编号：2024-3-1145141919　　级别：公开

优化基本理论与方法课程研究报告

Optimal Transport Based Distributed Optimization Research

（2024 年 1 月）

郑皓壬　　（3220103230）
郑俊达　　（3220103540）
李瀚轩　　（3220106039）

浙江大学计算机科学与技术学院

Contents

Abstract

1 论文介绍

1.1 问题背景

[?]

1.2 论文贡献

1.3 章节组织

2 相关工作

3 问题描述和常用记号

3.1 BFGS 算子与算法

3.2 Greedy-BFGS 算法

4 方法描述

Sharpened-BFGS 算法的核心思想就是同时使用经典的 BFGS 算法和贪心 BFGS 算法来更新 G_t。经典的 BFSG 算法可以提升牛顿方向的近似估计，而贪心 BFGS 算法能够提高整体 Heissan 矩阵的估计精确度。

4.1 二次规划

对于一般的优化问题 $\min\limits_{x\in\mathbb{R}^n} f(x)$, 我们考虑目标函数为二次函数的情形，即

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{2}x^T Ax + b^T x \tag{1}$$

其中 $A \in \mathbb{R}^{d\times d}$ 是对称正定矩阵，满足条件 $\mu I \preceq A \preceq LI, b \in \mathbb{R}^d$。

对于这类优化问题，Sharpened-BFGS 算法在每轮迭代的时候进行了两次更新。算法首先通过经典 BFGS 算法更新得到矩阵 $\overline{G}_t$，然后使用贪心 BFGS 算法选择下降方向 $\overline{u}$, 最后由 $\overline{G}_t$ 和 $\overline{u}$ 更新得到 G_{t+1}。即 Shapened-BFGS 通过经典 BFGS 方向对贪心 BFGS 得到的 Hessian 矩阵估计进行了改进。伪代码如下：

Algorithm. 1 Sharpened-BFGS applied to quadratic programming

Require: 初始化变量 x_0, 初始化近似矩阵 $G_0 = LI$

for $t = 0, 1, 2, \cdots$ **do**

更新变量 $x_{t+1} = x_t - G_t^{-1}\nabla f(x_t)$;

计算 BFGS 方向 $s_t = x_{t+1} - x_t$;

计算 BFGS 矩阵 $\overline{G}_t = BFGS(A, G_t, s_t)$;

计算贪心 BFGS 方向 $\overline{u} = \overline{u}(A, \overline{G}_t)$;

估计 Hessian 矩阵 $G_{t+1} = BFGS(A, \overline{G}_t, \overline{u}_t)$;

end for

为了定量地描述 Sharpened-BFGS 算法的收敛速度，验证其确实融合了经典 BFGS 和贪心 BFGS 算法的优点，我们首先需要定义牛顿衰减因子：

$$\lambda_f(x) = \sqrt{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)} \tag{2}$$

其是用于衡量优化问题迭代过程中每一步近似解的改进程度的指标。它比较当前点的目标函数值和通过二阶信息进行近似预测的下降量。对于其形式的推导如下：

我们想要最小化 $f(\boldsymbol{x})$，考虑它的二阶泰勒展开：

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_0}) + (\boldsymbol{x} - \boldsymbol{x_0})^T \nabla f(\boldsymbol{x_0}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_0})^T \nabla^2 f(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{x_0}) \tag{3}$$

由于海森矩阵 $\nabla^2 f(\boldsymbol{x_0})$ 是正定的，所以为了使 $f(\boldsymbol{x})$ 最小，我们将上式对 $\boldsymbol{x}$ 求偏导，得到：

$$\nabla f(\boldsymbol{x_0}) + \nabla^2 f(\boldsymbol{x_0})(\boldsymbol{x^*} - \boldsymbol{x_0}) = 0 \tag{4}$$

其中 $\boldsymbol{x^*}$ 是 $f(\boldsymbol{x})$ 的极小值点，所以有：

$$\boldsymbol{x^*} = \boldsymbol{x_0} - \nabla^2 f(\boldsymbol{x_0})^{-1} \nabla f(\boldsymbol{x_0}) \tag{5}$$

