Pure Mathematics 理论数学, 2018, 8(3), 221-229
Published Online May 2018 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2018.83028

Hans汉斯

A Stochastic Sub-Gradient Mirror Descent Algorithm for Non-Smooth and Strongly Convex Functions

Qian Zhou, Xianbing Luo*, Xin Wang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
Email: *luoxb121@163.com

Received: Apr. 27th, 2018; accepted: May 11th, 2018; published: May 18th, 2018

Abstract

Mirror descent (MD) has been widely used to deal with the machine learning problems. For large scale data processing and non-smooth loss convex optimization problem, we proposed an improved mirror descent method, which is called modified stochastic sub-gradient mirror descent method. It combined an iterative average method with stochastic sub-gradient descent method. In the process of weighted average, the average iteration is not used to construct the algorithm, but occurs as a byproduct of our algorithm. The average weight is determined by the step size used by the algorithm. Our algorithm has good convergence. For strong convex functions, we show that the optimal convergence rate of the algorithm arrives at $o\left(\frac{1}{k}\right)$.

Keywords

Mirror Descent Method, Non-Smooth Loss Convex Optimization, Stochastic Sub-Gradient Mirror Descent Method, Iterative Weighted Average

一种非光滑强凸函数的随机次梯度镜面下降算法

周　倩，罗贤兵*，王　鑫

贵州大学数学与统计学院，贵州　贵阳
Email: *luoxb121@163.com

收稿日期：2018年4月27日；录用日期：2018年5月11日；发布日期：2018年5月18日

*通讯作者。

文章引用：周倩，罗贤兵，王鑫．一种非光滑强凸函数的随机次梯度镜面下降算法[J]．理论数学, 2018, 8(3): 221-229.
DOI: 10.12677/pm.2018.83028

摘　要

镜面下降法(MD)在机器学习问题中已有些实际应用，针对大规模数据的处理和非光滑损失凸优化问题，本文将迭代平均与随机次梯度镜面下降方法相结合，得到了一种改进的方法，通过对问题域的特殊处理，利用它们的结构，提出一种加权平均的随机次梯度镜面下降算法。在这个加权平均过程中，平均迭代不用于构造算法，而是作为算法的副产品出现，其中平均权重由算法使用的步长确定。该算法有很好的收敛性。对于强凸函数，我们证明了该算法的最佳收敛速度达到$o\left(\frac{1}{k}\right)$。

关键词

镜面下降法，非光滑损失凸优化，随机次梯度镜面下降法，迭代加权平均

1. 引言

随着机器学习的火热发展，机器学习中的大规模优化问题成为近年来被广泛研究的热门问题之一，大部分的机器学习问题的本质都是建立优化模型，通过最优化方法对目标函数(或损失函数)进行优化，从而训练出最好的模型。其主要模式是通过优化经验损失函数，模拟出最优参数，从而获得损失最小化模型。正则化损失最小化问题是机器学习问题的常见范例，在当今机器学习的应用和研究中占主导地，强凸不可微优化问题作为正则化随机学习问题在机器学习中最为突出。大规模非光滑凸优化是机器学习和计算机虚拟等计算领域中的一个常见问题。这些领域的问题包括特殊的领域结构和特征。对这些问题域的特殊处理，利用它们的结构，可以大大提高计算效率。

众所周知，在低迭代次数下，凸优化问题可以用内点法在多对数时间内求解。然而，这些方法中的大多数都不能很好地适应优化问题的维数。内点法的单次迭代代价随问题的大小呈非线性增长。一阶方法计算上廉价的迭代成为大规模优化问题的一个可行的选择。本文提出了一种求解凸集笛卡儿乘积上一般大规模非光滑强凸优化问题的一阶自适应方法[1]。

