。
- $E_s||\nabla f_s(x)||^2 \le M$
  随机取的x的二阶导数不能无限大，有一个上界M。

证明会用到以下信息:
(1).α_k为固定长度α
(2). E[X] = E[E[X|Y]]
(3).strong monotonicity (coercivity) of ∇f : $(\nabla f(x) - \nabla f(y))^T(x-y) \ge \mu||x-y||_2^2, \forall x, y \in dom f$
在SGD收敛性证明中可以得到$\mu\nabla_k^2 \le < \nabla f(x_k), x_k - x^* >$

- $$\begin{aligned}\Delta_{k+1}^2 &= \|x_{k+1}-x^*\|_2^2 = \|x_k - \alpha_k \nabla f_{s_k}(x_k) - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\alpha_k \langle \nabla f_{s_k}(x_k), x_k - x^* \rangle + \alpha_k^2 \|\nabla f_{s_k}(x_k)\|_2^2 \\ &= \Delta_k^2 - 2\alpha_k \langle \nabla f_{s_k}(x_k), x_k - x^* \rangle + \alpha_k^2 \|\nabla f_{s_k}(x_k)\|_2^2\end{aligned}$$
- Using $E[X] = E[E[X|Y]]$:

$$\begin{aligned}& \mathbb{E}_{s_1,\ldots,s_k}[\langle \nabla f_{s_k}(x_k), x_k - x^* \rangle] = \mathbb{E}_{s_1,\ldots,s_{k-1}}[\mathbb{E}_{s_k}[\langle \nabla f_{s_k}(x_k), x_k - x^* \rangle]] \\ = \; & \mathbb{E}_{s_1,\ldots,s_{k-1}}[\langle \mathbb{E}_{s_k}[\nabla f_{s_k}(x_k)], x_k - x^* \rangle] \\ = \; & \mathbb{E}_{s_1,\ldots,s_{k-1}}[\langle \nabla f(x_k), x_k - x^* \rangle] \\ = \; & \mathbb{E}_{s_1,\ldots,s_k}[\langle \nabla f(x_k), x_k - x^* \rangle]\end{aligned}$$

- By the strongly convexity

$$\mathbb{E}_{s_1,\ldots,s_k}(\Delta_{k+1}^2) \leq (1-2\alpha\mu)\mathbb{E}_{s_1,\ldots,s_k}(\Delta_k^2) + \alpha^2 M^2 \qquad (49)$$

上面就得到了一个∇_{k+1}^2的递归关系，并且利用$0 \leq 2\alpha\mu \leq 1$的假设可以得到如下结果:

Proof of Theorem

- Taking induction from $k = 1$ to $k = T$, we have

$$\mathbb{E}_{s_1,\ldots,s_T}(\Delta_{T+1}^2) \leq (1-2\alpha\mu)^T \Delta_1^2 + \sum_{i=0}^{T-1} (1-2\alpha\mu)^i \alpha^2 M^2 \quad (50)$$

- under the assumption that $0 \leq 2\alpha\mu \leq 1$, we have

$$\sum_{i=0}^{\infty} (1-2\alpha\mu)^i = \frac{1}{2\alpha\mu}$$

- Then

$$\mathbb{E}_{s_1,\ldots,s_T}(\Delta_{T+1}^2) \leq (1-2\alpha\mu)^T \Delta_1^2 + \frac{\alpha M^2}{2\mu} \qquad (51)$$

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:

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

Stochastic Gradient Descend:

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

$$\mathbb{E}\Delta_{k+1}^2 \le \underbrace{(1-2\alpha\mu+2\alpha^2L^2)\mathbb{E}\Delta_k^2}_{A} + \underbrace{\boxed{2\alpha^2\mathbb{E}||\nabla f_{s_k}(x_k)-\nabla f(x_k)||_2^2}}_{B} \quad (54)$$