将 $\boldsymbol{x^*}$ 代入 $f(\boldsymbol{x})$ 的二阶泰勒展开式，得到：

$$\begin{aligned} f(\boldsymbol{x^*}) = f(\boldsymbol{x_0}) &- (\nabla^2 f(\boldsymbol{x_0})^{-1} \nabla f(\boldsymbol{x_0}))^T \nabla f(\boldsymbol{x_0}) \\ &+ \frac{1}{2}(\nabla^2 f(\boldsymbol{x_0})^{-1} \nabla f(\boldsymbol{x_0}))^T \nabla^2 f(\boldsymbol{x_0})(\nabla^2 f(\boldsymbol{x_0})^{-1} \nabla f(\boldsymbol{x_0})) \end{aligned} \tag{6}$$

因此我们得到牛顿减量的表达形式。$\lambda_f(x)$ 越小，说明当前点附近的二次模型逼近越可信，算法使用 λ_f 来判断是否终止迭代。在本节中，我们使用 $\lambda_t = \lambda_f(x_t)$ 的记号来代表牛顿减量。下面我们介绍在二次规划问题中，Sharpened-BFGS 算法的更新结果，收敛速度（牛顿减量 λ_t 是如何收敛到零的）。

Lemma 1. 对于形如式 (1) 的二次规划问题和步长为 1 的拟牛顿法产生的迭代序列，我们有：

$$\lambda_{t+1} = \theta(A, G_t, x_{t+1} - x_t)\lambda_t \tag{7}$$

其中，

$$\theta(A, G, u) := \left(\frac{u^\top (G - A)A^{-1}(G - A)u}{u^\top GA^{-1}Gu}\right)^{\frac{1}{2}}. \tag{8}$$

该引理通过收缩因子 θ 量化了相邻两次迭代的牛顿减量之间的关系。观察收缩因子 θ 的表达式，反映了矩阵 G 和 A 在非零方向 u 上的相似程度，用于评估迭代方向上 Hessian 矩阵的改进性能。注意到连接两次牛顿减量的收缩因子表达式中的 $x_{t+1} - x_t$ 项，它强调牛顿减量收敛的因子与 $G_t(x_{t+1} - x_t), A(x_{t+1} - x_t)$ 之间的差距有关。

由于我们关心的是收敛速度问题，因此下面的定理给出 Sharpened-BFGS 算法下收缩因子的上界。

Theorem 1. 使用 Sharpened-BFGS 算法求解二次规划问题，则：

$$\theta(A, G_t, x_{t+1} - x_t) \leq 1 - \frac{\mu}{L}, \qquad \forall t \geq 0, \tag{9}$$

因此我们有：

$$\lambda_t \leq \left(1 - \frac{\mu}{L}\right)^t \lambda_0, \qquad \forall t \geq 0. \tag{10}$$

该定理指出，在每次迭代过程中收缩因子的上界与问题参数 μ, L 有关 (μ 强凸，L 光滑)，因此牛顿减量 λ_t 满足线性速率的衰减。虽然能够以线性速率收敛，但是由于这个上界仅仅由 G_t, A 的特征值都有界的性质得出，因此这个估计可能比实际的上限更加宽松。在接下来的引理和定理中，我们可以看到收缩因子序列最终收敛于零，Sharpened-BFGS 算法的收敛速度呈超线性。

Lemma 2. 考虑 Sharpened-BFGS 算法求解二次规划问题，进一步定义 $\theta_t := \theta(A, G_t, x_{t+1} - x_t)$,$\sigma_t := \sigma(A, G_t)$。则对于任意 $t \geq 0$：

$$\sigma_{t+1} \leq \left(1 - \frac{\mu}{dL}\right)\left(\sigma_t - \theta_t^2\right) \tag{11}$$

并且我们有：

$$\sum_{i=0}^{t-1} \frac{\theta_i^2}{(1 - \frac{\mu}{dL})^i} \leq \sigma_0, \qquad \forall t \geq 1. \tag{12}$$

回忆贪心 BFGS 算法中引入的函数 $\sigma(A, G)$，它可以用来衡量矩阵 A, G 的距离 (相似程度)。该引理证明了 Sharpened-BFGS 算法中的 σ_t 与贪心 BFGS 算法相比具