我们主要关注于求解一个形式如下的优化问题：考虑约束集Ω上的凸函数 F 的最小化问题(其中 F 不一定是可微函数)：

$$\min_{w\in\Omega} F(w) = E\left[f(w,\xi)\right]. \tag{1}$$

设约束集Ω是闭凸集，函数F是凸的，且在任意点$w\in\Omega$是连续的。此外，对任意的$w\in\Omega$，函数$F(w)$的次梯度$g(w)$存在，即，对任意的$w\in\Omega$，存在向量$g(w)$，使得：

$$F(w)+\langle g(w),u-w\rangle \le F(u)，\text{对任意的}\ u\in\Omega$$

2. 基本的镜面下降算法

镜面下降算法[2] [3]是投影次梯度法[4]的推广。标准次梯度法在投影步长上采用欧氏距离函数，选

用适当步长。镜面下降算法在非线性投影步长中，采用具有最佳步长的非线性距离函数，对标准投影次梯度法进行扩展。在这一部分中，我们回顾一个基本的镜面下降算法来解决问题(1)，不考虑区域几何。

设 $B(\bullet,\bullet)$ 表示集合 Ω 中任意两点的 Bregman 距离函数，一个基本的镜面下降算法使用了一个非线性投影序列：

$$w_{k+1}=\arg\min_{w\in\Omega}\left\{\langle g_k,w\rangle+\frac{1}{\alpha_k}B_\varphi(w,w_k)\right\}.$$

其中，g_k 是函数 F 在点 w_k 处的次梯度，$\alpha_k>0$ 为迭代步长。

MD 方法要求距离函数的范数满足如下性质[5] [6]：

1) 空间 E 上的范数 $\|\bullet\|$ 嵌入于 Ω，且 E^* 上的对偶范数 $\|\bullet\|_*$ 满足：

$$\|\xi\|_*=\max_w\{\langle\xi,w\rangle:\|w\|\le 1\};$$

2) 集合 Ω 上的距离生成函数及范数 $\|\bullet\|$，即有连续凸函数 $\varphi(w):\Omega\to R$ 使得：

- 存在 $\varphi(\bullet)$ 的一个次梯度 $\varphi'(\bullet)$，在集合 $\Omega^0=\{w\in\Omega:\partial\varphi(w)\ne\varnothing\}$ 上连续；
- $\varphi(\bullet)$ 关于 1 范数 $\|\bullet\|$ 是强凸的：

$$\forall(w,w'\in\Omega^0):\langle\varphi(w)-\varphi(w'),w-w'\rangle\ge\|w-u\|^2.$$

设最大距离为 $V=\max_{u\in\Omega}B(w_c,u)$，假设 $F(w)$ 在集合 Ω 上是 Lipschitz 连续的，Lipschitz 常数为 $L=\max_{w\in\Omega}\|F'_w\|_*<\infty$，MD 方法有如下收敛性：

定理 1：设 F^* 记为目标函数的全局最优解，$\bar{w}=\arg\min_{w=\{w_1,w_2,\cdots,w_k\}}F(w)$。选取最优步长为 $\alpha=\frac{\sqrt{2V}}{L\sqrt{k}}$，则 k 次迭代的最优界为：

$$F^*-F(\bar{w})\le\frac{L\sqrt{2V}}{\sqrt{k}}.$$

上述定理的证明参考[1]。

3. 随机镜面下降算法

在这一节中，我们假设问题(1)的目标函数 F 具有强凸性，且设其强凸系数为 μ_F，则有：

$$F(u)\ge F(w)+\langle g(w),u-w\rangle+\frac{\mu_F}{2}\|u-w\|^2,\ \text{对任意}\ u,w\in\Omega$$

接下来，我们考虑求解问题(1)的随机次梯度镜面下降算法。首先给出 Bregman 距离函数[7]的定义如下：

设 $\varphi(\cdot)$ 为集合 Ω 上连续可微的强凸函数，即存在 $\mu_\varphi>0$，使得

$$\varphi(v)\ge\varphi(w)+\langle\nabla\varphi(w),v-w\rangle+\frac{\mu_\varphi}{2}\|v-w\|^2,\ \text{对任意}\ w,v\in\Omega \tag{2}$$

则由函数 $\varphi(\cdot)$ 生成的 Bregman 距离函数记为 B_φ 定义为：

$$B_\varphi(w,u)=\varphi(u)-\varphi(w)-\langle\nabla\varphi(w),u-w\rangle,\ \text{对任意}\ u,w\in\Omega \tag{3}$$

