

近似点梯度法

文再文

北京大学北京国际数学研究中心

教材《最优化：建模、算法与理论》配套电子教案

<http://bicmr.pku.edu.cn/~wenzw/optbook.html>

致谢：本教案由**陈乐恒**、**朱桢源**协助准备

1 近似点梯度法

2 应用

- LASSO问题
- 低秩矩阵恢复
- 小波模型求解

3 收敛性分析

4 拓展

- 非凸函数的近似点梯度法
- 镜像下降算法
- 惯性近似点梯度算法
- 条件梯度法

复合优化问题

我们将考虑如下复合优化问题：

$$\min_{x \in \mathbb{R}^n} \psi(x) = f(x) + h(x)$$

- 函数 f 为可微函数，其定义域 $\text{dom } f = \mathbb{R}^n$
- 函数 h 为凸函数，可以是非光滑的，并且邻近算子容易计算
- LASSO问题： $f(x) = \frac{1}{2} \|Ax - b\|^2$ ， $h(x) = \mu \|x\|_1$
- 次梯度法计算的复杂度： $\mathcal{O}(1/\sqrt{k})$

是否可以设计复杂度为 $\mathcal{O}(1/k)$ 的算法？

邻近算子：回顾

定义

对于一个凸函数 h ，定义它的邻近算子为

$$\text{prox}_h(x) = \underset{u}{\operatorname{argmin}} \left(h(u) + \frac{1}{2} \|u - x\|_2^2 \right)$$

例子

- ℓ_1 范数: $h(x) = \|x\|_1$, $\text{prox}_{th}(x) = \operatorname{sign}(x) \max\{|x| - t, 0\}$
- ℓ_2 范数: $h(x) = \|x\|_2$, $\text{prox}_{th}(x) = \begin{cases} \left(1 - \frac{t}{\|x\|_2}\right) x, & \|x\|_2 \geq t, \\ 0, & \text{其他.} \end{cases}$
- 二次函数(其中 A 对称正定):
 $h(x) = \frac{1}{2} x^T A x + b^T x + c$, $\text{prox}_{th}(x) = (I + tA)^{-1}(x - tb)$
- 负自然对数的和:
 $h(x) = -\sum_{i=1}^n \ln x_i$, $\text{prox}_{th}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}$, $i = 1, 2, \dots, n$

近似点梯度法

对于光滑部分 f 做梯度下降，对于非光滑部分 h 使用邻近算子，则近似点梯度法的迭代格式为

$$x^{k+1} = \text{prox}_{t_k h} (x^k - t_k \nabla f (x^k)) \quad (1)$$

其中 $t_k > 0$ 为每次迭代的步长，它可以是一个常数或者由线搜索得出。

Algorithm 1 近似点梯度法

- 1: 输入：函数 $f(x)$, $h(x)$, 初始点 x^0 .
 - 2: **while** 未达到收敛准则 **do**
 - 3: $x^{k+1} = \text{prox}_{t_k h} (x^k - t_k \nabla f (x^k))$.
 - 4: **end while**
-

对近似点梯度法的理解

根据定义, (1)式等价于

$$\begin{aligned}x^{k+1} &= \arg \min_u \left\{ h(u) + \frac{1}{2t_k} \|u - x^k + t_k \nabla f(x^k)\|^2 \right\} \\ &= \arg \min_u \left\{ h(u) + f(x^k) + \nabla f(x^k)^\top (u - x^k) + \frac{1}{2t_k} \|u - x^k\|^2 \right\}\end{aligned}$$

根据邻近算子与次梯度的关系, 又可以形式地写成

$$x^{k+1} = x^k - t_k \nabla f(x^k) - t_k g^k, \quad g^k \in \partial h(x^{k+1}).$$

即对光滑部分做显式的梯度下降, 关于非光滑部分做隐式的梯度下降.

步长选取

- 当 f 为梯度 L -利普希茨连续函数时, 可取固定步长 $t_k = t \leq \frac{1}{L}$. 当 L 未知时可使用线搜索准则

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{1}{2t_k} \|x^{k+1} - x^k\|^2$$

- 利用 **BB** 步长作为 t_k 的初始估计并用非单调线搜索进行校正:

$$\alpha_{\text{BB1}}^k \stackrel{\text{def}}{=} \frac{(s^{k-1})^T y^{k-1}}{(y^{k-1})^T y^{k-1}} \quad \text{或} \quad \alpha_{\text{BB2}}^k \stackrel{\text{def}}{=} \frac{(s^{k-1})^T s^{k-1}}{(s^{k-1})^T y^{k-1}},$$

其中 $s^{k-1} = x^k - x^{k-1}$ 以及 $y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$.

