Advances in Applied Mathematics 应用数学进展, 2019, 8(7), 1243-1250
Published Online July 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.87144

Hans汉斯

The Bundle Methods Based on the Regularized Risk Minimization Model

Jia Qin, Yanni Li

School of Mathematics and Information Science, Guangxi University, Nanning Guangxi
Email: qinjia1992@qq.com, ynli@st.gxu.edu.cn

Received: July 4th, 2019; accepted: July 19th, 2019; published: July 26th, 2019

Abstract

Based on the regularized risk function model (SRM) of machine learning (ML), this paper combines the bundle method for solving non-smooth functions, presents an algorithm for solving empirical risk function model. The objective function is approximated by the cut-plane model, and the step size is obtained by the inexact line search. Under appropriate assumptions, the global convergence and convergence speed of the algorithm are analyzed.

Keywords

Machine Learning, Bundle Methods, Risk Minimization, Global Convergence

基于正则化风险函数模型的束方法

覃　嘉，李艳妮

广西大学数学与信息科学学院，广西　南宁
Email: qinjia1992@qq.com, ynli@st.gxu.edu.cn

收稿日期：2019年7月4日；录用日期：2019年7月19日；发布日期：2019年7月26日

摘　要

本文在机器学习(ML)的正则化风险函数模型(SRM)的基础上，结合求解非光滑函数的束方法，提出了一种求解经验风险函数模型的算法，用割平面模型去近似目标函数，用非精确线搜索去获得步长，在适当的假设下，分析了算法的全局收敛和收敛速度。

文章引用: 覃嘉, 李艳妮. 基于正则化风险函数模型的束方法[J]. 应用数学进展, 2019, 8(7): 1243-1250.
DOI: 10.12677/aam.2019.87144

关键词

机器学习，束方法，风险最小化，全局收敛

1. 引言

本文主要考虑以下最优化问题(文[1]):

$$\min_{w} J(w) := \lambda\Omega(w) + R_{emp}(w), \tag{1.1}$$

其中，$R_{emp}(w) := \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, w)$ 是经验风险函数。$x_i \in X \subseteq R^d$ 是训练实例，$y_i \in Y$ 是相应的标签。l 是一个凸的损失函数，用来衡量 y 与通过 w 产生的预测值之间的差异，例如，$l(x, y, w) = (\langle w, x\rangle - y)^2$，其中 $\langle\cdot,\cdot\rangle$ 定义为标准的欧几里得内积。最后，$\Omega(w)$ 是一个凸函数且作为正则化项，而 $\lambda \geq 0$ 是正则化参数。特别的，$\Omega(w)$ 是可微的且易于计算。与之相反的，经验风险函数项 $R_{emp}(w)$ 通常是不可微的，而且它的计算代价总是很大的(文[2])。

目前，各种各样的机器学习问题都可以描述成一类最小化正则风险函数问题，具体地，不同的机器学习算法会形成不同的风险函数和正则化项，例如：线性支持向量机(SVMs)，逻辑回归，条件随机场(CRFs)等等。因而研究此类凸优化问题具有重要的理论和现实意义。本文在束方法的基础上，结合一类 BMRM 算法，对 BMRM 算法的精确线搜索进行改进，推广到非精确的情形，给出一个改进的 BMRM 算法来求解正则化风险函数问题，在一定条件下，证明算法的全局收敛性。

2. 基础知识

机器学习问题建立在最优化基础理论上，为了更好的研究机器学习，下面将简要介绍与最优化理论相关的基本概念和一些重要结论，次梯度和束方法等。先介绍一些记号，其中：序列或者集合元素的索引记在下标，如 u_1, u_2。向量 u 的第 i 个分量记为 $u^{(i)}$，$[k]$ 表示集合 $\{1,2,\cdots,k\}$，L_p 范数定义为：$\|u_p\| = \left(\sum_{i=1}^{d}\left|u^{(i)}\right|^p\right)^{1/p}$，这里的 $p \geq 1$，然后使用 $\|\cdot\|$ 来定义 $\|\cdot\|_2$。

次梯度(文[3]**)**：对于一类非光滑的凸函数，次梯度是对于梯度的一个近似，假设 w' 是凸函数 J 上的一点，s' 是 J 在点 w' 处的次梯度，当且仅当：

对任意的 w，有 $J(w) \geq J(w') + \langle w - w', s'\rangle$ 成立 (2.1)