从定义可以看出，Bregman 距离函数具有以下性质：

$$B_\varphi(w,u)-B_\varphi(v,u)=B_\varphi(w,v)+\langle\nabla\varphi(v)-\nabla\varphi(w),u-w\rangle,\ \text{对任意}\ u,v,w\in\Omega \tag{4}$$

$$B_\varphi(w,u)\geq\frac{\mu_\varphi}{2}\|w-u\|^2,\ \text{对任意}\ u,w\in\Omega \tag{5}$$

上述性质都可由函数 $\varphi(\cdot)$ 的强凸性质得到。又函数 $\varphi(\cdot)$ 连续可微，可知函数 $B_\varphi(w,u)$ 关于 u 可微，设 $\nabla_w B_\varphi(\cdot,\cdot)$ 表示 $B_\varphi(w,u)$ 关于 u 的偏导数，则有：

$$\nabla_w B_\varphi(w,u)=\nabla\varphi(u)-\nabla\varphi(w)。\ \text{对任意}\ u,w\in\Omega \tag{6}$$

设 $w_0\in\Omega$ 为初始点，则问题(1)的次梯度镜面下降法的迭代公式如下：

$$w_{k+1}=\arg\min_{w\in\Omega}\{\alpha_k\langle g_k,w-w_k\rangle+B_\varphi(w_k,w)\},\ \text{对任意}\ k>0$$

其中 $\alpha_k>0$ 为迭代步长，g_k 为函数 $F(w)$ 在点 $w=w_k$ 的次梯度。当集合 Ω 具有的结构能允许对 w_{k+1} 有效计算的前提下，该算法有效，例如，w_{k+1} 的闭形式是有效的。为了处理更一般的形式，我们采用随机思想，用目标函数次梯度的无偏估计代替其次梯度。即用任意选取的某一个损失函数 $f(w,\xi)$ 的次梯度 $\tilde{g}$ 代替目标函数 $F(w)$ 的次梯度，从而得到如下形式的随机次梯度镜面下降算法：

$$w_{k+1}=\arg\min_{w\in\Omega}\{\alpha_k\langle\tilde{g}_k,w-w_k\rangle+B_\varphi(w_k,w)\},\ \text{对任意}\ k>0 \tag{7}$$

其中任意选取初始点 $w_0\in\Omega$，满足 $E\left[\|w_0\|^2\right]\in\Omega<\infty$，但是与随机次梯度序列 $\{\tilde{g}_k\}$ 无关。

迭代步长 $\alpha_k\in(0,1]$，设其满足

$$\frac{1-\alpha_{k+1}}{\alpha_{k+1}^2}\leq\frac{1}{\alpha_k^2},\ \text{对任意}\ k\geq0,\ \alpha_0=1. \tag{8}$$

为了使上述假设看起来更有意义，我们讨论步长的选取问题。由设定的步长条件(8)，我们有：

$$0<\alpha_{k+1}\leq\frac{\sqrt{\alpha_k^4+4\alpha_k^2}-\alpha_k^2}{2},\ \text{对任意}\ k\geq0$$

初始化 $\alpha_0=1$。我们将考虑下面两种特殊的步长选择：

$$\alpha_k=\frac{1}{k+1}\ \text{和}\ \alpha_{k+1}=\frac{\sqrt{\alpha_k^4+4\alpha_k^2}-\alpha_k^2}{2},\ \text{对任意}\ k\geq0 \tag{9}$$

上述的第一个步长由 Tseng [8]提出，设第二个式子中 $\alpha_{k+1}=\frac{1}{t_{k+1}}$，则得到 Nesterov 序列 $\{t_{k+1}\}$ [9]，它用于构造梯度 Lipschitz 连续的凸函数的快速一阶算法，即

$$t_{k+1}=\frac{\sqrt{1+4t_k^2}+1}{2},\ \text{对任意}\ k\geq0$$

且 $t_0=1$ [10]。通过归纳，我们可以发现 $t_{k+1}=\frac{k+2}{2}$，对任意 $k\geq0$，从而推出(9)中的两个步长公式都满足：

$$0<\alpha_k\leq\frac{2}{k+2},\ \text{对任意}\ k\geq0,\ \text{且}\ \alpha_0=1$$

由于步长 α_k 满足(8)，则有

$$\alpha_k^2\geq1\Big/\left(\sum_{t=0}^{k}\frac{1}{\alpha_t}\right),\ \text{对任意}\ k\geq0.$$