- a worst case convergence rate of $\sim 1/T$ for SGD
- In practice, the actual convergence rate may be somewhat better than this bound.
- Initially, $B << A$ and we observe the linear rate regime, once $B > A$ we observe $1/T$ rate.
- How to reduce variance term B to speed up SGD?
  - SAG (Stochastic average gradient)
  - SAGA
  - SVRG (Stochastic variance reduced gradient)

5 随机优化算法在深度学习中的应用

6 总结

7 附录

7.1 基本性质

一些有用的链接:

https://www.cs.ubc.ca/~schmidtm/Documents/2013_Notes_ConvexOptim.pdf

http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf

http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf

http://bicmr.pku.edu.cn/~wenzw/opt2015/lect-gm.pdf

http://bicmr.pku.edu.cn/~wenzw/bigdata/lect-sto.pdf

- convex function

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

上面的式子也称为jensen不等式，

对于凸函数一阶可微的情况下等价于:

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

将f(y)在x处二阶展开,可以得到如下结果:

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

很显然有:

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

对于凸函数二阶可微的情况下等价于:

$$\nabla^2 f(x) \geq 0$$

凸函数梯度存在单调性(monotonicity):

a differentiable function f is convex if and only if dom f is convex and $(f(x) - f(y))^T(x - y) \geq 0$,for all x,y $\in$ dom f,$x \neq y$

注意“单调性,定义域集合为凸”和“函数f为凸”是等价的。

如果f(x)为凸函数,那么

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

结合两者即得证。

- M-Lipschitz Continuous

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

M-Lipschitz Continuous有如下性质(利用梯度的定义即可得证):

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

- L-Lipschitz Smoothness 也称为L-Lipschitz continuous gradient，意思是f的梯度满足L-Lipchitz Continuous

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

此处不要求f为凸函数

L-Lipschitz Smoothness有以下性质:

(1).$\nabla^2 f(x) \leq LI$

(2).$\frac{L}{2}x^T x - f(x)$为凸函数.

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

(4).$\frac{1}{2L}\|\nabla f(x)\|_2^2 \le f(x) - f(x^*) \le \frac{L}{2}\|x-x^*\|_2^2$

(1).
利用导数的定义很容易证明。
(2).
利用∇f的Lipschitz Continuity:
$\|\nabla f(x) - \nabla f(y)\|\|x-y\| \le L\|x-y\|_2^2$
利用柯西不等式可得:
$(\nabla f(x) - \nabla f(y))^T(x-y) \le \|\nabla f(x) - \nabla f(y)\|\|x-y\| \le L\|x-y\|_2^2$
移到同一边，整理得:
$(\nabla f(x) - Lx - \nabla f(y) + Ly)^T(x-y) \le 0$
换号:
$(Lx - \nabla f(x) - (Ly - \nabla f(y)))^T(x-y) \le 0$
其中$Lx - \nabla f(x)$是$\frac{L}{2}x^T x - f(x)$的梯度。
且$\frac{L}{2}x^T x - f(x)$的dom是convex set，由于gradient monotonicity的等价性知:
$\frac{L}{2}x^T x - f(x)$是凸函数。
(3).
将f(y)在x出泰勒展开:
$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$
并且由于$\nabla^2 f(z) \le LI$:
$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$
,此处的$f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$是f(y)的二阶上界(quadratic upper bound),是关于y的二次函数。
(4).首先证明$f(x) - f(x^*) \le \frac{L}{2}\|x-x^*\|_2^2$:
由于x^*为f的极值点，所以$f(x^*) = 0$,并且:
$f(x) \le f(x^*) + \nabla f(x^*)^T(x-x^*) + \frac{L}{2}\|x-x^*\|_2^2$
化简得证，接下来证明$\frac{1}{2L}\|\nabla f(x)\|_2^2 \le f(x) - f(x^*)$
因为x^*为极值点,所以$f(x^*) \le f(y)$:
$f(x^*) \le inf_{y \in domf}(f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2)$
因为f的定义域为R^n,这里不妨令$y = x - \frac{1}{L}\nabla f(x)$
化简整理即可证明左边不等式。

L-LipschitzContinuous相当于确定了f的二阶上界。

- μ-strongly convex characterization

$$f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2, \text{for } \forall \text{ y,x}$$