则在 w' 处所有的次梯度组成的集合叫做次微分，记为 $\partial_w J(w')$。

切平面模型(CPM) (文[4]**)**：

给定点 $w_0, w_1, \cdots, w_{t-1}$ 处的次梯度 $s_1, s_2, \cdots, s_t$，产生一系列分段线性下界来迫近 J：

$$J(w) \geq J_t^{CP}(w) := \max_{1\leq i\leq t}\{J(w_{i-1}) + \langle w - w_{i-1}, s_i\rangle\}. \tag{2.2}$$

迭代点 w_t 的更新：

$$w_t := \arg\min_w J_t^{CP}(w). \tag{2.3}$$

切平面模型(CPM)的模型迭代过程不断改进分段线性下界 J^{CP}，且使整个模型逐步逼近 J 的最小值。

邻近束方法(Proximal Bundle Methods) (文[5]**)：**

虽然切平面模型是收敛的，但是当新的迭代点与上一个迭代点距离较大时，切平面模型收敛速度很慢。为了改进切平面模型的不稳定性，邻近束方法通过增加一个邻近函数来改进下一个点的选取，其中，一般选取 $\frac{1}{2}\|\cdot\|^2$ 作为邻近函数：

$$w_t := \arg\min_w \left\{\frac{\zeta_t}{2}\|w-\hat{w}_{t-1}\|^2 + J_t^{CP}(w)\right\}. \tag{2.4}$$

这里 $\hat{w}_{t-1}$ 是最近的邻近中心，ζ_t 是正邻近项参数。但是，这样的改进会使束方法需要对参数进行仔细地调整。

3. 一个新的最小化正则风险函数算法(BMRM)

结合次梯度和邻近束方法，文章[6]提出了一种新的机器学习凸优化算法：BMRM 算法。在求解一类机器学习问题(1.1)时，对于每一次迭代 t，BMRM 算法都会构建下界 R_t^{CP} 来近似经验风险函数 R_{emp}，而新的迭代点 w_t 则由最小化 J_t 更新，J_t 由 R_t^{CP} 与正则项 Ω 组成。这是 BMRM 算法和邻近束方法的最大不同。

实际上，BMRM 算法不涉及邻近中心，进一步，BMRM 算法甚至不包括邻近项。

BMRM 算法具体如下：

算法 1 (BMRM 算法)

步骤 1：输入及初始化：$\varepsilon \geq 0, w_0, t := 0$

步骤 2：循环：$t := t+1$，计算：

$$a_t \in \partial_\omega R_{emp}(w_{t-1}),$$
$$b_t := R_{emp}(w_{t-1}) - \langle w_{t-1}, a_t \rangle,$$

步骤 3：(更新模型) $R_t^{CP} := \max_{1\leq i\leq t}\{\langle w, a_i\rangle + b_i\}$

步骤 4：(迭代点更新) $w_t := \arg\min_w J_t(w) := \lambda\Omega(w) + R_t^{CP}(w)$

步骤 5：$\varepsilon_t := \min_{0\leq i\leq t} J(w_i) - J_t(w_t)$

步骤 6：直到 $\varepsilon_t \leq \varepsilon$，返回 w_t

与邻近束方法相比较，BMRM 算法有两点不同：

a) BMRM 算法是用一个分段线性下界来接近 $R_{emp}(w)$ 而不是 $J(w)$。

b) 在邻近束方法中，邻近项($\frac{\zeta}{2}\|w-\hat{w}_t\|^2$)被正则化项 $\Omega(w)$ 替代，因此，就不需要调整邻近参数了。

进一步，考虑存在有效的精确线搜索时，BMRM 算法可以实现更快速的收敛，所以，结合线搜索的 BMRM 算法如下(文[4])：

算法 2 (结合精确线搜索的 BMRM 算法)

步骤 1：输入及初始化：$\varepsilon \geq 0, \theta \in (0,1], w_0^b, w_0^c := w_0^b, t := 0$

步骤 2：循环：$t := t+1$，计算：

$$a_t \in \partial_\omega R_{emp}\left(w_{t-1}^c\right),$$
$$b_t := R_{emp}\left(w_{t-1}^c\right) - \left\langle w_{t-1}^c, a_t \right\rangle,$$

步骤 3：(更新模型) $R_t^{CP} := \max_{1\le i\le t}\left\{\left\langle w, a_i\right\rangle + b_i\right\}$

