多种优化算法的收敛性和收敛速率分析

惠成煊
信息科学与工程学院
武汉科技大学
carr_001@163.com

Abstract

本文主要对包括梯度下降法，子梯度方法等多种梯度下降法的收敛性进行总结。本文大部分内容是对 CMU 2018 凸优化课程的重新推导, 少部分参考其他网站内容, 均已在最后引用。

1 基本知识

1.1 常用概念

1.2 常用公式

1.2.1 Sherman-Morrison 公式

$$\left(A+uv^T\right)^{-1}=A^{-1}-\frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u} \tag{1}$$

证明 [5]: 令 $X=A+uv^T, Y=A^{-1}-\frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$:

$$\begin{aligned}
XY&=\left(A+uv^T\right)\left(A^{-1}-\frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}\right)\\
&=AA^{-1}+uv^TA^{-1}-\frac{AA^{-1}uv^TA^{-1}+uv^TA^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}\\
&=I+uv^TA^{-1}-\frac{uv^TA^{-1}+uv^TA^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}\\
&=I+uv^TA^{-1}-\frac{u\left(1+v^TA^{-1}u\right)v^TA^{-1}}{1+v^TA^{-1}u}\\
&=I+uv^TA^{-1}-uv^TA^{-1}\\
&=I
\end{aligned} \tag{2}$$

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.2.2 Woodbury 公式

此公式是对 Sherman-Morrison 公式的泛化

$$(A+UDV)^{-1}=A^{-1}-A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}VA^{-1} \tag{3}$$

证明：

$$\begin{aligned}
&\left(A^{-1}-A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right)(A+UDV)\\
=&I+A^{-1}UDV-A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}V-A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}VA^{-1}UDV\\
=&I+A^{-1}UDV-A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}\left(D^{-1}+VA^{-1}U\right)DV\\
=&I
\end{aligned} \tag{4}$$

1.3 常用性质

性质 1.1: 凸函数一阶性质

$$f(y)\geq f(x)+\partial f(x)^T(y-x) \tag{5}$$

性质 1.2: 如果凸函数二次导数小于等于 L 的，那么

$$f(y)\leq f(x)+\nabla f(x)^T(y-x)+\frac{1}{2}L\left\|y-x\right\|_2^2 \tag{6}$$

最优条件: 对于任意的函数 f，当且仅当在 $x^\star$ 时满足:

$$f\left(x^\star\right)=\min_x f(x)\Longleftrightarrow 0\in\partial f\left(x^\star\right) \tag{7}$$

,$x^\star$ 是一个最优点. 例子 1: 如果 f 可微，那么 $0=\partial f(x^\star)$. 例子 2: 如果 f 不可微，那么 $\partial f(x^\star)$ 是次微分集合，当 $x^\star$ 是一个最优点意味着 0 在这个集合中。

定理 1 如果 f 是 $m-strongconvex$ 的，那么

$$f(x)-f^*\leq\frac{1}{2m}\|\nabla f(x)\|_2^2 \tag{8}$$

证明: 根据条件容易写出:

$$f(y)\geqslant f(x)+\nabla f(x)^\top(y-x)+\frac{m}{2}\|y-x\|_2^2 \tag{9}$$

两边同时对 y 最小化，可以得到:

$$f^*\min_y f(x)+\nabla f(x)^\top(y-x)+\frac{m}{2}y-x_2^2 \tag{10}$$

右边求解 $y=-\frac{1}{m}\nabla f(x)+x$，带入右边整理后可得:

$$f(x)-\frac{1}{2m}\nabla f(x)_2^2\leq f^* \tag{11}$$

。

1.3.1 共轭函数及其性质

定义 1 给定函数 $f : \mathbb{R}^n \to \mathbb{R}$，函数:

$$f^*(y) = \max_x y^T x - f(x) \tag{12}$$

称为共轭函数。**性质 1** 如果函数 f 是闭凸的 (一个函数是封闭的当且仅当所有 sub-level set 是封闭的)，那么 $f^{\star\star} = f$。证明:

性质 2 如果 f 是闭凸的，那么 $x \in \partial f^*(y) \Longleftrightarrow y \in \partial f(x) \Longleftrightarrow x \in \underset{z}{\operatorname{argmin}} f(z) - y^T z$

性质 3 如果 f 是严格凸的，那么 $\nabla f^*(y) = \underset{z}{\operatorname{argmin}} f(z) - y^T z$。

1.4 收敛速率名词

1.4.1 第一种定义方法

假设 $\left\{x^{(k)}\right\} \in R^n$ 是一个收敛数列，一般这样定义收敛数列：

$$\lim_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k} = \lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} = C \tag{13}$$

如果 $C = 0$，那么说数列 $\left\{x^{(k)}\right\}$ 超线性收敛；如果 $C \in (0,1)$，那么说数列 $\left\{x^{(k)}\right\}$ 线性收敛；如果 $C = 1$，那么说数列 $\left\{x^{(k)}\right\}$ 次线性收敛。

二阶收敛:

$$\lim_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k} = \lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|^2} = C \tag{14}$$

如果 $C > 0$，那么说数列 $\left\{x^{(k)}\right\}$ 二次收敛；

1.4.2 第二种定义方法

迭代 k 次之后，ϵ 是多少: 如果 $\epsilon_k = e^{-e^k}$, 那么我们说超线性收敛；如果 $\epsilon_k = e^k$, 那么我们说线性收敛；如果 $\epsilon_k = \frac{1}{k}$、$\epsilon_k = \frac{1}{\sqrt{k}}$、$\epsilon_k = \frac{1}{k^2}$, 那么我们说线性收敛；

1.4.3 第三种定义方法

为了使误差小于 ϵ, 需要 k 为多少: 如果 $k = O\left(\log\left(\log\left(\frac{1}{\epsilon}\right)\right)\right)$, 那么说超线性收敛；如果 $k = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$, 那么说线性收敛；如果 $k = O\left(\frac{1}{\epsilon}\right)$、$k = O\left(\frac{1}{\epsilon^2}\right)$、$k = O\left(\frac{1}{\sqrt{\epsilon}}\right)$, 那么说次线性收敛.

1.4.4 三种方法总结

三种方法本质是相同的。这里收敛速率的定义与计算机中的定义方式有所不同，计算机中的线性收敛是 $O(n)$，意味着时间复杂度与问题的规模是线性的。而这里是可以理解为误差 ϵ 与迭代次数 k 之间的关系，如果是误差随着迭代次数是指数衰减的，那么就是就是线性收敛 (针对取对数之后的结果)，如果比指数收敛慢，那么都是次线性收敛，如果比指衰减快，那么都是超线性收敛。

为什么将指数衰减定义为线性收敛？个人认为是误差一般比较小，因此需要取对数查看波动情况，那么在取对数后的结果就是对收敛性的定义。

迭代次数随着问题规模的变化: 在计算机中，计算复杂度反映了计算复杂度随着问题规模的增长情况，从而通过计算复杂度来反映算法的优劣; 在优化中，考虑将误差 ϵ 减少 100 倍数之后需要增加的迭代次数，对于次线性收敛 $k=O\left(\frac{1}{\epsilon}\right)$，需要增加 100 倍迭代次数; 对于次线性收敛 $k=O\left(\frac{1}{\epsilon^2}\right)$，需要增加 10000 倍迭代次数; 对于次线性收敛 $k=O\left(\frac{1}{\sqrt{\epsilon}}\right)$，需要增加 10 倍迭代次数; 对于线性收敛, 需要增加常数次数 $\log 100$。

例子，令数列为 $x_k=\frac{1}{2^k}$，$x^\star=0$。在计算机可以理解为二分法。查找 0 的时间复杂度是 $log(n)$，那么在这里就是线性收敛。

2 一阶算法

2.1 梯度下降法

假设函数 $f(x)$ 凸可微，其微分满足 Lipschitz 连续性 $\|\nabla f(x)-\nabla f(y)\|_2 \leq L\|x-y\|_2$, 选择步长 $t \leq 1/L$。