有更快的向零收敛的速率。值得注意的是，引理也给出了收缩因子随迭代轮次增加的上界变化情况，整个 θ_t 序列最终收敛到零。因此与前述定理相比，收缩因子具有更加紧密的上界。由此我们可以得到牛顿减量超线性收敛的结论，由下面的定理给出。

Theorem 2. *若使用 Sharpened-BFGS 算法求解二次规划问题，那么对于 $t \geq 1$ 的情况：*

$$\lambda_t \leq \left(1 - \frac{\mu}{dL}\right)^{\frac{t(t-1)}{4}} \left(\frac{dL}{t\mu}\right)^{\frac{t}{2}} \lambda_0. \tag{13}$$

观察牛顿减量收缩的表达式，其中 $\left(1 - \frac{\mu}{dL}\right)^{\frac{t(t-1)}{4}}$ 线性收敛，而 $\left(\frac{dL}{t\mu}\right)^{\frac{t}{2}}$ 在 $t > d\frac{L}{\mu}$ 时 (即底数小于 1 时) 超线性收敛。因此，对于二次规划问题，Sharpened-BFGS 算法在迭代次数小于 $d\frac{L}{\mu}$ 时，以线性速度收敛，而在迭代次数大于 $d\frac{L}{\mu}$ 时，以超线性速度收敛，λ_t 以 $\mathcal{O}((1 - \frac{\mu}{dL})^{t^2}(\frac{dL}{\mu t})^t)$ 的速率收敛到零。

4.2 一般的强凸光滑场景

在本节中，我们将 Sharpened-BFGS 算法扩展到一般的情形。为了建立算法的超线性收敛速率，我们需要如下的两个假设。

Assumption 4.1. *目标函数 f 二阶可微。f 是强凸函数，强凸参数为 $\mu > 0$, 并且 f 的梯度 ∇f 是 Lipschitz 连续的，Lipschitz 参数为 $L > 0$。*

Assumption 4.2. *目标函数 f 是具有参数 M 的强自和谐函数。即对于任意 $x, y, z, w \in \mathbb{R}^n$, 我们有：*

$$\nabla^2 f(y) - \nabla^2 f(x) \preceq M\|y - x\|_z \nabla^2 f(w) \tag{14}$$

其中 $\|y - x\|_z := \sqrt{(y - x)^\top \nabla^2 f(z)(y - x)}$

假设二中引入的强自和谐函数是为了分析拟牛顿法的二次收敛速率。观察它的表达式，$\nabla^2 f(y) - \nabla^2 f(x)$ 表示 f 在 y 和 x 处的 Hessian 矩阵的差异，即函数在这两点附近局部性质的变化。$\|y - x\|_z$ 表示向量 $y - x$ 通过 z 点 Hessian 矩阵的作用后产生变化的范数。该范数的引入确保了函数在点 x, y 之间的局部变化受到 Hessian 矩阵的适当控制。这个性质使得牛顿法在局部收敛时更为有效，可以证明算法在逼近最优解时的收敛速度至少是二次的。

下面给出一般的 Sharpened-BFGS 算法的伪代码:

Algorithm. 2 General Sharpened-BFGS

Require: 初始化变量 x_0, 初始化近似矩阵 $G_0 = LI$

for $t = 0, 1, 2, \cdots$ **do**

更新变量 $x_{t+1} = x_t - G_t^{-1}\nabla f(x_t)$;

计算 BFGS 方向 $s_t = x_{t+1} - x_t$;

计算沿 s_t 方向的平均海森矩阵, 作为 G_t 需要近似的值 $J_t = \int_0^1 \nabla^2 f(x_t + \tau s_t)d\tau$;

计算 $\bar{G}_t = BFGS(J_t, G_t, s_t)$;

计算修正项 $r_t = \|x_{t+1} - x_t\|_{x_t}$;

计算 $\hat{G}_t = (1 + Mr_t/2)^2\bar{G}_t$;

计算贪心 BFGS 方向 $\bar{u} = \bar{u}(\nabla^2 f(x_{t+1}), \hat{G}_t)$;