我们可以建立基本的随机次梯度镜面下降算法如下：

a) 任意选取初始点 $w_0 \in \Omega$，满足 $E\left[\|w_0\|^2\right] \in \Omega < \infty$，一般我们选取 $w_0 = \arg\min_{w\in\Omega} \varphi(w)$。

b) $w_{k+1} = \arg\min_{w\in\Omega} \left\{\alpha_k \left\langle \tilde{g}_k, w - w_k \right\rangle + B_\varphi\left(w_k, w\right)\right\}$。

c) $\hat{w}_k = \left(\sum_{t=0}^{k} \frac{1}{\alpha_t}\right)^{-1} \sum_{t=0}^{k} \left(\frac{1}{\alpha_t} w_t\right)$。

这里，我们取 $\alpha_k = \frac{1}{k+1}$，且 $\alpha_0 = 1$。观察发现，上述迭代中，每次循环都需存储 w_k，加权平均值 $\hat{w}_{k-1}$，以及步长的相关和 $S = \sum_{t=0}^{k} \frac{1}{\alpha_t}$。

我们假设 $S_{k+1} = S_k + \frac{1}{\alpha_k}$，其中 $S_0 = 0$，这样我们可以将上述 c 改成：

$$\hat{w}_k = \frac{S_k}{S_{k+1}} \hat{w}_{k-1} + \left(1 - \frac{S_k}{S_{k+1}}\right) w_k .$$

这也就是我们说的加权平均迭代的随机次梯度镜面下降算法。

4. 算法的收敛性分析

接下来，我们对算法的收敛性进行理论分析。首先，讨论步长满足条件(8)的随机次梯度镜面下降法的迭代特点。这里我们假设 Bregman 距离函数满足如下要求：

$$B_\varphi(w,u) \le \frac{1}{2}\|w-u\|^2, \text{ 对任意 } w,u \in \Omega \tag{10}$$

上述关系成立，当 Bregman 距离生成函数 φ 在集合 Ω 上梯度 Lipschitz 连续，且 Lipschitz 参数 $L = \frac{1}{2}$，即，

$$\left\langle \nabla\varphi(w) - \nabla\varphi(u), w-u \right\rangle \le \frac{1}{2}\|w-u\|^2, \text{ 对任意 } w,u \in \Omega$$

观察发现，基于函数 φ 的梯度 Lipschitz 连续的性质，则对任意 $w,u \in \Omega$，我们有

$$\frac{1}{2}\|w-u\|^2 \ge \left\langle \nabla\varphi(w) - \nabla\varphi(u), w-u \right\rangle \ge \varphi(u) - \varphi(w) - \left\langle \nabla\varphi(w), w-u \right\rangle = B_\varphi(w,u),$$

其中后面的不等式可由函数 φ 在 Ω 上的凸性得到。而且，若 $\tilde{\varphi}$ 在 Ω 上满足梯度 Lipschitz 连续，且 Lipschitz 常数为 $L > 0$，

$$\left\langle \nabla\tilde{\varphi}(w) - \nabla\tilde{\varphi}(u), w-u \right\rangle \le L\|w-u\|^2, \text{ 对任意 } w,u \in \Omega$$

则存在函数 $\varphi = \frac{1}{2L}\tilde{\varphi}$，其梯度的 Lipschitz 常数为 $L = \frac{1}{2}$。这样的函数可以用做满足式(10)的距离生成函数。对于欧式空间，我们取 $\varphi(w) = \frac{1}{2}\|w\|_2^2$，得 Bregman 距离函数为 $B_\varphi(w,u) = \frac{1}{2}\|w-u\|^2$，同样满足式(10)。

假设 1：设随机次梯度 $\tilde{g}(w)$ 满足 $E\left[\tilde{g}(w) \mid w\right] = g(w)$，对任意 $w \in \Omega$，则存在某一标量 $\tilde{C} > 0$ 使得对任意 $w \in \Omega$，有 $E\left[\|\tilde{g}(w)\|_*^2 \mid w\right] \le \tilde{C}^2$。