同时也等价于：

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

μ-strongly convex characterization有以下性质:

Properties of Lipschitz-Continuous Gradient

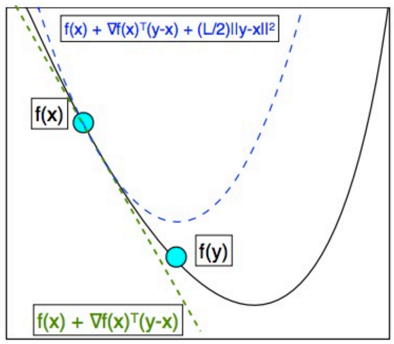
- From Taylor's theorem, for some z we have:

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(z)(y-x)$$

- Use that $\nabla^2 f(z) \preceq LI$.

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$$

- *Global quadratic upper bound on function value.*



(1).$\nabla^2 f(x) \geq \mu I$

(2). strong monotonicity (coercivity) of ∇f : $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \mu\|x-y\|_2^2, \forall x, y \in dom f$

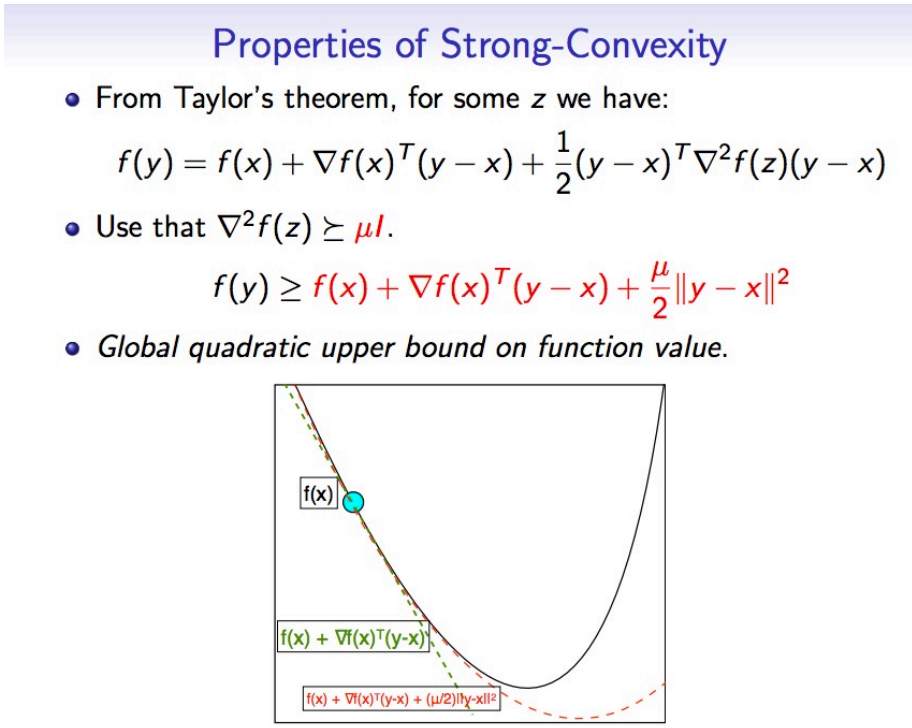
(3).$f(x) - \frac{\mu}{2}x^T x$为凸函数.

(4).$f(y) \geq f(x) + < \nabla f(x), y-x > + \frac{\mu}{2}\|x-y\|_2^2$

(5).$\frac{\mu}{2}\|x-x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$

(1). $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$,for ∀ y,x
同时又有在x点处泰勒展开:
$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{\nabla^2 f(z)}{2}\|y-x\|_2^2$,for ∀ y,x
所以:
$f(x) + \nabla f(x)^T(y-x) + \frac{\nabla^2 f(z)}{2}\|y-x\|_2^2 \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$
整理即得证。
(2).由二阶导数定义知:
(3).
(4).定义
(5). 证法和上面类似。相当于二阶下界:

μ-strongly convex相当于确定了f的二阶下界。



L-Lipschitz-Smooth和μ**-strongly convex**共同对应着f的二阶上界和二阶下界，两者之间的性质有遥相呼应的关系。

7.2 Co-coercivity of gradient

http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf

> if f is convex with dom f = R^n, and $\frac{L}{2}x^Tx - f(x)$ is convex,then
> $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2, \forall x, y$
> this property is known as co-coercivity of ∇f with parameter $\frac{1}{L}$

co-coercivity的证明如下:

构造以下两个函数:
$f_x(z) = f(z) - \nabla f(x)^T z, f_y(z) = f(z) - \nabla f(y)^T z,$
那么可以得到以下两个结论:
(1). 通过求导数可以知道: $f_x(z)$在z=x处取极值, $f_y(z)$在z=y处取极值.
(2). $\frac{L}{2}z^T z - f_x(z)$为凸函数。（因为其二阶导数等于L，大于0）

因为$f_x(z)$满足L-Lipschitz continuity，且在x处取极值，所以有: $f_x(y) - f_x(x) \geq \frac{1}{2L}\|\nabla f_x(y)\|_2^2$

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

上面两个式子相加即得证。
$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2, \forall x, y$

co-coercivity之间的等价性:

Lipschitz continuity of $\nabla f(x)$ =>
convexity of $\frac{L}{2}x^T x - f(x)$ =>
co-coercivity of $\nabla f(x)$ =>
Lipschitz continuity of $\nabla f(x)$

下面给出证明：

(1).Lipschitz continuity of $\nabla f(x)$ 到convexity of $\frac{L}{2}x^T x - f(x)$ 的充分性:

TODO
(2). convexity of $\frac{L}{2}x^T x - f(x)$ 到co-coercivity of $\nabla f(x)$的充分性上面已经证明。
(3). co-coercivity of $\nabla f(x)$到Lipschitz continuity of $\nabla f(x)$ 的充分性证明如下:
$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L(\nabla f(x) - \nabla f(y))^T(x-y)$
利用柯西不等式:
$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L(\nabla f(x) - \nabla f(y))^T(x-y) \leq L\|\nabla f(x) - \nabla f(y))\|_2\|(x-y)\|_2$
化简即得证。

上面的等价性刻画co-coercivity of $\nabla f(x)$,Lipschitz continuity of $\nabla f(x)$,convexity of $\frac{L}{2}x^T x - f(x)$ 三者之间的关系。

7.2.1 co-coercivity的扩展

if f is strongly convex and ∇f is Lipschitz Continuous, then
(1). $g(x) = f(x) - \frac{\mu}{2}x^T x$ is convex.
(2).∇g is Lipschitz Continuous with parameter $L - \mu$
(3). co-coercvity of ∇g gives
$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{\mu L}{\mu+L}\|x - y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2$

下面依次给出证明:

(1).
直接对g(x)求二阶导数得到$\nabla^2 g(x) = \nabla^2 f(x) - \mu > 0$,因此为凸函数。
(2).这里看到网上的一个答案证明的是小于等于$L + \mu$, https://math.stackexchange.com/questions/1645272/extension-of-co-coercivity-in-strongly-convex-functions
这个上界太宽泛，需要进一步压缩。证明如下:
由∇g的co-coercivity的性质知道:
$(\nabla g(x) - \nabla g(y))^T(x - y) \geq \frac{1}{L}\|\nabla g(x) - \nabla g(y)\|_2^2, \forall x, y$
将g(x)定义带入不等式，整理得到: $(\nabla g(x) - \nabla g(y))^T(x - y) \leq \frac{L\mu+\mu^2}{L+2\mu}\|x - y\|_2^2 + \frac{1}{L+2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2$
可以发现这是一个比$\frac{\mu L}{\mu+L}\|x - y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2$更紧的上界，所以(3)得证。

7.3 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:

variational characterization :

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

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:

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

所以有:

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

8 Reference

Machine Learning Summer School: `http://learning.mpi-sws.org/mlss2016/slides/mlss_2016_cadiz_slides.pdf`