- 可构造如下适用于近似点梯度法的非单调线搜索准则:

$$\psi(x^{k+1}) \leq C^k - \frac{c_1}{2t_k} \|x^{k+1} - x^k\|^2,$$

$c_1 \in (0, 1)$ 为正常数. 注意, 定义 C^k 时需要使用整体函数值 $\psi(x^k)$.

1 近似点梯度法

2 应用

- LASSO问题
- 低秩矩阵恢复
- 小波模型求解

3 收敛性分析

4 拓展

- 非凸函数的近似点梯度法
- 镜像下降算法
- 惯性近似点梯度算法
- 条件梯度法

LASSO问题

考虑用近似点梯度法求解 LASSO 问题

$$\min_x \mu \|x\|_1 + \frac{1}{2} \|Ax - b\|^2.$$

令 $f(x) = \frac{1}{2} \|Ax - b\|^2$, $h(x) = \mu \|x\|_1$, 则

$$\begin{aligned} \nabla f(x) &= A^T(Ax - b), \\ \text{prox}_{t_k h}(x) &= \text{sign}(x) \max\{|x| - t_k \mu, 0\}. \end{aligned}$$

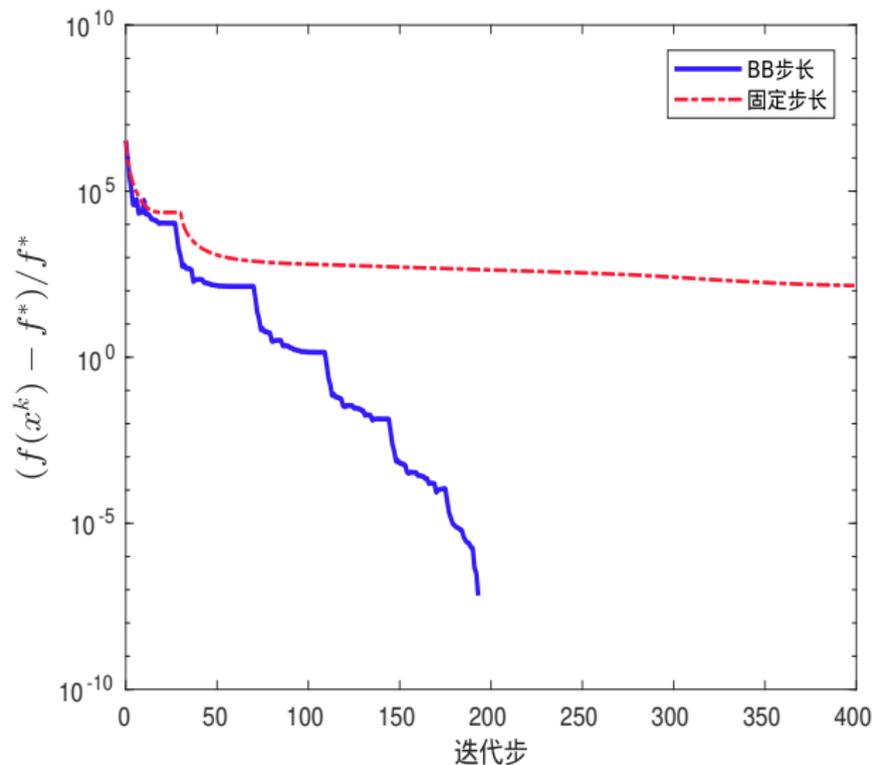
故相应的迭代格式为：

$$\begin{aligned} y^k &= x^k - t_k A^T (Ax^k - b), \\ x^{k+1} &= \text{sign}(y^k) \max\{|y^k| - t_k \mu, 0\}, \end{aligned}$$

即第一步做梯度下降，第二步做收缩

LASSO问题

我们还可以使用BB步长加速收敛



低秩矩阵恢复

考虑低秩矩阵恢复模型:

$$\min_{X \in \mathbb{R}^{m \times n}} \mu \|X\|_* + \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2,$$