当 $\tilde{g}(w) = g(w)$ 时，假设 1 中次梯度在 Ω 上一致有界，即存在 $C > 0$ 使得对任意 $w \in \Omega$，有 $\|g(w)\| \le C$。若目标函数的次梯度一致有界，且满足 $E\left[\tilde{g}(w) \mid w\right] = g(w)$，存在某个 $\gamma > 0$，使得对任意 $w \in \Omega$，

$E\left[\left\|\tilde{g}(w)-g(w)\right\|_*^2 \mid w\right]\le\gamma^2$，则对于一般范数，假设 1 成立，且 $\tilde{C}^2=2C^2+2\gamma^2$。显然欧式范数下有同样结论。

我们定义由算法生成的 σ-场如下：

$$\Gamma_k=\sigma\{w_0,\tilde{g}_0,\cdots,\tilde{g}_{k-1}\},\ \text{对任意}\ k\ge 1,$$

且 $\Gamma_0=\sigma\{w_0\}$。根据 σ-场，在假设 1 下，我们得到随机次梯度镜面下降法生成的序列 $\{w_k\}$ 有如下基本性质：

引理 1：若假设 1 成立，则根据算法(7)，对任意 $u\in\Omega$，$k\ge 0$，有

$$E\left[B_\varphi(w_{k+1},u)\mid\Gamma_k\right]+\alpha_k\langle g_k,w_k-u\rangle\le B_\varphi(w_k,u)+\frac{\alpha_k^2\tilde{C}^2}{2\mu_\varphi}.$$

证明：由点 w_{k+1} 处的一阶必要性条件得：

$$0\le\left\langle\alpha_k\tilde{g}_k+\nabla_u B_\varphi(w_k,w_{k+1}),u-w_{k+1}\right\rangle,\ \text{对任意}\ u\in\Omega$$

其中 $\nabla_u B_\varphi(\cdot,\cdot)$ 表示 Bregman 距离函数关于第二个变量的偏导数。由式(6)，我们有，

$$0\le\left\langle\alpha_k\tilde{g}_k+\nabla\varphi(w_{k+1})-\nabla\varphi(w_k),u-w_{k+1}\right\rangle,\ \text{对任意}\ u\in\Omega$$

等价地，

$$\alpha_k\langle\tilde{g}_k,w_{k+1}-u\rangle\le\left\langle\nabla\varphi(w_{k+1})-\nabla\varphi(w_k),u-w_{k+1}\right\rangle. \tag{11}$$

根据式(5)，令 $w=w_k,v=w_{k+1}$，则对任意 $u\in\Omega$，我们有

$$\left\langle\nabla\varphi(w_{k+1})-\nabla\varphi(w_k),u-w_{k+1}\right\rangle=B_\varphi(w_k,u)-B_\varphi(w_{k+1},u)-B_\varphi(w_k,w_{k+1}).$$

将其带入(11)式，可以发现，对任意 $u\in\Omega$，

$$\alpha_k\langle\tilde{g}_k,w_{k+1}-u\rangle\le B_\varphi(w_k,u)-B_\varphi(w_{k+1},u)-B_\varphi(w_k,w_{k+1}).$$

由 $\varphi(w)$ 的强凸性质，以及式(5)可推出，

$$\alpha_k\langle\tilde{g}_k,w_{k+1}-u\rangle\le B_\varphi(w_k,u)-B_\varphi(w_{k+1},u)-\frac{\mu_\varphi}{2}\left\|w_k-w_{k+1}\right\|^2. \tag{12}$$