2.1.1 收敛性分析

由公式6, 令 $x^+=x-t\nabla f(x)$, 可得

$$\begin{aligned}
f\left(x^{+}\right) &\leq f(x)+\nabla f(x)^T\left(x^{+}-x\right)+\frac{1}{2}L\left\|x^{+}-x\right\|_2^2 \\
&= f(x)+\nabla f(x)^T(x-t\nabla f(x)-x)+\frac{1}{2}L\|x-t\nabla f(x)-x\|_2^2 \\
&= f(x)-\nabla f(x)^T t\nabla f(x)+\frac{1}{2}L\|t\nabla f(x)\|_2^2 \\
&= f(x)-t\|\nabla f(x)\|_2^2+\frac{1}{2}Lt^2\|\nabla f(x)\|_2^2 \\
&= f(x)-\left(1-\frac{1}{2}Lt^2\right)\|\nabla f(x)\|_2^2
\end{aligned} \tag{15}$$

选择 $t \leq 1/L$, 那么

$$f\left(x^{+}\right) \leq f(x)-\frac{1}{2}t\|\nabla f(x)\|_2^2 \tag{16}$$

由于 $\frac{1}{2}t\|\nabla f(x)\|_2^2$ 大于零，所有每次迭代之后，$f(x_+)$ 都是严格减的，所有能够保证收敛。

2.1.2 收敛速率分析

定理 2.1 假设函数 $f(x)$ 可微，其微分满足 Lipschitz 连续性 $\|\nabla f(x)-\nabla f(y)\|_2 \leq L\|x-y\|_2$, 选择步长 $t \leq 1/L$, 那么我们有

$$f(x^{(k)})-f(x^*) \leq \frac{\left\|x^{(0)}-x^*\right\|_2^2}{2tk} \tag{17}$$

证明由性质 1.1 我们可以写出:

$$
\begin{aligned}
f\left(x^{*}\right) \geq f(x)+\nabla f(x)^{T}\left(x^{*}-x\right) \\
f(x) \leq f\left(x^{*}\right)+\nabla f(x)^{T}\left(x-x^{*}\right)
\end{aligned} \tag{18}
$$

现在将此公式带入到公式16：

$$
\begin{aligned}
& f\left(x^{+}\right) \leq f\left(x^{*}\right)+\nabla f(x)^{T}\left(x-x^{*}\right)-\tfrac{t}{2}\|\nabla f(x)\|_{2}^{2} \\
& f\left(x^{+}\right)-f\left(x^{*}\right) \leq \tfrac{1}{2t}\left(2t \nabla f(x)^{T}\left(x-x^{*}\right)-t^{2}\|\nabla f(x)\|_{2}^{2}\right) \\
& f\left(x^{+}\right)-f\left(x^{*}\right) \leq \tfrac{1}{2t}\left(2t \nabla f(x)^{T}\left(x-x^{*}\right)-t^{2}\|\nabla f(x)\|_{2}^{2}-\left\|x-x^{*}\right\|_{2}^{2}+\left\|x-x^{*}\right\|_{2}^{2}\right) \\
& f\left(x^{+}\right)-f\left(x^{*}\right) \leq \tfrac{1}{2t}\left(\left\|x-x^{*}\right\|_{2}^{2}-\left\|x-t \nabla f(x)-x^{*}\right\|_{2}^{2}\right)
\end{aligned} \tag{19}
$$

我们得到：

$$
f\left(x^{+}\right)-f\left(x^{*}\right) \leq \frac{1}{2t}\left(\left\|x-x^{*}\right\|_{2}^{2}-\left\|x^{+}-x^{*}\right\|_{2}^{2}\right) \tag{20}
$$

进行迭代求和：

$$
\begin{aligned}
\sum_{i=1}^{k} f\left(x^{(i)}-f\left(x^{*}\right) \right. & \leq \sum_{i=1}^{k} \frac{1}{2t}\left(\left\|x^{(i-1)}-x^{*}\right\|_{2}^{2}-\left\|x^{(i)}-x^{*}\right\|_{2}^{2}\right) \\
& =\frac{1}{2t}\left(\left\|x^{(0)}-x^{*}\right\|_{2}^{2}-\left\|x^{(k)}-x^{*}\right\|_{2}^{2}\right)
\end{aligned} \tag{21}
$$

得到：

$$
\begin{aligned}
\textstyle\sum_{i=1}^{k} f\left(x^{(i)}-f\left(x^{*}\right) \right. & \leq \tfrac{1}{2t}\left(\left\|x^{(0)}-x^{*}\right\|_{2}^{2}\right) \\
\tfrac{1}{k} \textstyle\sum_{i=1}^{k} f\left(x^{(i)}-f\left(x^{*}\right) \right. & \leq \tfrac{1}{2kt}\left(\left\|x^{(0)}-x^{*}\right\|_{2}^{2}\right)
\end{aligned} \tag{22}
$$

由于 $f(x)$ 每次迭代都递减，所以：

$$
f\left(x^{(k)}\right)-f\left(x^{*}\right) \leq \frac{1}{k} \sum_{i=1}^{k} f\left(x^{(i)}\right)-f\left(x^{*}\right) \tag{23}
$$

带入公式22，我们得到公式17。证毕。

2.2 梯度下降法-Backtracking

待完成

2.3 梯度下降法-满足强凸性

2.4 非凸函数的梯度下降法

2.4.1 收敛速率分析

目标函数是非凸函数，我们自然不能够要求梯度下降法到达全局最优点，因此我们分析到达的 ϵ 不动点的情况，即 $\|\nabla f(x)\|_{2} \leq \epsilon$。假设函数 $f(x)$ 非凸可微，其微分满足 Lipschitz 连续性 $\|\nabla f(x)-\nabla f(y)\|_{2} \leq L\|x-y\|_{2}$, 选择步长 $t \leq 1/L$, 我们可以得到与公式16相同的结果。我们将公式16转化为：

$$
\|\nabla f(x)\|_{2}^{2} \leq \frac{2}{t}\left(f(x)-f\left(x^{+}\right)\right) \tag{24}
$$

进行累加，我们得到：

$$\sum_{i=0}^{k}\left\|\nabla f\left(x^{(i)}\right)\right\|_2^2 \leq \frac{2}{t}\left(f\left(x^{(0)}\right)-f\left(x^{(k+1)}\right)\right) \leq \frac{2}{t}\left(f\left(x^{(0)}\right)-f^{\star}\right) \tag{25}$$

$$(k+1)\min_{i=0,\ldots,k}\left\|\nabla f\left(x^{(i)}\right)\right\|_2^2 \leq \frac{2}{t}\left(f\left(x^{(0)}\right)-f^{\star}\right) \tag{26}$$

$$\min_{i=0,\ldots,k}\left\|\nabla f\left(x^{(i)}\right)\right\|_2 \leq \sqrt{\frac{2\left(f\left(x^{(0)}\right)-f^{\star}\right)}{t(k+1)}} \tag{27}$$

由于非凸函数寻求局部最低点，这意味我们寻求 $\|\nabla f(x)\|_2 \leq \epsilon$，从公式27我们知道，收敛速度为：$O(\frac{1}{\sqrt{k}})$。

2.5 子梯度法

定义 2.5.1 子梯度定义：对于一个凸函数，其子梯度是满足下面的 $g \in \mathbb{R}^n$：

$$f(y) \geq f(x) + g^T(y-x) \tag{28}$$

2.5.1 收敛性分析

子梯度方法无法得到类似公式16的迭代下降的保证，但是能够得到 k 次迭代之后与最优点的上界。下面得到子梯度方法所谓的基本不等式。对下面式子展开：

$$\begin{aligned}\left\|x^{(k)}-x^{\star}\right\|_2^2 &\leq \left\|x^{(k-1)}-t_k g^{(k-1)}-x^{\star}\right\|_2^2 \\ &= \left\|x^{(k-1)}-x^{\star}\right\|_2^2 - 2t_{k-1}g^{(k-1)}(x^{(k-1)}-x^{\star}) + t_k^2 g^{(k-1)}\end{aligned} \tag{29}$$