其中 M 是想要恢复的低秩矩阵, 但是只知道其在下标集 Ω 上的值. 令

$$f(X) = \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2, \quad h(X) = \mu \|X\|_*.$$

定义矩阵 $P \in \mathbb{R}^{m \times n}$:

$$P_{ij} = \begin{cases} 1, & (i,j) \in \Omega, \\ 0, & \text{其他,} \end{cases}$$

则

$$f(X) = \frac{1}{2} \|P \odot (X - M)\|_F^2$$

低秩矩阵恢复

进一步可以得到

$$\begin{aligned}\nabla f(X) &= P \odot (X - M), \\ \text{prox}_{t_k h}(X) &= U \text{Diag}(\max\{|d| - t_k \mu, 0\}) V^T,\end{aligned}$$

其中 $X = U \text{Diag}(d) V^T$ 为矩阵 X 的约化的奇异值分解.

由此可以得到近似点梯度法的迭代格式:

$$\begin{aligned}Y^k &= X^k - t_k P \odot (X^k - M) \\ X^{k+1} &= \text{prox}_{t_k h}(Y^k)\end{aligned}$$

小波模型求解

考虑小波分解模型：

$$\min_u \|\lambda \odot (Wu)\|_1 + \frac{1}{2} \|Au - b\|^2,$$

其中 W 是紧小波框架算子，即满足 $W^T W = I$ ，第二项为问题的损失函数。引入 $d = Wu$ ，则 $u = W^T d$ ，即有合成模型：

$$\min_d \|\lambda \odot d\|_1 + \frac{1}{2} \|AW^T d - b\|^2. \quad (2)$$

此外还有平衡小波模型：

$$\min_{\alpha} \|\lambda \odot \alpha\|_1 + \frac{\kappa}{2} \|(I - WW^T)\alpha\|^2 + \frac{1}{2} \|AW^T \alpha - b\|^2 \quad (3)$$

其中不要求 W 是紧框架

小波模型求解

对于合成模型(2)

$$\min_d \quad \|\lambda \odot d\|_1 + \frac{1}{2} \|AW^T d - b\|^2$$

今 $f(d) = \frac{1}{2} \|AW^T d - b\|^2$, $h(d) = \|\lambda \odot d\|_1$, 则

$$\begin{aligned}\nabla f(d) &= WA^T (AW^T d - b) \\ \text{prox}_{t_k h}(d) &= \text{sign}(d) \max\{|d| - t_k \lambda, 0\}\end{aligned}$$

近似点梯度算法的迭代格式为：

$$\begin{aligned}y^k &= d^k - t_k WA^T (AW^T d^k - b) \\ d^{k+1} &= \text{sign}(y^k) \max\{|y^k| - t_k \lambda, 0\}\end{aligned}$$

小波模型求解

对于平衡小波模型(3)

$$\min_{\alpha} \quad \|\lambda \odot \alpha\|_1 + \frac{\kappa}{2} \|(I - WW^T) \alpha\|^2 + \frac{1}{2} \|AW^T \alpha - b\|^2$$

我们令

$$f(\alpha) = \frac{\kappa}{2} \|(I - WW^T) \alpha\|^2 + \frac{1}{2} \|AW^T \alpha - b\|^2, \quad h(\alpha) = \|\lambda \odot \alpha\|_1$$

则

$$\begin{aligned} \nabla f(\alpha) &= \kappa (I - WW^T) \alpha + WA^T (AW^T \alpha - b) \\ \text{prox}_{t_k h}(\alpha) &= \text{sign}(\alpha) \max \{|\alpha| - t_k \lambda, 0\} \end{aligned}$$

近似点梯度算法的迭代格式为：

$$\begin{aligned} y^k &= \alpha^k - t_k (\kappa (I - WW^T) \alpha^k + WA^T (AW^T \alpha^k - b)) \\ \alpha^{k+1} &= \text{sign}(y^k) \max \{|y^k| - t_k \lambda, 0\} \end{aligned}$$