下面求内积项，根据 Fenchel's 不等式 $\left|\langle p,q\rangle\right|\le\frac{1}{2}\left\|p^2\right\|+\frac{1}{2}\|q\|_*^2$，可得：

$$\begin{aligned}\alpha_k\langle\tilde{g}_k,w_{k+1}-u\rangle&=\alpha_k\langle\tilde{g}_k,w_{k+1}-w_k\rangle+\alpha_k\langle\tilde{g}_k,w_k-u\rangle\\&\ge-\left|\left\langle\frac{\alpha_k}{\sqrt{\mu_\varphi}}\tilde{g}_k,\sqrt{\mu_\varphi}(w_{k+1}-w_k)\right\rangle\right|+\alpha_k\langle\tilde{g}_k,w_k-u\rangle\\&\ge-\frac{\alpha_k^2}{2\mu_\varphi}\|\tilde{g}_k\|_*^2-\frac{\mu_\varphi}{2}\|w_{k+1}-w_k\|^2+\alpha_k\langle\tilde{g}_k,w_k-u\rangle\end{aligned}, \tag{13}$$

$\|\bullet\|_*$ 为范数 $\|\bullet\|$ 的共轭范数。消去 $\|w_{k+1}-w_k\|^2$，我们有

$$\alpha_k\langle\tilde{g}_k,w_k-u\rangle\le B_\varphi(w_k,u)-B_\varphi(w_{k+1},u)+\frac{\alpha_k^2}{2\mu_\varphi}\|\tilde{g}_k\|_*^2.$$

在 Γ_k 上取条件期望，再根据假设 1，我们得到

$$\alpha_k\langle\tilde{g}_k,w_k-u\rangle\le B_\varphi(w_k,u)-E\left[B_\varphi(w_{k+1},u)\mid\Gamma_k\right]+\frac{\alpha_k^2\tilde{C}^2}{2\mu_\varphi}.$$

从而推论得证。

引理 2：设 F 为 Ω 上的强凸函数，强凸系数 $\mu_F>0$，若假设 1 成立，且 Bregman 距离函数 B_φ 满足式(10)，步长取值满足条件(8)，则随机次梯度镜面下降迭代生成的序列 $\{w_k\}$ 与问题(1)的最优解 w_* 满足如下关系：

$$E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\mu_F}\frac{1}{\left(\sum_{t=0}^{k}\frac{1}{\alpha_t}\right)}\sum_{t=0}^{k}\frac{1}{\alpha_t}\left(E\left[F\left(w_t\right)\right]-F\left(w_*\right)\right)\le\left(k+1\right)\frac{\alpha_k^2\tilde{C}^2}{2\mu_F^2\mu_\varphi}.$$

证明：在假设 1 下，根据引理 1，令 $u=w_*$，则对任意 $k\ge0$，有

$$E\left[B_\varphi\left(w_{k+1},w_*\right)|\Gamma_k\right]+\frac{\alpha_k}{\mu_F}\left\langle g_k,w_k-w_*\right\rangle\le B_\varphi\left(w_k,w_*\right)+\frac{\alpha_k^2\tilde{C}^2}{2\mu_F^2\mu_\varphi},$$

由函数 F 的强凸性可得

$$\frac{\alpha_k}{\mu_F}\left\langle g_k,w_k-w_*\right\rangle\ge\frac{\alpha_k}{\mu_F}\left(F\left(w_k\right)-F\left(w_*\right)\right)+\frac{\alpha_k}{2}\left\|w_k-w_*\right\|^2\ge\frac{\alpha_k}{\mu_F}\left(F\left(w_k\right)-F\left(w_*\right)\right)+\alpha_kB\left(w_k,w_*\right).$$

结合上述两不等式，并对它们求总期望，进一步得：对任意 $k\ge0$，

$$E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{\alpha_k}{\mu_F}\left(E\left[F\left(w_k\right)\right]-F\left(w_*\right)\right)\le\left(1-\alpha_k\right)E\left[B_\varphi\left(w_k,w_*\right)\right]+\frac{\alpha_k^2\tilde{C}^2}{2\mu_F^2\mu_\varphi},\tag{14}$$

不等式两边同时乘以 $1/\alpha_k^2$，由 $\frac{1-\alpha_k}{\alpha_k^2}\le\frac{1}{\alpha_{k-1}^2}$，则对任意 $k\ge1$，我们有

$$\begin{aligned}&\frac{1}{\alpha_k^2}E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\alpha_k\mu_F}\left(E\left[F\left(w_k\right)\right]-F\left(w_*\right)\right)\\&\le\frac{\left(1-\alpha_k\right)}{\alpha_k^2}E\left[B_\varphi\left(w_k,w_*\right)\right]+\frac{\tilde{C}^2}{2\mu_F^2\mu_\varphi}\\&\le\frac{1}{\alpha_{k-1}^2}E\left[B_\varphi\left(w_k,w_*\right)\right]+\frac{\tilde{C}^2}{2\mu_F^2\mu_\varphi}\end{aligned}.$$