由子梯度定义我们可以得到 $g^{(k-1)}(x^{k-1}-x^{\star}) \geq f\left(x^{(k-1)}\right)-f\left(x^{\star}\right)$, 那么上式：

$$\left\|x^{(k)}-x^{\star}\right\|_2^2 \leq \left\|x^{(k-1)}-x^{\star}\right\|_2^2 - 2t_k\left(f\left(x^{(k-1)}\right)-f\left(x^{\star}\right)\right) + t_k^2\left\|g^{(k-1)}\right\|_2^2 \tag{30}$$

重复右边第一项，可以得到：

$$\left\|x^{(k)}-x^{\star}\right\|_2^2 \leq \left\|x^{(0)}-x^{\star}\right\|_2^2 - 2\sum_{i=1}^{k} t_i\left(f\left(x^{(i-1)}\right)-f\left(x^{\star}\right)\right) + \sum_{i=1}^{k} t_i^2\left\|g^{(i-1)}\right\|_2^2 \tag{31}$$

$$0 \leq \left\|x^{(k)}-x^{\star}\right\|_2^2 \leq \left\|x^{(0)}-x^{\star}\right\|_2^2 - 2\sum_{i=1}^{k} t_i\left(f\left(x^{(i-1)}\right)-f\left(x^{\star}\right)\right) + \sum_{i=1}^{k} t_i^2\left\|g^{(i-1)}\right\|_2^2 \tag{32}$$

令 $R = \left\|x^{(0)}-x^{\star}\right\|_2$, 结合 Lipshcitz 假设, 重写上式：

$$0 \leq R^2 - 2\sum_{i=1}^{k} t_i\left(f\left(x^{(i-1)}\right)-f\left(x^{\star}\right)\right) + G^2\sum_{i=1}^{k} t_i^2 \tag{33}$$

做简单变换：

$$2\sum_{i=1}^{k} t_i\left(f\left(x^{(i-1)}\right)-f\left(x^{\star}\right)\right) \leq R^2 + G^2\sum_{i=1}^{k} t_i^2 \tag{34}$$

引入 $f\left(x_{\text{best}}^{(k)}\right)=\min_{i=0,\ldots k} f\left(x^{(i)}\right)$, 可得

$$2(f\left(x_{\text{best}}^{(k)}\right)-f\left(x^{\star}\right))\sum_{i=1}^{k} t_i \leq 2\sum_{i=1}^{k} t_i\left(f\left(x^{(i-1)}\right)-f\left(x^{\star}\right)\right) \leq R^2+G^2\sum_{i=1}^{k} t_i^2 \tag{35}$$

由此可得：

$$f\left(x_{\text{best}}^{(k)}\right)-f\left(x^{\star}\right) \leq \frac{R^2+G^2\sum_{i=1}^{k} t_i^2}{2\sum_{i=1}^{k} t_i} \tag{36}$$

这意味着，对于固定步长或者非固定步长，第 k 次迭代结果，我们可以通过约束右边项来限制与 $f\left(x^{\star}\right)$ 的距离，具体收敛性要看下面的步长选择。

2.5.2 收敛速率分析

对于公式36, 如果固定步长, 我们可以得到:

$$f\left(x_{\text{best}}^{(k)}\right)-f^{\star} \leq \frac{R^2}{2kt}+\frac{G^2 t}{2} \tag{37}$$

若想要 $\frac{R^2}{2kt}+\frac{G^2 t}{2} \leq \epsilon$, 我们使每一项 $\leq \epsilon$, 令 $t=\epsilon/G^2$ 满足条件，从而:

$$k=R^2/t \cdot 1/\epsilon=R^2 G^2/\epsilon^2 \tag{38}$$

这意味着收敛速率为 $O(\frac{1}{\epsilon^2})$。

对于自适应步长，以 Polyak step sizes:

$$t_k=\frac{f\left(x^{(k-1)}\right)-f^{\star}}{\|g^{(k-1)}\|_2^2} \tag{39}$$

为例。Polyak step sizes 是通过最小化公式30得到的。收敛速度仍然为 $O(\frac{1}{\epsilon^2})$。证明过程暂时略过。

2.6 近端梯度法

近端估计法针对问题：

$$f(x)=g(x)+h(x) \tag{40}$$

满足以下条件:1.g 是凸可微的;2.∇g 满足 Lipschitz 条件 $\|\nabla f(x)-\nabla f(y)\|_2 \leq L\|x-y\|_2$;3.$h$ 是不一定可微的凸函数。

近端梯度法：

$$x^{+}=\operatorname{prox}_t(x-t\nabla g(x)) \tag{41}$$

其中：

$$\operatorname{prox}_t(x)=\underset{u}{\operatorname{argmin}}\frac{1}{2t}\|x-u\|_2^2+h(u) \tag{42}$$

2.6.1 收敛性分析

定理 2.6 使用近端梯度下降法，我们可以得到如下结论:

$$f\left(x^{(k)}\right)-f\left(x^{\star}\right) \leq \frac{L}{2k}\left\|x^{(0)}-x^{\star}\right\|^{2} \tag{43}$$

证明: 首先，由性质 1.2（公式6）我们得到,

$$g\left(x^{+}\right) \leq g(x)+\nabla g(x)^{T}\left(x^{+}-x\right)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \tag{44}$$

由性质 1.1（公式5）, 可以得到:

$$g(z) \geq g(x)+\nabla g(x)^{T}(z-x) \Rightarrow g(x) \leq g(z)-\nabla g(x)^{T}(z-x) \tag{45}$$

结合公式45与公式44, 可得:

$$\begin{aligned} g\left(x^{+}\right) &\leq g(z)-\nabla g(x)^{T}(z-x)+\nabla g(x)^{T}\left(x^{+}-x\right)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \\ &= g(z)+\nabla g(x)^{T}\left(x^{+}-z\right)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \end{aligned} \tag{46}$$

对于 $h(x)$, 有子梯度定义，我们可以得到:

$$h(z) \geq h(x^{+})+g(z-x^{+}) \tag{47}$$

接下来我们尝试得到 g, 由公式41，我们可以得到:

$$x^{+}=\underset{u}{\operatorname{argmin}}\frac{1}{2t}\|x-t\nabla g(x)-u\|_{2}^{2}+h(u) \tag{48}$$

由最优条件可得:

$$0 \in \partial h(x)+\frac{1}{t}(x^{+}-x+t\nabla g(x)) \Rightarrow -\frac{1}{t}(x^{+}-x)-\nabla g(x) \in \partial h(x) \tag{49}$$

这意味着 $-\frac{1}{t}(x^{+}-x)-\nabla g(x)$ 是 $h(x)$ 在 x 处的次梯度,我们令 $g=-\frac{1}{t}(x^{+}-x)-\nabla g(x)$。记

$$G(x)=\frac{1}{t}\left(x-x^{+}\right) \tag{50}$$

得到 $g=G(x)-\nabla g(x)$。带入公式47, 进行移项，我们得到:

$$h(x^{+}) \leq h(z)+(\nabla g(x)-G(x))(z-x^{+}) \tag{51}$$

将公式46与公式51的左边与左边相加，右边与右边相加，得到:

$$g\left(x^{+}\right)+h(x^{+}) \leq g(z)+h(z)+\nabla g(x)^{T}\left(x^{+}-z\right)++(\nabla g(x)-G(x))^{T}(z-x^{+})+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \tag{52}$$

$$f(x^{+}) \leq f(z)+G(x)^{T}\left(x^{+}-z\right)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \tag{53}$$

有公式50, 可以得到:

$$x^{+}=x+tG(x) \tag{54}$$

带入公式53，可得:

$$
\begin{aligned}
f(x^{+}) &\leq f(z)+G(x)^{T}(x+tG(x)-z)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2} \\
&= f(z)+G(x)^{T}(x-z)+tG(x)^{T}G(x)+\frac{1}{2}L\left\|x^{+}-x\right\|_{2}^{2}
\end{aligned} \tag{55}
$$