估计海森矩阵 $G_{t+1} = BFGS(\nabla^2 f(x_{t+1}), \hat{G}_t, \bar{u})$;

end for

算法的整体流程和 4.1 节所述的二次规划问题类似，都是在每一轮迭代中进行两次 BFGS 的更新估计，其中第一次 BFGS 得到的矩阵用来贪婪计算下降方向。与二次规划中的 Sharpened-BFGS 不同的地方在于，一般情形的算法在两次 BFGS 更新之间添加了修正项 $r_t = \|x_{t+1} - x_t\|_{x_t}$。由于在一般情形中，函数的 Hessian 矩阵并不是固定的 (二次规划中 Hessian 矩阵始终为 A)，我们并不能保证 $\nabla^2 f(x) \preceq G$ 始终成立，而只有当 $\nabla^2 f(x) \preceq G$ 时，$\sigma(\nabla^2 f(x), G)$ 才是良定义的。因此我们需要添加修正项确保在经过一次 BFGS 更新之后，新的点 x_+ 和新的 Hessian 逼近矩阵 G_+ 仍满足 $\nabla^2 f(x_+) \preceq G_+$。

Remark 1. 我们需要考虑算法的迭代计算成本是否因为修正项的增加而改变。注意到二次规划中的 Sharpened-BFGS 每轮迭代的计算成本为 $\mathcal{O}(d^2)$。而计算修正向量 r_t 和修正矩阵 $\hat{G}_t$ 的成本也是 $\mathcal{O}(d^2)$，因此修正项的引入后算法每轮迭代的计算成本并没有发生变化。

对一般情形的 Sharpened-BFGS 算法收敛速率的分析大体上与二次规划的情形类似。但是正如上文所述，一般情形下目标函数的 Hessian 矩阵会发生变化，这是我们需要考虑的地方。除此之外，需要注意的是在一般情形下，只有当初始点在最优解的局部邻域内，我们才能保证算法的收敛性。和二次规划问题相似，接下来一步步给出一般情形下 Sharpened-BFGS 算法收敛速率的结论。首先我们给出牛顿减量的收缩因子。

Lemma 3. 考虑满足假设 4.1 和 4.2 的目标函数以及步长为 1 的拟牛顿法产生的迭代序列，我们有：

$$\lambda_{t+1} \leq \left(1 + \frac{Mr_t}{2}\right) \theta(J_t, G_t, x_{t+1} - x_t)\lambda_t, \tag{15}$$

其中 $J_t := \int_0^1 \nabla^2 f(x_t + \tau(x_{t+1} - x_t))d\tau$, $r_t := \|x_{t+1} - x_t\|_{x_t}$。

在给出了相邻两次迭代牛顿减量的关系之后，我们需要考虑收缩因子 θ 的上界。与二次规划的步骤一致，下面的定理先给出较为宽松的上界。

Theorem 3. 考虑一般情形下的 Sharpened-BFGS 算法, 以及满足假设 4.1 和 4.2 的目标函数。假设初始点 x_0 满足：

$$\lambda_0 \leq \frac{C_0\mu}{ML} \tag{16}$$

其中 $C_0 = \frac{1}{4}\ln\frac{3}{2}$，则对任意 $t \geq 0$，我们有：

$$\theta(J_t, G_t, x_{t+1} - x_t) \leq 1 - \frac{2\mu}{3L}, \tag{17}$$

因此得出收缩因子的上界：

$$\lambda_t \leq (1 - \frac{\mu}{2L})^t \lambda_0. \tag{18}$$

由于限制了初始点的牛顿减量，因此上述定理给出的是在最优解的局部邻域内，Sharpened-BFGS 算法以 $1 - \frac{\mu}{2L}$ 的线性速率收敛。接下来的引理和定理将给出收缩因子更精确的上界，使得算法在一般情形下也能达到超线性的收敛速率。