由于 $\alpha_0=1$，将上面不等式从 $k,k-1,\cdots,1$ 求和得

$$\frac{1}{\alpha_k^2}E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\mu_F}\sum_{t=1}^{k}\frac{1}{\alpha_t}\left(E\left[F\left(w_t\right)\right]-F\left(w_*\right)\right)\le E\left[B_\varphi\left(w_1,w_*\right)\right]+k\frac{\tilde{C}^2}{2\mu_F^2\mu_\varphi},\tag{15}$$

由于 $\alpha_0=1$，由(14)式知，当 $k=1$ 时，$E\left[B_\varphi\left(w_1,w_*\right)\right]$ 满足如下关系：

$$E\left[B_\varphi\left(w_1,w_*\right)\right]+\frac{1}{\mu_F}\left(E\left[F\left(w_0\right)\right]-F\left(w_*\right)\right)\le\frac{\tilde{C}^2}{2\mu_F^2\mu_\varphi}\text{。对任意 }k\ge1\tag{16}$$

联立(15)和(16)式，我们可得到，对任意 $k\ge1$

$$\frac{1}{\alpha_k^2}E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\mu_F}\sum_{t=0}^{k}\frac{1}{\alpha_t}\left(E\left[F\left(w_t\right)\right]-F\left(w_*\right)\right)\le\left(k+1\right)\frac{\tilde{C}^2}{2\mu_F^2\mu_\varphi},$$

将上式两边同时乘以 α_k^2，然后运用关系 $\alpha_k^2\ge1\Big/\left(\sum_{t=0}^{k}\frac{1}{\alpha_t}\right)$，从而推论得证。

定理 2：设 F 为 Ω 上的强凸函数，强凸系数 $\mu_F>0$。若假设 1 成立，且 Bregman 距离函数 B_φ 满足式(10)，步长取(9)中任何一个，则对随机次梯度镜面下降迭代法生成的序列 $\{w_k\}$，我们有：

$$E\left[\left\|w_t-w_*\right\|^2\right]\le\frac{4}{k+1}\frac{\tilde{C}^2}{\mu_F^2\mu_\varphi^2},\text{ 对任意 }k\ge0,$$

则对加权平均迭代序列 $\{\hat{w}_k\}$，我们有：

$$E\left[F\left(\hat{w}_t\right)\right]-F\left(w_*\right)\le\frac{2}{k+1}\frac{\tilde{C}^2}{\mu_F\mu_\varphi},\ \text{对任意}\ k\ge 0,$$

$$E\left[\left\|\hat{w}_t-w_*\right\|^2\right]\le\frac{4}{k+1}\frac{\tilde{C}^2}{\mu_F^2\mu_\varphi},\ \text{对任意}\ k\ge 0,$$

其中 w_* 为问题(1)的最优解。

证明：根据引理 2，对任意 $k\ge 0$，我们有：

$$E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\mu_F}\frac{1}{\left(\sum_{t=0}^{k}\frac{1}{\alpha_t}\right)}\sum_{t=0}^{k}\frac{1}{\alpha_t}\left(E\left[F\left(w_t\right)\right]-F\left(w_*\right)\right)\le(k+1)\frac{\alpha_k^2\tilde{C}^2}{2\mu_F^2\mu_\varphi}.$$

有步长序列 $\{\alpha_k\}$ 满足：对任意 $k\ge 0$，$\alpha_k\le\frac{2}{k+1}$，所以

$$E\left[B_\varphi\left(w_{k+1},w_*\right)\right]+\frac{1}{\mu_F}\frac{1}{\left(\sum_{t=0}^{k}\frac{1}{\alpha_t}\right)}\sum_{t=0}^{k}\frac{1}{\alpha_t}\left(E\left[F\left(w_t\right)\right]-F\left(w_*\right)\right)\le\frac{2}{k+1}\frac{\alpha_k^2\tilde{C}^2}{\mu_F^2\mu_\varphi},\tag{17}$$