为了方便分析，令 $t=\frac{1}{L}$，并将公式50带入上式，可以得到:

$$
f\left(x^{+}\right)-f(z) \leq G(x)^{T}(x-z)-\frac{1}{2L}\|G(x)\|^{2} \tag{56}
$$

令 $x^{\star}=z$，使用与公式19类似的配方法，可以得到：

$$
f\left(x^{+}\right)-f\left(x^{\star}\right) \leq \frac{L}{2}\left(\left\|x-x^{\star}\right\|^{2}-\left\|x^{+}-x^{\star}\right\|^{2}\right) \tag{57}
$$

两个进行累加，可以得到：

$$
f\left(x^{(k)}\right)-f\left(x^{\star}\right) \leq \frac{L}{2k}\left\|x^{(0)}-x^{\star}\right\|^{2} \tag{58}
$$

与子梯度法一样，近端梯度法通过限制 k 次迭代之后的误差来保证收敛性。

2.6.2 收敛速率分析

公式58意味着近端梯度法的的收敛速度是 $O(\frac{1}{\epsilon})$。

Nesterov 加速法: 可以提升到 $O(\frac{1}{\sqrt{\epsilon}})$(见 proximal_gradient proof.pdf);

如果可微函数 $g(x)$ 具有强凸性: 可以证明线性速率 (见 homework2.pdf);

2.7 统计梯度下降法

从子梯度法、梯度下降法我们可以知道: 如果凸函数不可微，那么收敛速率为 $O(\frac{1}{\epsilon^2})$; 如果可微而且假设 Lipschitz 连续性，那么收敛速度为 $O(\frac{1}{\epsilon})$；如果满足强凸性，那么能够达到线性收敛。对于 SGD，如果可微而且假设 Lipschitz 连续性，那么收敛速率依然为 $O(\frac{1}{\epsilon^2})$；如果满足强凸性，那么收敛速度为 $O(\frac{1}{\epsilon})$。

3 二阶算法

3.1 牛顿法

3.1.1 收敛性分析

牛顿法的递归过程可以分成两个部分，第一部分叫做 damp phase, 另一部分叫做 Pure phase。首先，给出总的收敛性分析结果:

$$
f\left(x^{(k)}\right)-f^{\star} \leq \begin{cases} \left(f\left(x^{(0)}\right)-f^{\star}\right)-\gamma k & \text{if } k \leq k_0 \\ \frac{2m^3}{M^2}\left(\frac{1}{2}\right)^{2^{k-k_0+1}} & \text{if } k > k_0 \end{cases} \tag{59}
$$

其中，k_0 满足 $\left\|\nabla f\left(x^{(k_0+1)}\right)\right\|_2 < \eta$，$\gamma, \eta$ 满足 $\gamma=\alpha\beta^2\eta^2 m/L^2, \eta=\min\{1,3(1-2\alpha)\}m^2/M$。

下面只分析 Pure phase。**定理 1** 如果 f 满足 $\nabla^2 f$ 满足 $M-Lipschitz$，而且 f 是 $m-strongconvex$ 的。当 $\left\|\nabla f\left(x^{(k)}\right)\right\|_2<\eta$ 且步长 $t=1$，那么

$$\frac{M}{2m^2}\|\nabla f(x^\star)\|_2 \leq \left(\frac{M}{2m^2}\|\nabla f(x)\|_2\right)^2 \tag{60}$$

这个公式告诉我们，当 $\left\|\nabla f\left(x^{(k)}\right)\right\|_2<\eta$ 且步长 $t=1$，$\left\|\nabla f\left(x^{(k)}\right)\right\|_2$ 将不会再变大。证明:

$$\begin{aligned}
\|\nabla f(x^+)\|_2 &= \|\nabla f(x+v)\|_2 \text{ where } v=-\left(\nabla^2 f(x)\right)^{-1}\nabla f(x)\\
&= \left\|\nabla f(x+v)-\nabla f(x)-\nabla^2 f(x)v\right\|_2\\
&= \left\|\int_0^1 \nabla^2 f(x+tv)vdt-\nabla^2 f(x)v\right\|_2\\
&= \left\|\int_0^1 \left(\nabla^2 f(x+tv)-\nabla^2 f(x)\right)vdt\right\|_2\\
&\leq \int_0^1 \underbrace{\left\|\left(\nabla^2 f(x+tv)-\nabla^2 f(x)\right)v\right\|_2}_{\leqslant\|\nabla^2 f(x+tv)-\nabla^2 f(x)\|_{op}\cdot\|v\|_2\leq Mt\|v\|_2^2} dt\\
&\leq\leqslant M\|v\|_2^2\int_0^1 tdt\\
&= \frac{1}{2}M\|-\left(\nabla^2 f(x)\right)^{-1}\nabla f(x)\|\\
&\leq \frac{1}{2}M\|-\left(\nabla^2 f(x)\right)^{-1}\|_{op}^2\|\nabla f(x)\|_2^2\\
&= \frac{M}{2m^2}\|\nabla f(x)\|_2^2
\end{aligned} \tag{61}$$

对最后的式子两边同时乘以 $\frac{M}{2m^2}$ 后得证。**定理 2** 如果 f 满足 $\nabla^2 f$ 满足 $M-Lipschitz$，而且 f 是 $m-strongconvex$ 的，那么可得:

$$f\left(x^{(k)}\right)-f^\star \leq \frac{2m^3}{M^2}\left(\frac{1}{2}\right)^{2^{k-k_0+1}} \tag{62}$$

证明: 我们令 $a_k=\frac{M}{2m^2}\|\nabla f(x^\star)\|_2$，根据定理 1，可以得到:

$$a_k \leq a_{k-1}^2 \tag{63}$$

迭代此式子，可得 $a_k \leq a_{k_0}^{2^{k-k_0}}$，也就是:

$$\frac{M}{2m^2}\|\nabla f(x^k)\|_2 \leq \left(\frac{M}{2m^2}\|\nabla f(x_{k_0})\|_2\right)^{2^{k-k_0}} \tag{64}$$

我们知道 $\|\nabla f(x_{k_0})\|_2 \leq \eta$ ，我们也知道 $\eta \leq \frac{m^2}{M}$(尚未证明)，因此:

$$\frac{M}{2m^2}\|\nabla f(x^k)\|_2 \leq \left(\frac{1}{2}\right)^{2^{k-k_0}} \tag{65}$$

根据公式8,

$$
\begin{aligned}
f\left(x^{(h)}\right)-f^{*} & \leqslant \frac{1}{2 m}\left\|\nabla f\left(x^{(L)}\right)\right\|_{2}^{2} \\
& \leqslant \frac{1}{2 m}\left(\frac{2 m^{2}}{M}\right)^{2}\left(\frac{1}{2}\right)^{2^{k-k_{0}+1}} \\
& =\frac{2 m^{3}}{m^{2}}\left(\frac{1}{2}\right)^{2 k-k_{0}+1}
\end{aligned} \tag{66}
$$

进一步，令 $c\cdot\left(\frac{1}{2}\right)^{2^{k-k_0+1}}=\varepsilon$，可以得到 $k-k_0=\log\log\left(\varepsilon_0/c\right)$。boyd 书上说 $\log\log\left(1/c\right)=6$ 实际上是个常数。

注意: 虽然这证明了牛顿法是二次收敛的，但是这是一种局部收敛速度 (只在接近 $f*$ 时的收敛速度)，实际上最差的情况下，牛顿法可能变成线性收敛的 (内点法能够保证总是二次收敛，暂不讨论)。