步骤 4： $w_t := \arg\min_w J_t(w) := \lambda\Omega(w) + R_t^{CP}(w)$

步骤 5：(线搜索)

$$\eta_t := \arg\min_{\eta\in R} J\left(w_{t-1}^b + \eta\left(w_t - w_{t-1}^b\right)\right)$$
$$w_t^b := w_{t-1}^b + \eta_t\left(w_t - w_{t-1}^b\right)$$
$$w_t^c := (1-\theta)w_t^b + \theta w_t$$

步骤 6： $\varepsilon_t := \min_{0\le i\le t} J\left(w_t^b\right) - J_t\left(w_t\right)$

直到 $\varepsilon_t \le \varepsilon$，返回 w_t^b。

注意到，对所有的 t，如果设 $\eta_t = 1$，则算法 2 可以简化为算法 1，且两个算法使用相同的终止准则. 值得注意的是，带线搜索的 BMRM 算法并不能应用于结构化的机器学习问题，例如：最大边际马尔科夫网络问题。

4. 改进的非精确线搜索 BMRM 算法

在机器学习的优化问题中，算法 2 中的精确线搜索往往需要计算很多的函数值和梯度值，这样的话会产生较大的计算代价，特别是当迭代点离最优点比较远的时候，精确线搜索往往不是最有效和合理的。所以，对于许多像问题(1.1)的机器学习凸优化问题，非精确线搜索既能保证目标函数具有可接受的下降量，又能使最终形成的迭代序列收敛，以下将考虑非精确线搜索和 BMRM 算法的结合。

算法 3 (结合非精确线搜索的 BMRM 算法)

步骤 1：输入及初始化： $\varepsilon \ge 0, \theta \in (0,1], w_0^b, w_0^c := w_0^b, t := 0$

步骤 2：循环： $t := t+1$，计算：

$$a_t \in \partial_\omega R_{emp}\left(w_{t-1}^c\right),$$
$$b_t := R_{emp}\left(w_{t-1}^c\right) - \left\langle w_{t-1}^c, a_t \right\rangle,$$

步骤 3：(更新模型) $R_t^{CP} := \max_{1\le i\le t}\left\{\left\langle w, a_i\right\rangle + b_i\right\}$

步骤 4： $w_t := \arg\min_w J_t(w) := \lambda\Omega(w) + R_t^{CP}(w)$

步骤 5：(线搜索)内循环： $m := 0, \beta \in (0,1), \sigma \in (0, 0.5)$

$$\text{if } J\left(w_{t-1}^b + \beta^m\left(w_t - w_{t-1}^b\right)\right) \le J\left(w_{t-1}^b\right) + \sigma\beta^m\left(a_t + \lambda\Omega'(w)\right)\left(w_t - w_{t-1}^b\right)$$
$$\eta_t = \beta^m$$
$$m := m+1$$

步骤 6：更新点列：

$$w_t^b := w_{t-1}^b + \eta_t\left(w_t - w_{t-1}^b\right)$$
$$w_t^c := (1-\theta)w_t^b + \theta w_t$$

步骤 7： $\varepsilon_t := \min_{0\le i\le t} J\left(w_t^b\right) - J_t\left(w_t\right)$

直到 $\varepsilon_t \le \varepsilon$，返回 w_t^b。

在证明算法 3 的收敛性之前，先介绍相关引理。

定义 4.1 (Fenchel 对偶) 定义 $\Omega: W \to R$ 是一个在凸集 W 上的凸函数，则 Ω 的对偶 Ω^* 定义为：

$$\Omega^*(\mu) := \sup_{w \in W} \langle w, \mu \rangle - \Omega(w).$$

定理 4.1 定义 $A = [a_1, \cdots, a_t]$ 的列向量为(次)梯度，$b = [b_1, \cdots, b_t]$，则问题：

$$w_t := \arg\min_{w \in R^d} \left\{ J_t(w) := \max_{1 \le i \le t} \langle w, a_i \rangle + b_i + \lambda \Omega(w) \right\} \tag{4.1}$$

的对偶问题为：

$$\alpha_t = \arg\max_{\alpha \in R^t} \left\{ J_t^*(\alpha) := -\lambda \Omega^*\left(-\lambda^{-1} A\alpha\right) + \alpha^{\mathrm{T}} b \,|\, \alpha \ge 0, \|\alpha\|_1 = 1 \right\} \tag{4.2}$$

此外，w_t 和 α_t 之间有对偶关系式 $w_t = \partial \Omega^*\left(-\lambda^{-1} A\alpha_t\right)$。

证明： 首先将(4.1)式改写成一个带约束的最优化问题：

$$\min_{w,\xi} \lambda \Omega(w) + \xi$$
$$\text{st. } \xi \ge \langle w, a_i \rangle + b_i$$