由 F 的强凸性以及加权平均迭代序列 $\{\hat{w}_k\}$ 的定义，上式可总结为：

$$E\left[F\left(\hat{w}_k\right)\right]-F\left(w_*\right)\le\frac{2}{k+1}\frac{\tilde{C}^2}{\mu_F\mu_\varphi},\ \text{对任意}\ k\ge 0,$$

又 F 在 Ω 上的强凸性，我们有 $F\left(\hat{w}\right)-F\left(w_*\right)\ge\frac{\mu_F}{2}\left\|\hat{w}-w_*\right\|^2$，则

$$E\left[\left\|\hat{w}_k-w_*\right\|^2\right]\le\frac{4}{k+1}\frac{\tilde{C}^2}{\mu_F^2\mu_\varphi},\ \text{对任意}\ k\ge 0,$$

而且由式(17)我们可得：

$$E\left[B_\varphi\left(w_{k+1},w_*\right)\right]\le\frac{2}{k+1}\frac{\tilde{C}^2}{\mu_F^2\mu_\varphi},\ \text{对任意}\ k\ge 0,$$

则由函数 φ 的强凸性，我们有

$$E\left[\left\|w_k-w_*\right\|^2\right]\le\frac{4}{k+1}\frac{\tilde{C}^2}{\mu_F^2\mu_\varphi^2},\ \text{对任意}\ k\ge 0.$$

5. 总结

本文在基本的镜面下降方法的基础上引入随机思想，得到随机次梯度镜面下降方法，并将该方法与加权平均迭代法相结合，进一步探讨了带有加权平均迭代[11]的随机次梯度下降镜面下降方法。我们考虑了随机次梯度镜像下降法的最优性条件，通过使用自适应步长求得权值，从而得到该方法的加权平均迭代值。本文的创新点是在迭代平均值的构造中使用自适应步长选择来获得权重，通过使用所提出的权值，我们可以恢复强凸函数和仅凸函数的已知速率。

基金项目

国家自然科学基金(11461013)，贵州省公共大数据重点实验开放课题项目(2017BDKFJJ012)。

参考文献

[1] Juditsky, A. and Nemirovski, A.S. (2010) First Order Methods for Nonsmooth Convex Large-Scale Optimization, II:

Utilizing Problem's Structure. MIT Press, Cambridge.

[2] Ben-Tal, A., Margalit, T. and Nemirovski, A. (2001) The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography. *Society for Industrial and Applied Mathematics*, **12**, 79-108.

[3] Rakhlin, A., Shamir, O. and Sridharan, K. (2012) Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. arXiv:1109.5647 [cs.LG]

[4] Beck, A. and Teboulle, M. (2003) Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters*, **31**, 167-175. https://doi.org/10.1016/S0167-6377(02)00231-6

[5] Duchi, J.C., Shalev-Shwartz, S. and Singer, Y. (2010) Composite Objective Mirror Descent. 23*rd Conference on Learning Theory*, Haifa, 27-29 June 2010, 14-26.

[6] Sra, S., Nowozin, S. and Wright, S.J. (2011) Optimization for Machine Learning. MIT Press, Cambridge.

[7] Brègman, L.M. (1967) The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics & Mathematical Physics*, **7**, 200-217. https://doi.org/10.1016/0041-5553(67)90040-7

[8] Tseng, P. (2008) On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. *SIAM Journal on Optimization*.

[9] 陶蔚, 潘志松, 陶卿. 使用 Nesterov 步长策略投影次梯度方法的个体收敛性[J]. 计算机学报. 2018 (1).

[10] Beck, A. and Teboulle, M. (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**, 183-202. https://doi.org/10.1137/080716542

[11] Luong, D.V.N., Parpas, P. and Rueckert, D. (2016) A Weighted Mirror Descent Algorithm for Non-Smooth Convex Optimization Problem. *Journal of Optimization Theory & Applications*, **170**, 900-915. https://doi.org/10.1007/s10957-016-0963-5