3.2 内点法

3.2.1 Barrier method

3.2.2 Primal-dual interior-point methods

3.3 准牛顿法 Quasi-Newton

牛顿法的主要缺点：求解 Hessian 矩阵计算复杂度和空间复杂度非常大

3.3.1 基本思想

准牛顿法的基本思想：用矩阵 B 估计 Hessian 矩阵。准牛顿法的一般步骤：1. 求解

$$
B^{(k-1)}\Delta x^{(k-1)}=-\nabla f\left(x^{(k-1)}\right) \tag{67}
$$

。2. 梯度更新

$$
x^{(k)}=x^{(k-1)}+t_k\Delta x^{(k-1)} \tag{68}
$$

。3. 更新对 B 的估计，即从 $B^{(k-1)}$ 得到 $B^{(k)}$。

割线方程：利用第 k 和 $k-1$ 步的梯度信息和上一次 B 的估计来更新 B 估计

$$
\nabla f\left(x^{+}\right)=\nabla f(x)+B^{+}\left(x^{+}-x\right) \tag{69}
$$

简写为 $B^{+}s=y$。用第 k 和 $k-1$ 步的两个梯度点形成的直线的斜率作为对 B 的估计 (见 note)。

3.3.2 SR1

SR1(symmetric rank-one) 是指对称秩 1 更新，即保证 B 的秩始终为 1 且对称。假设 B^{+} 与 B 具有以下关系:

$$
B^{+}=B+auu^{T} \tag{70}
$$

根据割线方程，我们可以得到:

$$\left(au^{T}s\right)u=y-Bs \tag{71}$$

求解此方程时，我们可以令人为地令 $u=y-Bs$。从而 $au^{T}s=1$, 求解可得 $a=1/(y-Bs)^{T}s$。从而对 B 的更新公式为:

$$B^{+}=B+\frac{(y-Bs)(y-Bs)^{T}}{(y-Bs)^{T}s} \tag{72}$$

在牛顿法中，得到了 Hessian 矩阵之后，求解 Hessian 矩阵的逆，实际上根据 Sherman-Morrison 公式 (公式1), 我们可以直接估计 $C=B^{-1}$ 而不估计 B:

$$C^{+}=C+\frac{(s-Cy)(s-Cy)^{T}}{(s-Cy)^{T}y} \tag{73}$$

证明:

$$\begin{aligned}\left(B+auu^{T}\right)^{-1}&=B^{-1}-\frac{B^{-1}auu^{T}B^{-1}}{1+u^{T}B^{-1}au}\\ C^{+}&=C-\frac{Cauu^{T}C}{1+u^{T}Cau}\\ &=C-\frac{Cuu^{T}C}{1/a+u^{T}Cu}\\ &=C-\frac{Cuu^{T}C}{u^{T}s+u^{T}Cu}\end{aligned} \tag{74}$$

注意:B 是对称矩阵，所有 $B=B^{T}$，其逆 $C=C^{T}$。

$$\begin{aligned}C^{+}&=C-\frac{Cuu^{T}C^{T}}{u^{T}s+u^{T}Cu}\\ &=C-\frac{C(y-Bs)(y-Bs)^{T}C^{T}}{u^{T}(s+Cu)}\end{aligned} \tag{75}$$

将 u 展开之后便可得到证。SR1 缺点：不能保证 B 是正定的。

3.3.3 BFGS

对称 2 秩估计:

$$B^{+}=B+auu^{T}+bvv^{T} \tag{76}$$

根据割线方程，我们可以得到:

$$y-Bs=\left(au^{T}s\right)u+\left(bv^{T}s\right)v \tag{77}$$

令 $u=y,v=Bs$, 可以得到 $au^{T}s=1,bv^{T}s=-1$, 解得 $a=\frac{1}{u^{T}s},b=-\frac{1}{v^{T}s}$，带入公式 (76) 可得:

$$B^{+}=B-\frac{Bss^{T}B}{s^{T}Bs}+\frac{yy^{T}}{y^{T}s} \tag{78}$$

与 SR1 算法类似，BFGS 算法也可以直接估计 B^{-1}, 这里要借用 Woodbury 公式 (公式3), 在我们的例子中:

$$U=V^{T}=\left[\begin{array}{ll}Bs & y\end{array}\right],\quad D=\left[\begin{array}{cc}-1/\left(s^{T}Bs\right) & 0\\ 0 & 1/\left(y^{T}s\right)\end{array}\right] \tag{79}$$

带入后，进行整理，可得:

$$C^+=\left(I-\frac{sy^T}{y^Ts}\right)C\left(I-\frac{ys^T}{y^Ts}\right)+\frac{ss^T}{y^Ts} \tag{80}$$

BFGS 能够保证 C 始终是正定的, 假设 $y^Ts=(\nabla f(x^+)-\nabla f(x))^T(x^+-x)>0$(也就是满足强凸性) 并且 $C\succ 0$，那么

$$x^TC^+x=\left(x-\frac{s^Tx}{y^Ts}y\right)^TC\left(x-\frac{s^Tx}{y^Ts}y\right)+\frac{\left(s^Tx\right)^2}{y^Ts} \tag{81}$$

始终大于等于零。

3.3.4 Davidon-Fletcher-Powell(DFP)

我们可以绕过 Woodbury 公式更新 Hessian 的逆:

$$C^+=C+auu^T+bvv^T \tag{82}$$

将割线方程写为 $s=C^+y$，令 $u=s,v=Cy$，可以得到:

$$C^+=C-\frac{Cyy^TC}{y^TCy}+\frac{ss^T}{y^Ts} \tag{83}$$

3.3.5 Limited memory BFGS

BFGS 和 DFP 都需要保存整个 Hessian 或者 Hessian inverse 矩阵，对于规模比较大的问题，这非常消耗内存。实际上，在梯度更新 (68) 步骤, 我们并不关注 Hessian 或者 Hessian inverse, 而是希望得到 $-(B^k)^{-1}\nabla f(x^k)$。也就是说，我们在实际运算过程中，并不真的需要 C^+ 矩阵的内容，而是希望得到 C^+g, 其中 g 是当前点梯度。公式80两边同时乘以 g, 我们可以得到 Implicit BFGS:

$$\begin{aligned}
cC^+g&=\left(I-\frac{sy^T}{y^Ts}\right)C\left(I-\frac{ys^T}{y^Ts}\right)g+\frac{ss^T}{y^Ts}g\\
&=\left(I-\frac{sy^T}{y^Ts}\right)C\left(g-\frac{ys^T}{y^Ts}g\right)+\frac{s^Tg}{y^Ts}s\\
&=\left(I-\frac{sy^T}{y^Ts}\right)\underbrace{C\underbrace{\left(g-\frac{s^Tg}{y^Ts}y\right)}_{q}}_{p}+\underbrace{\frac{s^Tg}{y^Ts}}_{\alpha}s\\
&=\left(I-\frac{sy^T}{y^Ts}\right)p+\alpha s\\
&=p-\underbrace{\frac{y^Tp}{y^Ts}}_{\beta}s+\alpha s\\
&=p+(\alpha-\beta)s
\end{aligned} \tag{84}$$

也就是:

$$C^+g=p+(\alpha-\beta)s \tag{85}$$

其中:

$$\alpha=\frac{s^Tg}{y^Ts}, q=g-\alpha y, p=Cq, \beta=\frac{y^Tp}{y^Ts} \tag{86}$$

理解 Implicit BFGS: 假设我们想要求 C^kg^k)，那么需要计算

$$\begin{aligned} \alpha^k &= \tfrac{(s^k)^Tg^k}{(y^k)^T(s^k)^T} \\ q^k &= g^k-\alpha^ky^k \\ p^k &= C^{k-1}q^k \end{aligned} \tag{87}$$

由于 C^{k-1} 未知，我们无法直接求 p^k, 进而无法求 β^k。但是 $p^k=C^{k-1}q^k$ 意味着我们可以迭代求解。也就是说，我们运用公式85求解 $p^k=C^{k-1}q^k$:

$$C^{k-1}q^k=p^{k-1}+(\alpha^{k-1}-\beta^{k-1})s^{k-1} \tag{88}$$

其中，

$$\begin{aligned} \alpha^{k-1} &= \tfrac{(s^{k-1})^Tq^k}{(y^k)^T(s^k)^T} \\ q^{k-1} &= q^k-\alpha^ky^k \\ p^{k-1} &= C^{k-2}q^{k-1} \end{aligned} \tag{89}$$