式中 $i = 1, 2, \cdots, t$。通过引入非负的拉格朗日乘子 α 和元素都是 1 的 t 维向量 1_t，相应的拉格朗日函数可以写成：

$$\begin{cases} L(w, \xi, \alpha) = \lambda \Omega(w) + \xi - \alpha^{\mathrm{T}} \left(\xi 1_t - A^{\mathrm{T}} w - b\right) \\ \alpha \ge 0 \end{cases} \tag{4.3}$$

式中，$\alpha \ge 0$ 表示向量 α 的每一个分量都是非负的，对 ξ 求导可得：$1 - \alpha^{\mathrm{T}} 1_t = 0$。此外，对变量 w 最小化 L 意味着求解问题：

$$\max_w \left\langle w, -\lambda^{-1} A\alpha \right\rangle - \Omega(w) = \Omega^*\left(-\lambda^{-1} A\alpha\right).$$

把上述两项插回到(4.3)式中，就可以消除了原始变量 ξ 和 w。得证。

推论 4.1 当 $\Omega(w) = \frac{1}{2}\|w\|_2^2$ 时，即取二次正则项的时候，(4.2)式可以写成：

$$\alpha_t = \arg\max_{\alpha \in R^t} \left\{ -\frac{1}{2\lambda} \alpha^{\mathrm{T}} A^{\mathrm{T}} A\alpha + \alpha^{\mathrm{T}} b \,|\, \alpha \ge 0, \|\alpha\|_1 = 1 \right\}$$

5. 收敛性分析

接下来证明结合非精确线搜索的 BMRM 算法具有较好的收敛速度。事实上，对于求解非光滑优化问题，水平束方法收敛速度为 $O\left(1/\varepsilon^2\right)$。本文将证明，对于求解具有非光滑损失函数的正则化风险函数问题，算法 3 将具有 $O(1/\varepsilon)$ 的收敛速度。更具体的说，我们将证明以下收敛性：

假设 $\max_{u \in \partial_w R_{emp}(w)} \|u\| \le G$，而对于正则化项 $\Omega(w)$ 有 $\left\|\partial_\mu^2 \Omega^*(\mu)\right\| \le H^*$，本文将证明在上述假设下，算法具有 $O(1/\varepsilon)$ 收敛速度，即算法将在 $O(1/\varepsilon)$ 的迭代次数中，得到其 ε-最优解。

在以下的证明过程中，我们使用了类似于文[7]的技术：

定理 5.1 假设 $\max_{u \in \partial_w R_{emp}(w)} \|u\| \le G$ 对于所有的 $w \in domJ$ 成立，且 $\left\|\partial_\mu^2 \Omega^*(\mu)\right\| \le H^*$ 对于所有的 $\mu \in \left\{ -\lambda^{-1} \sum_{i=1}^{t+1} \alpha_i a_i \right\}$ 成立，这里 $\alpha_i \ge 0$，且对 $\forall i$ 有 $\sum_{i=1}^{t+1} \alpha_i = 1$ 成立，即，存在：

$$\varepsilon_t-\varepsilon_{t+1}\geq\frac{\varepsilon_t}{2}\min\left(1,\frac{\lambda\varepsilon_t}{4G^2H^*}\right).$$

证明：定理 5.1 的证明参考文献[4]。

定理 5.2 假设对于所有的 w 有 $J(w)\geq 0$，在定理 5.1 成立的条件下，可以证明算法 1 的收敛性。即对任意的 $\varepsilon<\frac{4G^2H^*}{\lambda}$，算法会在至多：

$$n\leq\log_2\frac{\lambda J(0)}{G^2H^*}+\frac{8G^2H^*}{\lambda\varepsilon}-1$$

步后，收敛到所需精度。

证明：定理 5.2 的证明参考文献[4]。

在证明算法 3 收敛性之前先介绍相关引理。

引理 5.1 $J_{t+1}(w_t)=\lambda\Omega(w_t)+\langle w_t,a_{t+1}\rangle+b_{t+1}$。

证明：由 w_t^b 和 w_t 与 w_{t-1}^b 的关系，且 w_t^c 是 w_t 和 w_t^b 的凸组合，此外，根据 a_{t+1} 和 b_{t+1} 的定义，得到 $J(w_t^c)=J_{t+1}(w_t^c)$。因此有：

$$J(w_t^c)=J_{t+1}(w_t^c)=\lambda\Omega(w_t^c)+\langle a_{t+1},w_t^c\rangle+b_{t+1}\geq J(w_t^b).\tag{5.1}$$