Lemma 4. 考虑一般情形下的 Sharpened-BFGS 算法, 以及满足假设 4.1 和 4.2 的目标函数。假设初始点 x_0 满足：

$$\lambda_0 \leq \frac{C_0\mu}{ML} \tag{19}$$

其中 $C_0 = \frac{1}{4}\ln\frac{3}{2}$。定义 $\theta_t := \theta(\nabla^2 f(x_t), G_t, x_{t+1} - x_t)$，$\sigma_t := \sigma(\nabla^2 f(x_t), G_t)$. 则对任意 $t \geq 0$，我们有：

$$\sigma_{t+1} \leq (1 - \frac{\mu}{2dL}) \left[(1 + \frac{M\lambda_t}{2})^4(\sigma_t + 4Md\lambda_t) - \frac{1}{4}\theta_t^2\right]. \tag{20}$$

以及：

$$\sum_{i=0}^{t-1} \frac{\theta_i^2}{(1 - \frac{\mu}{2dL})^i} \leq 8(\sigma_0 + 4Md\lambda_0), \qquad \forall t \geq 1. \tag{21}$$

上述引理也是建立在最优解的局部邻域之内。具体而言，式 (20) 说明只要 λ_0 足够小 (初始点位于最优解的局部邻域)，那么 σ_t 可以收敛到零，即 G_t 能够很好地近似实际的 Hessian 矩阵 $\nabla^2 f(x_t)$，并且由于 θ_t^2 的存在，其收敛速率快于贪心 BFGS

算法。式 (21) 给出收缩因子 θ 更精确的上界，保证其能够以更快的速度收敛到零。有了上述的结论，我们可以得出如下的定理，给出一般情形下 Sharpened-BFGS 算法的收敛速率。

Theorem 4. *考虑一般情形下的 Sharpened-BFGS 算法, 以及满足假设 4.1 和 4.2 的目标函数。假设初始点 x_0 满足：*

$$\lambda_0 \leq \frac{C_1 \mu}{dML}, \tag{22}$$

其中 $C_1 = \frac{\ln 2}{20}$。那么对于任意的 $t \geq 1$，我们有：

$$\lambda_t \leq 2\left(1 - \frac{\mu}{2dL}\right)^{\frac{t(t-1)}{4}} \left(\frac{8dL}{t\mu}\right)^{\frac{t}{2}} \lambda_0. \tag{23}$$

类似于二次规划问题的分析，第一项 $\left(1 - \frac{\mu}{2dL}\right)^{\frac{t(t-1)}{4}}$ 线性收敛，第二项 $\left(\frac{8dL}{t\mu}\right)^{\frac{t}{2}}$ 在 $t \geq \frac{8dL}{t\mu}$ 时超线性收敛。总结来说，当迭代轮数 $t \leq \Theta(d\frac{L}{\mu})$ 时，算法具有线性收敛速率，当迭代轮数 $t \geq \Theta(d\frac{L}{\mu})$ 时，算法可以达到超线性的收敛速率。至此我们介绍了 Sharpened-BFGS 算法的具体思路方法，以及收敛性分析的推导过程。

5 理论结果

这一部分我们将 Sharpened-BFGS 的收敛结果与第三部分的经典 BFGS 和贪心 BFGS 进行比较。我们特别关注目标函数满足假设 4.1 和 4.2 的情况。为了简化比较，除了参数 μ, L, M, d 之外，其它的参数我们用常数 1 来替代，并且定义条件数 $\kappa = L/\mu \geq 1$。

Sharpened-BFGS. 根据第四部分的结果，如果我们设置初始近似矩阵 $G_0 = LI$, 并且初始点 x_0 满足：

$$\lambda_f(x_0) = \mathcal{O}\left(\frac{1}{dM\kappa}\right),$$

那么由 Sharpened-BFGS 算法产生的迭代序列满足：

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \min\left\{\left(1-\frac{1}{\kappa}\right)^t, \left(1-\frac{1}{d\kappa}\right)^{\frac{t(t-1)}{4}}\left(\frac{d\kappa}{t}\right)^{\frac{t}{2}}\right\}.$$