同样的，由于 C^{k-2} 未知，我们无法直接求 p^{k-1}, 进而无法求 β^{k-1}, 使用与上面同样的递归方法，我们可以得到 Implict BFGS 算法 [1]3.3.5: 第一个 for 循环一直在向下递归，

Algorithm 1 Implict BFGS

1: Let $q=-\nabla f\left(x^{(k)}\right)$
2: **for** $i=k-1,\ldots,0$ **do**
3: Compute $\alpha_i=\left(s^{(i)}\right)^Tq/\left(\left(y^{(i)}\right)^Ts^{(i)}\right)$
4: Update $q=q-\alpha y^{(i)}$
5: **end for**
6: Let $p=C^{(0)}q$
7: **for** $i=0,\ldots,k-1$ **do**
8: Compute $\beta=\left(y^{(i)}\right)^Tp/\left(\left(y^{(i)}\right)^Ts^{(i)}\right)$
9: Update $p=p+(\alpha_i-\beta)\,s^{(i)}$
10: **end for**
11: Return p

第二个 for 循环向上递归，最后的 p 为 C^kg^k) 运行此算法 k 次 (k 次牛顿更新) 的空间复杂度 Okn，时间复杂度 $O(k^2n)$,k 是总的迭代次数。这个算法只有当 k 小于 n 的时候才有提升，实际上，k 大于 n 经常出现。

LBFGS 算法是对 Implict BFGS 的改动3.3.5: 也就是说，在计算 C^kg^k) 时，我并不使用从 $0,\ldots,k-1$ 所有信息，而只使用 $k-m,\ldots,k-1$ 这 m 个信息。在第 6 行中，$\bar{C}^{(k-m)}$ 通常通过估计得来，一般设置为 I 的倍数，例如 [2]:

$$C^{0,k}:=\frac{\left(y^{k-1}\right)^\top s^{k-1}}{\left(y^{k-1}\right)^\top y^{k-1}}I \tag{90}$$

LBFGS 算法运行 k 次 (k 次牛顿更新) 的空间复杂度 Omn，时间复杂度 $O(kmn)$。

Algorithm 2 Limited memory BFGS

Let $q=-\nabla f\left(x^{(k)}\right)$
for $i=k-1,\ldots,k-m$ **do**
  Compute $\alpha_i=\left(s^{(i)}\right)^T q/\left(\left(y^{(i)}\right)^T s^{(i)}\right)$
  Update $q=q-\alpha y^{(i)}$
end for
Let $p=\bar{C}^{(k-m)}q$
for $i=k-m,\ldots,k-1$ **do**
  Compute $\beta=\left(y^{(i)}\right)^T p/\left(\left(y^{(i)}\right)^T s^{(i)}\right)$
  Update $p=p+\left(\alpha_i-\beta\right)s^{(i)}$
end for
Return p

3.3.6 BFGS 收敛性分析

暂时略

4 Advanced Topic

4.1 Coordinate Descent

定理 1 对于一个凸可微的函数 $f(x)$，如果 $x^\star$ 沿着每个轴都最小化 $f(x)$，那么 $x^\star$ 是全局最小点 (充要条件)。证明：如果 $x^\star$ 沿着每个坐标轴都最小化 $f(x)$，这意味着 $\left(\frac{\partial f}{\partial x_1}(x),\ldots\frac{\partial f}{\partial x_n}(x)\right)=0=\nabla f(x)$, 即满足凸可微的一阶最优条件。

对于定理 1, 如果 $f(x)$ 不是可微的，那么结果不成立，如图2, 在红色点位置，沿着任意方向都是最小值，但是并非全局最小值。

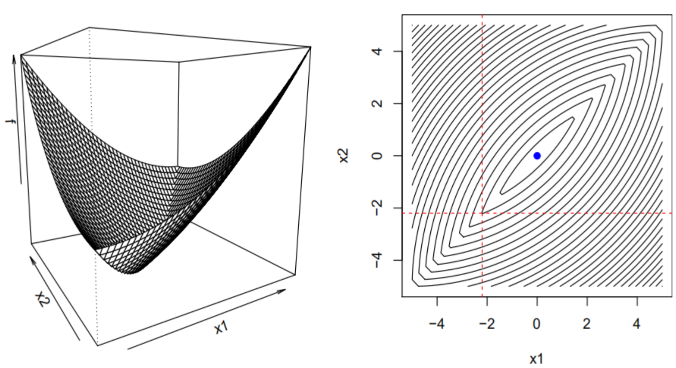

图 1: 凸不可微函数

定理 2 假设有一个凸函数 $f(x)=g(x)+\sum_{i=1}^{n}h_i(x_i)$(不一定凸，也不一定可微)，其中，$g(x)$ 凸可微，$h_i(x_i)$ 凸。如果 $x^\star$ 沿着每个轴都最小化 $f(x)$，那么 $x^\star$ 是全局最小点。

证明: 由于 $f(x)$ 是凸函数，那么我们只需要证明对于任意的 y, 都有 $f(y)-f(x^\star)>0$ 即可。即证明:

$$
\begin{aligned}
f(y)-f(x^\star) &\geq \nabla g(x^\star)^T(y-x^\star)+\sum_{i=1}^{n}\left[h_i\left(y_i\right)-h_i\left(x_i^\star\right)\right] \\
&=\sum_{i=1}^{n}\underbrace{\left[\nabla_i g(x^\star)\left(y_i-x_i^\star\right)+h_i\left(y_i\right)-h_i\left(x_i^\star\right)\right]}_{\geq 0} \geq 0
\end{aligned}
\tag{91}
$$

由于 $x^\star$ 是最小点，所以

$$
\begin{aligned}
0 \in \nabla_i g(x^\star)+\partial h_i(x_i^\star) \\
-\nabla_i g(x^\star) \in \partial h_i(x_i^\star)
\end{aligned}
\tag{92}
$$

由凸函数的一阶性质可得:

$$
\begin{aligned}
h_i\left(y_i\right) \geq h_i\left(x_i^\star\right)-\nabla_i g(x^\star)\left(y_i-x_i^\star\right) \\
h_i\left(y_i\right)-h_i\left(x_i^\star\right)+\nabla_i g(x^\star)\left(y_i-x_i^\star\right) \geq 0
\end{aligned}
\tag{93}
$$

坐标下降法:

$$
\begin{aligned}
x_i^{(k)}=\underset{x_i}{\operatorname{argmin}} f\left(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)}\right), \\
i=1, \ldots, n
\end{aligned}
\tag{94}
$$

4.2 共轭上升法

4.2.1 基本思想

考虑下面问题:

$$
\begin{aligned}
\min_x f(x) \\
\text{subject to } Ax=b
\end{aligned}
\tag{95}
$$

其对偶问题是

$$
\max_x -f^*\left(-A^T u\right)-b^T u \tag{96}
$$

我们使用子梯度法来求解对偶问题 (变量为对偶变量)。定义 $g(u)=-f^*\left(-A^T u\right)-b^T u$。根据求导链式法则，可以得到 $\partial g(u)=A \partial f^*\left(-A^T u\right)-b$，使用共轭函数性质 212, 我们可以得到

$$
\begin{aligned}
\partial g(u)=Ax-b \\
\text{where} \quad x \in \underset{z}{\operatorname{argmin}} f(z)+u^T A z
\end{aligned}
\tag{97}
$$

那么子梯度法可以写为:

$$
\begin{aligned}
x^{(k)} &\in \underset{x}{\operatorname{argmin}} f(x)+\left(u^{(k-1)}\right)^T A x \\
u^{(k)} &= u^{(k-1)}+t_k\left(Ax^{(k)}-b\right)
\end{aligned}
\tag{98}
$$

4.2.2 共轭分解

对于如下问题:

$$\begin{array}{c}\min_x \sum_{i=1}^{B} f_i\left(x_i\right) \\ \text{subject to } Ax=b \quad \text{where} A=\left[A_1 \ldots A_B\right]\end{array} \tag{99}$$