因为 Ω 是凸的，所以：

$$\Omega\left((1-\theta)w_t^b+\theta w_t\right)\leq(1-\theta)\Omega(w_t^b)+\theta\Omega(w_t),$$

即：

$$\theta\left(\Omega(w_t^b)-\Omega(w_t)\right)\leq\Omega(w_t^b)-\Omega(w_t^c),$$

上式两边乘以 λ，然后加上 $\theta R_{emp}(w_t^b)$ 和减去 $\theta R_t(w_t)$ 得：

$$\begin{aligned}&\lambda\theta\,\Omega(w_t^b)+\theta R_{emp}(w_t^b)-\lambda\theta\,\Omega(w_t)-\theta R_t(w_t)\\&\leq\lambda\Omega(w_t^b)+R_{emp}(w_t^b)-\lambda\Omega(w_t^c)-(1-\theta)R_{emp}(w_t^b)-\theta R_t(w_t),\end{aligned}$$

结合算法 3 的步骤 7：

$$\theta\varepsilon_t\leq J(w_t^b)-\lambda\Omega(w_t^c)-(1-\theta)R_{emp}(w_t^b)-\theta R_t(w_t),\tag{5.2}$$

联立(5.1)和(5.2)：

$$\langle a_{t+1},w_t^c\rangle+b_{t+1}\geq J(w_t^b)-\lambda\Omega(w_t^c)\geq(1-\theta)R_{emp}(w_t^b)+\theta R_t(w_t)+\theta\varepsilon_t$$

由 $w_t^c=(1-\theta)w_t^b+\theta w_t$ 可得：

$$(1-\theta)\langle a_{t+1},w_t^b\rangle+\theta\langle a_{t+1},w_t\rangle+b_{t+1}\geq J(w_t^b)-\lambda\Omega(w_t^c)\geq(1-\theta)R_{emp}(w_t^b)+\theta R_t(w_t)+\theta\varepsilon_t,$$

化简得：

$$(1-\theta)\left(\langle a_{t+1},w_t^b\rangle-R_{emp}(w_t^b)\right)+\theta\left(\langle a_{t+1},w_t\rangle-R_t(w_t)\right)+b_{t+1}\geq J(w_t^b)-\lambda\Omega(w_t^c)\geq\theta\varepsilon_t.\tag{5.3}$$

因为 $\langle w_t^b,a_{t+1}\rangle+b_{t+1}$ 是凸函数 R_{emp} 的泰勒近似，所以有：

$$R_{emp}(w_t^b)\geq\langle w_t^b,a_{t+1}\rangle+b_{t+1}\tag{5.4}$$

把(5.4)带入(5.3)，得到：

$$(1-\theta)(-b_{t+1})+\theta\left(\left\langle w_t^b, a_{t+1}\right\rangle - R_t(w_t)\right)+b_{t+1}\geq\theta\varepsilon_t$$

两边除以$\theta>0$，得到：

$$\left\langle w_t^b, a_{t+1}\right\rangle + b_{t+1}\geq R_t(w_t)+\varepsilon_t$$

又：

$$R_{t+1}(w_t)=\max\left(\left\langle w_t, a_{t+1}\right\rangle + b_{t+1}, R_t(w_t)\right)=\left\langle w_t, a_{t+1}\right\rangle + b_{t+1}$$

$$J_{t+1}(w_t)=\lambda\Omega(w_t)+R_{t+1}(w_t)$$

故：$J_{t+1}(w_t)=\lambda\Omega(w_t)+\left\langle w_t, a_{t+1}\right\rangle + b_{t+1}$.

引理 5.2 $\varepsilon_t-\varepsilon_{t+1}\geq J_{t+1}(w_{t+1})-J_t(w_t)$。

证明：

$$\begin{aligned}\varepsilon_t-\varepsilon_{t+1}&=J\left(w_t^b\right)-J_t(w_t)-J\left(w_{t+1}^b\right)+J_{t+1}(w_{t+1})\\&=J\left(w_t^b\right)-J\left(w_{t+1}^b\right)+J_{t+1}(w_{t+1})-J_t(w_t)\\&\geq J_{t+1}(w_{t+1})-J_t(w_t).\end{aligned}$$