从表达式可以看出，当 $t < d\kappa$ 时，第一个上界项更小，牛顿减量以线性速率 $\left(1-\frac{1}{\kappa}\right)^t$ 收敛。当 $t > d\kappa$ 时，第二个上界项更小，牛顿减量以 $\left(1-\frac{1}{d\kappa}\right)^{\frac{t(t-1)}{4}}\left(\frac{d\kappa}{t}\right)^{\frac{t}{2}}$ 的超线性速率收敛，并且可以看出其收敛速度快于二次收敛。

Greedy-BFGS. 如果我们设置初始近似矩阵 $G_0 = LI$, 并且初始点 x_0 满足：

$$\lambda_f(x_0) = \mathcal{O}\left(\frac{1}{dM\kappa}\right),$$

那么由 Greedy-BFGS 算法产生的迭代序列满足：

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \min\left\{\left(1-\frac{1}{\kappa}\right)^t, \left(1-\frac{1}{d\kappa}\right)^{\frac{t(t-1)}{2}}\left(\frac{1}{2}\right)^t\right\}.$$

将其结果与 Sharpened-BFGS 的结果进行对比，我们可以发现：

1. Greedy-BFGS 算法的迭代轮数达到 $d\kappa \ln(d\kappa)$ 之后，收敛速度达到超线性。这是慢于 Sharpened-BFGS 算法 ($d\kappa$ 轮后达到超线性) 的。

2. 由于当 t 充分大时，除了共同的二次收敛项 $(1-\frac{1}{d\kappa})^{t^2}$ 之外，$(\frac{d\kappa}{t})^{\frac{t}{2}} \ll (\frac{1}{2})^t$, 因此 Sharpened-BFGS 算法最终的超线性收敛速率快于 Greedy-BFGS 算法。

BFGS. 如果我们设置初始近似矩阵 $G_0 = LI$, 并且初始点 x_0 满足：

$$\lambda_f(x_0) = \max\left\{\mathcal{O}\left(\frac{1}{M\kappa}\right), \mathcal{O}\left(\frac{1}{Md\ln\kappa}\right)\right\},$$

那么由 BFGS 算法产生的迭代序列满足：

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \min\left\{\left(1-\frac{1}{\kappa}\right)^t, \left(\frac{d\ln\kappa}{t}\right)^{\frac{t}{2}}\right\}.$$

我们可以看到，BFGS 算法在经过 $d\ln\kappa$ 轮后达到超线性收敛速率，这是快于 Sharpened-BFGS 算法的。但是，当 t 充分大时我们有

$$\left(1-\frac{1}{d\kappa}\right)^{\frac{t(t-1)}{4}}\left(\frac{d\kappa}{t}\right)^{\frac{t}{2}} \ll \left(\frac{d\ln\kappa}{t}\right)^{\frac{t}{2}}.$$

因此 BFGS 算法的超线性收敛速率是慢于 Sharpened-BFGS 算法的。

6 实验结果

7 问题分析与挑战

8 总结

在本篇论文中，作者提出了一种用于解决无约束凸优化问题新的拟牛顿方法——Sharpened-BFGS。其中目标函数具有 μ-强凸性，其梯度具有 L-光滑性，并且是 M-强自和谐的。Sharpened-BFGS 算法充分利用了经典 BFGS 算法在牛顿方向的近似和贪心 BFGS 算法的 Heissan 矩阵近似。利用这些性质，作者证明了该算法达到 $\mathcal{O}((1-\frac{\mu}{dL})^{\frac{t(t-1)}{4}}(\frac{dL}{t\mu})^{\frac{t}{2}})$ 的超线性收敛速率，并且其收敛速度快于二次收敛。作者也

通过理论分析和数值实验将该方法和经典 BFGS 和 Greedy-BFGS 算法的收敛速率进行了比较，凸显了 Sharpened-BFGS 算法的优越性。

在学习这篇论文的工作时，我们体会到了不同优化方法的差异以及如何充分利用它们各自的优势得到表现更好的优化方法。同时，我们在跟着论文一步步推导并最终得到结果的过程中也学习到了如何从理论上分析算法的收敛速度，以及如何通过数值实验来验证理论结果，受益匪浅。

Reference

[] YE H, LIN D, ZHANG Z, et al. Explicit superlinear convergence rates of the sr1 algorithm[J]. arXiv preprintarXiv:2105.07162, 2021.