那么我们可以分布式地更新对偶变量:

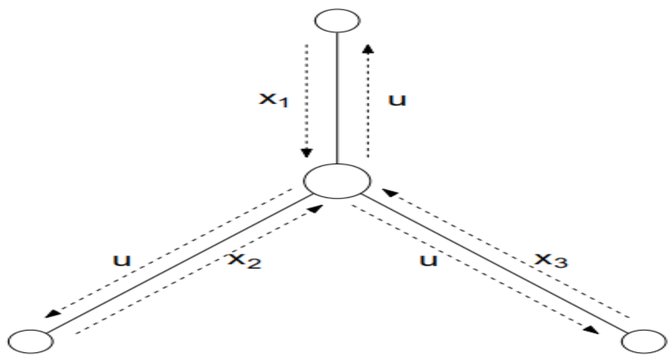

图 2: 共轭分解法分布式计算

整个共轭分解法分成两步:1) 广播: 中心节点将对偶变量 u 传递给各个节点;2) 收集: 各个节点计算后返回到中心节点。

4.2.3 收敛速率分析

在进行收敛性分析时，对偶上升法要求具有强凸性。

4.2.4 增广拉格朗日法 (乘子法)

对偶上升法要求目标函数具有强凸性，一般通过乘子法可以达到此目的, 将原问题95转化为:

$$\begin{array}{ll}\min\limits_x & f(x)+\frac{\rho}{2}\|Ax-b\|_2^2 \\ \text{subject to} & Ax=b\end{array} \tag{100}$$

对应的梯度上升法为:

$$\begin{aligned} x^{(k)} &= \underset{x}{\operatorname{argmin}} f(x)+\left(u^{(k-1)}\right)^T Ax+\frac{\rho}{2}\|Ax-b\|_2^2 \\ u^{(k)} &= u^{(k-1)}+\rho\left(Ax^{(k)}-b\right)\end{aligned} \tag{101}$$

注意，为什么步长选取 ρ? 观察问题100, 其 KKT 条件中静态条件是

$$\begin{aligned} 0 &\in \partial f\left(x^{(k)}\right)+A^T\left(u^{(k-1)}+\rho\left(Ax^{(k)}-b\right)\right) \\ &= \partial f\left(x^{(k)}\right)+A^T u^{(k)}\end{aligned} \tag{102}$$

这正是原问题95的静态条件。这意味着我们用对偶上升法求解增广拉格朗日问题的时候，当 $k \rightarrow \infty$，对偶问题等价原问题，对偶间隙为零。

乘子法的缺点: 使其可分解性，意味着难以分布式地计算。

4.2.5 交换方向乘子法 (ADMM)

交换方向乘子法保留可分解性。考虑如下问题:

$$\begin{array}{ll} \min\limits_{x,z} & f(x)+g(z) \\ \text{subject to} & Ax+Bz=c \end{array} \tag{103}$$

其增广拉格朗日形式:

$$\begin{array}{ll} \min\limits_{x} & f(x)+g(z)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \\ \text{subject to} & Ax+Bz=c \end{array} \tag{104}$$

对偶增广拉格朗日函数:

$$L_\rho(x,z,u)=f(x)+g(z)+u^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \tag{105}$$

使用对偶上升法和坐标下降法:

$$\begin{aligned} x^{(k)} &= \underset{x}{\operatorname{argmin}} L_\rho\left(x, z^{(k-1)}, u^{(k-1)}\right) \\ z^{(k)} &= \underset{z}{\operatorname{argmin}} L_\rho\left(x^{(k)}, z, u^{(k-1)}\right) \\ u^{(k)} &= u^{(k-1)}+\rho\left(Ax^{(k)}+Bz^{(k)}-c\right) \end{aligned} \tag{106}$$

4.2.6 尺度化的 ADMM

令 $w=u/\rho$,

$$\begin{array}{ll} L_\rho(x,z,u) & =f(x)+g(z)+u^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \\ L_\rho(x,z,u) & =f(x)+g(z)+\rho w^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \\ L_\rho(x,z,u) & =f(x)+g(z)+\rho w^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2+\frac{\rho}{2}\|w\|_2^2-\frac{\rho}{2}\|w\|_2^2 \\ L_\rho(x,z,w) & =f(x)+g(z)+\frac{\rho}{2}\|Ax-Bx+c+w\|_2^2-\frac{\rho}{2}\|w\|_2^2 \end{array} \tag{107}$$

那么求解对偶问题的迭代公式为:

$$\begin{aligned} x^{(k)} &= \underset{x}{\operatorname{argmin}} f(x)+\frac{\rho}{2}\left\|Ax+Bz^{(k-1)}-c+w^{(k-1)}\right\|_2^2 \\ z^{(k)} &= \underset{z}{\operatorname{argmin}} g(z)+\frac{\rho}{2}\left\|Ax^{(k)}+Bz-c+w^{(k-1)}\right\|_2^2 \\ w^{(k)} &= w^{(k-1)}+Ax^{(k)}+Bz^{(k)}-c \end{aligned} \tag{108}$$

4.2.7 consensus ADMM

4.3 动量法

RMSProp 和指数动量法的初衷有一部分是对步长鲁棒。

我们可以将这一部分分成两部分: 对步长的修改 (RMSprop) 和利用 momentum 梯度 v_t 代替当前梯度 g_t(指数加权平均)。

[8] 中，AdaGrad 和 AdaDelta 算法是对步长进行修改的两个变种，这里不考虑这两个算法。

4.3.1 指数加权平均

如图3，以温度预测为例：假设现在有 n 天的温度数据 $\{\theta_1 \dots \theta_n\}$(图中蓝色点)，这个数据波动比较大，我们想要得到一个变化比较平缓的序列 (图中红线)。

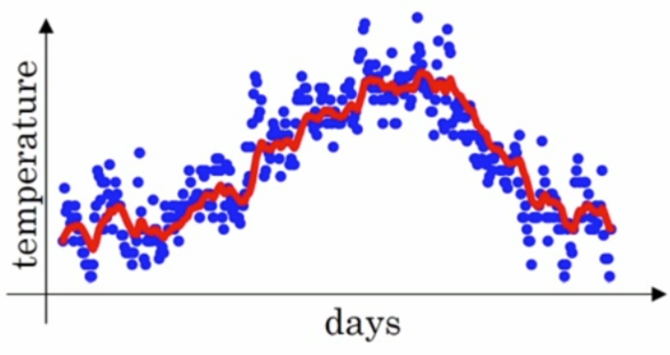

图 3: 指数加权平均平滑温度曲线

指数加权平均递归公式:$v_t = \beta v_{k-1} + (1-\beta)\theta_t$。展开之后:

$$
\begin{aligned}
v_t &= \beta(\beta v_{k-2} + (1-\beta)\theta_{k-1}) + (1-\beta)\theta_t \\
&= \beta(\beta(\beta v_{k-3} + (1-\beta)\theta_{k-2}) + (1-\beta)\theta_{k-1}) + (1-\beta)\theta_t \\
&= \beta v_{k-3} + \beta\beta(1-\beta)\theta_{k-2} + \beta(1-\beta)\theta_{k-1} + (1-\beta)\theta_t \\
&= \beta^k(1-\beta)v_0 + \dots + \beta^2(1-\beta)\theta_{k-2} + \beta(1-\beta)\theta_{k-1} + (1-\beta)\theta_t \\
&= (\beta^k v_0 + \dots + \beta^2\theta_{k-2} + \beta\theta_{k-1} + \theta_t)(1-\beta)
\end{aligned} \tag{109}
$$