引理 5.3 令α_t如定理 4.1 中的定义，$\bar{A}:=[a_1,\cdots,a_{t+1}]$和$\bar{b}:=[b_1,\cdots,b_{t+1}]$，则在定理 5.1 里下述条件成立$\max_{u\in\partial_w R_{emp}(w)}\|u\|\leq G$情况下，下式成立：

$$[-\alpha_t,1]^{\mathrm{T}}\,\bar{A}^{\mathrm{T}}\bar{A}[-\alpha_t,1]\leq 4G^2.$$

证明：首先，有对偶关系$\partial_w\lambda\Omega(w_t)=-A\alpha_t$，且(4.2)在第$t$迭代时，$\alpha_t\geq 0$和$\|\alpha_t\|_1=1$，随后，因为$\partial_w\lambda\Omega(w_t)$在$a_{t'}\in\partial_w R_{emp}\left(w_{t'}^c\right),\forall t'\leq t$的凸包里面，因此$\left\|\partial_w\lambda\Omega(w_t)\right\|\leq G$，接下来，由 Cauchy-Schwarz 不等式得：

$$\begin{aligned}[-\alpha_t,1]^{\mathrm{T}}\,\bar{A}^{\mathrm{T}}\bar{A}[-\alpha_t,1]&=\left\|\partial_w\lambda\Omega(w_t)+a_{t+1}\right\|^2\\&=\left\|\partial_w\lambda\Omega(w_t)\right\|^2+2\partial_w\lambda\Omega(w_t)^{\mathrm{T}}a_{t+1}+\left\|a_{t+1}\right\|^2\\&\leq 4G^2.\end{aligned}$$

引理 5.4 对满足以下递推条件：$c>0$是一个常数，当$t\geq 1$时

$$\rho_t-\rho_{t+1}\geq c(\rho_t)^2$$

的非负序列$\langle\rho_1,\rho_2,\cdots\rangle$，有当整数$t\geq 1$时：

$$\rho_t\leq\frac{1}{c\left(t-1+\dfrac{1}{\rho_1 c}\right)}.$$

此外，当$t\geq\dfrac{1}{c\rho}-\dfrac{1}{\rho_1 c}+1$时：总有$\rho_t\leq\rho$成立。

证明：此引理通过归纳法易得，参考文献[8]。

定理 5.3 在定理 5.1 成立条件下，算法 3 收敛到任意$\varepsilon<4G^2H^*/\lambda$精度需要：

$$n\leq\frac{8G^2H^*}{\lambda\varepsilon}\text{ 步。}$$

证明：对于$\varepsilon<\dfrac{4G^2H^*}{\lambda}$，结合定理 5.1 及其证明过程，易知：

$$\varepsilon_t - \varepsilon_{t+1} \geq \frac{\lambda \varepsilon_t^2}{8G^2H^*}$$

又由引理 5.4 可知 $\varepsilon_t \leq \frac{1}{c\left(t-1+\frac{1}{\varepsilon_1 c}\right)}$，这里 $c = \frac{\lambda}{8G^2H^*}$。令 $\frac{1}{c\left(t-1+\frac{1}{\varepsilon_1 c}\right)} = \varepsilon$，假设 $\varepsilon_1 > 0$，然后解出 n，有：

$$n \leq \frac{1}{c\varepsilon} = \frac{8G^2H^*}{\lambda\varepsilon}.$$

参考文献

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[2] 李航. 统计机器学习[M]. 北京: 清华大学出版社, 2012.

[3] 袁亚湘, 孙文瑜. 最优化理论与方法[M]. 武汉: 科学出版社, 2002.

[4] Kelley, J.E. (1960) The Cutting-Plane Method for Solving Convex Programs. *Journal of the Society for Industrial & Applied Mathematics*, **8**, 703-712. https://doi.org/10.1137/0108053

[5] 陈韵梅, 张维. 基于近似一阶信息的加速的 Bundle Level 算法[J]. 中国科学: 数学, 2017(10): 1119-1142.

[6] Teo, C.H., Viswanathan, S.V.N., Smola, A.J. and Le, Q. (2010) Bundle Methods for Regularized Risk Minimization, *Journal of Machine Learning Research*, **11**, 311-365.

[7] Shalev-Shwartz and Singer, Y. (2006) Online Learning Meets Optimization in the Dual. In: Simon, H.U. and Lugosi, G., Eds., *Computational Learning Theory*, Springer, Berlin, 423-437. https://doi.org/10.1007/11776420_32

[8] Abe, N., Takeuchi, J. and Warmuth, M.K. (2001) Polynomial Learn Ability of Stochastic Rules with Respect to the KL-Divergence and Quadratic Distance. *IEICE Transactions on Information and Systems*, **84**, 299-316.