上式最后一行意味着 v_t 是前面温度的指数加权平均。**理解指数加权平均**：利用如下事实：当 $\epsilon \to 1$，那么 $(1-\varepsilon)^{1/\varepsilon} \approx \frac{1}{e} \approx 0.367$。假设我们忽略权重系数小于 $1/e$ 的项，那么相当于我们对过去 $1/\varepsilon$ 次的温度进行指数加权平均。例子 1: 如果选取 $\varepsilon = 0.1$ 也就是 $\beta = 0.9$，相当于 v_t 的值是过去 10 天的指数加权平均 (更精确地说，其实不止 10 天，但是我们忽略了权重系数小于 $1/e$ 的项)。例子 2: 如果选取 $\varepsilon = 0.01$ 也就是 $\beta = 0.99$，相当于 v_t 的值是过去 100 天的指数加权平均。**偏差纠正**：问题描述:Pure 指数加权平

Algorithm 3 Exponential average

Input: temperatures $\{\theta_1 \dots \theta_n\}$,$\beta$

1: $v_0 = 0$
2: **for** $i = 1 \to k$ **do**
3: Compute $v_t = \beta v_{k-1} + (1-\beta)\theta_t$
4: **end for**

Return v_t

均的问题主要是前几项。如图4所示。假设 $\beta = 0.98$，根据算法4.3.1, 我们会得到紫色的数列，可以看出，在早期帧是有偏差的。实际上，偏差估计后的数列是绿色线。

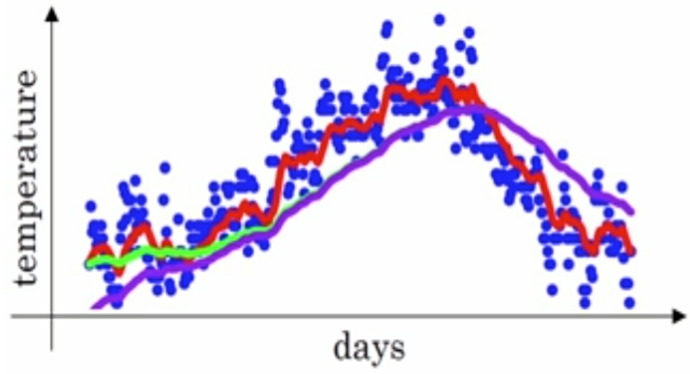

图 4: 指数加权平均平滑温度曲线

取消偏差估计方法：

$$
\begin{aligned}
v_t &= \beta v_{k-1} + (1-\beta)\theta_t \\
v_k &= \frac{v_k}{1-\beta^k}
\end{aligned}
\tag{110}
$$

下面的公式的意思是除以权重系数的总和 (可以通过等比数列求和公式得到 $\frac{1}{1-\beta^k}$ 这一项)。

4.3.2 使用指数加权平均的梯度下降法

问题描述: 如图5中紫色线所示，梯度下降法的问题是步长如果比较大，那么抖动比较大，而使用指数加权平均法可以降低抖动，使得求解过程更加稳定。

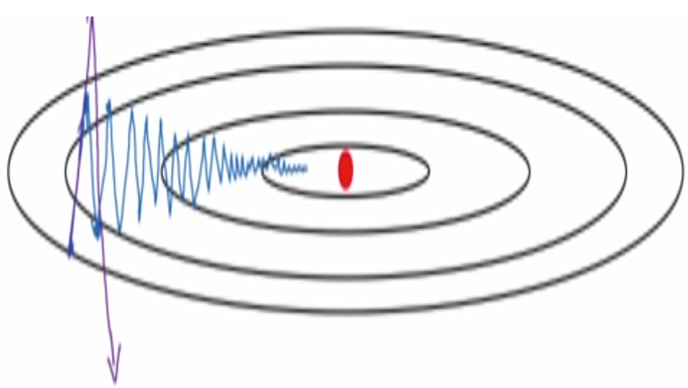

图 5: 使用指数加权平均的梯度下降法可以平稳梯度更新方向

指数加权平均的梯度下降法一次更新如算法4.3.2。另外注意，这里我们没有用偏差纠正，在实际中这个方法很少用偏差纠正 [7]。

Algorithm 4 Exponential average gradient descent

Input: gradient g_k after epoch k, average value from previous epoch v_{k-1} ,current position x_{k-1},β

1: compute $v_k = \beta v_{k-1} + (1-\beta) g_k$

2: compute $x_k = x_{k-1} - v_k$

Return x_k, v_k

4.3.3 RMSprop

基本想法：一种更新步长的策略，更新策略中使用到了指数加权平均。如图6。蓝色是梯度下降法，绿色是 RMSprop，可以看到 RMSprop 不仅抑制垂直的抖动，而且增加朝着中心的步长。(效果有点类似把弹簧拉伸，纵向变窄，横向变长)

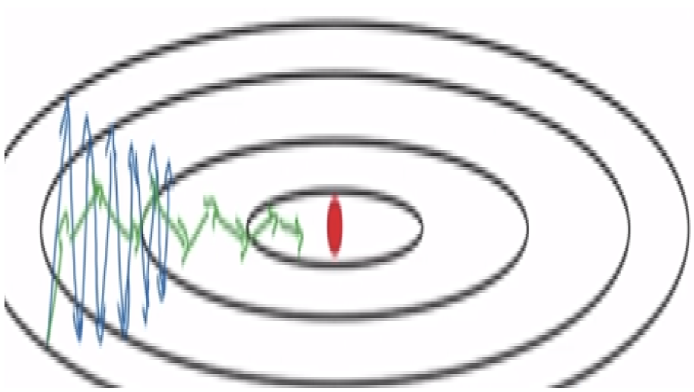

图 6: RMSprop 更快地朝着最优点前进

算法如4.3.3所示。

Algorithm 5 Rmsprop

Input: gradient g_k after epoch k, average value from previous epoch s_{k-1} ,current position x_{k-1},β

1: compute $s_k = \beta s_{k-1} + (1-\beta) g_k \odot g_k, where \odot denote \quad hadamard \quad product$

2: compute $x_k = x_{k-1} - \frac{t}{\sqrt{s_k + \epsilon}} g_k$

Return x_k, s_k

4.3.4 Adam

Adam 其实是指数加权平均梯度和 RMSprop 步长更新法的结合。不过，指数加权梯度更新和指数加权步长更新都进行了偏差纠正。

Algorithm 6 Adam

Input: gradient g_k after epoch k, average value from previous epoch s_{k-1}, v_{k-1} ,current position x_{k-1},β_1, β_2

1: compute $s_k = \beta_1 s_{k-1} + (1-\beta_1) g_k \odot g_k$

2: compute $v_k = \beta_2 v_{k-1} + (1-\beta_2) g_k$

3: compute $s_k^{corrected} = \frac{s_k}{1-\beta_1^k}$

4: compute $v_k^{corrected} = \frac{v_k}{1-\beta_2^k}$

5: compute $x_k = x_{k-1} - \frac{t}{\sqrt{s_k + \epsilon}} v_k$

Return x_k, s_k, v_k

5 总结

下面总结什么情况下应该选择什么算法

5.1 无约束

5.1.1 优化目标不可微

一阶方法: 子梯度法

5.1.2 优化目标可微

一阶方法: 梯度下降法、梯度下降法-backline tracking

5.1.3 优化目标部分可微、部分不可微

一阶方法: 子梯度法、近端梯度法

5.2 只有等式约束 (线性等式约束)

一阶方法: 二阶方法:1) 转化为无约束;2) 利用等式约束的牛顿法，使用 KKT 条件求解。

5.3 不等式约束 (含等式约束)

通用方法: 投影梯度法，即计算梯度更新之后，投影到可行集中。二阶方法:barrier 方法

参考文献

[1] http://www.stat.cmu.edu/ ryantibs/convexopt-F18/

[2] http://www.stat.cmu.edu/ ryantibs/convexopt-F16/

[3] https://www.zhihu.com/question/296828990/answer/502071622

[4] Nemirosvki et al. (2009). Robust stochastic optimization approach to stochastic programming

[5] https://blog.csdn.net/jclian91/article/details/80254568

[6] https://blog.csdn.net/zhangping1987/article/details/24365455

[7] https://mooc.study.163.com/smartSpec/detail/1001319001.htm?

[8] http://zh.gluon.ai/index.html