1 近似点梯度法

2 应用

- LASSO问题
- 低秩矩阵恢复
- 小波模型求解

3 收敛性分析

4 拓展

- 非凸函数的近似点梯度法
- 镜像下降算法
- 惯性近似点梯度算法
- 条件梯度法

收敛性分析

基本假设：

- f 在 \mathbb{R}^n 上是凸的； ∇f 为 L -利普希茨连续，即

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

- h 是适当的闭凸函数 (因此 prox_{th} 的定义是合理的)；
- 函数 $\psi(x) = f(x) + h(x)$ 的最小值 ψ^* 是有限的，并且在点 x^* 处可取到 (并不要求唯一)。

梯度映射

在基本假设的基础上，我们定义**梯度映射**：

$$G_t(x) = \frac{1}{t} (x - \text{prox}_{th}(x - t\nabla f(x))) \quad (t > 0) \quad (4)$$

不难推出梯度映射具有以下性质：

- “负搜索方向”： $x^{k+1} = \text{prox}_{th}(x^k - t\nabla f(x^k)) = x^k - tG_t(x^k)$
- 根据邻近算子和次梯度的关系，我们有

$$G_t(x) - \nabla f(x) \in \partial h(x - tG_t(x)) \quad (5)$$

- 与算法的收敛性的关系：

$$G_t(x) = 0 \iff x \text{ 为 } \psi(x) = f(x) + h(x) \text{ 的最小值点}$$

收敛性分析

$G_t(x) = 0 \iff x$ 为 $\psi(x) = f(x) + h(x)$ 的最小值点

证明:

$$\begin{aligned} 0 \in \nabla f(x) + \partial h(x) &\iff x - t\nabla f(x) \in x + t\partial h(x) \\ &\iff x - (x - t\nabla f(x)) \in t\partial h(x) \\ &\iff x = \text{prox}_{th}(x - t\nabla f(x)) \\ &\iff G_t(x) = 0. \end{aligned}$$

定理 1 (固定步长近似点梯度法的收敛性)

取定步长为 $t_k = t \in (0, \frac{1}{L}]$, 设 $\{x^k\}$ 由迭代格式(1)产生, 则

$$\psi(x^k) - \psi^* \leq \frac{1}{2kt} \|x^0 - x^*\|^2$$

定理证明

证明 根据利普希茨连续“二次上界”的性质，得到

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n$$

令 $y = x - tG_t(x)$, 有

$$\begin{aligned} f(x - tG_t(x)) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t^2 L}{2} \|G_t(x)\|^2 \\ &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|^2 \end{aligned} \quad (6)$$

此外，由 $f(x), h(x)$ 为凸函数，结合(4)式我们有

$$h(x - tG_t(x)) \leq h(z) - (G_t(x) - \nabla f(x))^T (z - x + tG_t(x)) \quad (7)$$

$$f(x) \leq f(z) - \nabla f(x)^T (z - x) \quad (8)$$

将(6)(7)(8)式相加可得对任意 $z \in \text{dom } \psi$ 有

$$\psi(x - tG_t(x)) \leq \psi(z) + G_t(x)^T (x - z) - \frac{t}{2} \|G_t(x)\|^2 \quad (9)$$

定理证明

由 $x^i = x^{i-1} - tG_t(x^{i-1})$, 在不等式(9)中, 取 $z = x^*$, $x = x^{i-1}$ 得到

$$\begin{aligned}\psi(x^i) - \psi^* &\leq G_t(x^{i-1})^\top (x^{i-1} - x^*) - \frac{t}{2} \|G_t(x^{i-1})\|^2 \\ &= \frac{1}{2t} \left(\|x^{i-1} - x^*\|^2 - \|x^{i-1} - x^* - tG_t(x^{i-1})\|^2 \right) \\ &= \frac{1}{2t} \left(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2 \right)\end{aligned}\quad (10)$$

取 $i = 1, 2, \dots, k$ 并累加, 得

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

定理证明

注意到在不等式(9)中, 取 $z = x^{i-1}$ 即得:

$$\psi(x^i) \leq \psi(x^{i-1}) - \frac{t}{2} \|G_t(x^{i-1})\|^2$$

即 $\psi(x^i)$ 不增, 因此

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

步长选取

定理1中要求 $t \leq \frac{1}{L}$ ，而根据定理证明的过程，也可以用线搜索准则：

- 从某个 $t = \hat{t} > 0$ 开始进行回溯 ($t \leftarrow \beta t$)，直到满足不等式

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|^2 \quad (11)$$

- 这等价于算法部分提到的线搜索准则：

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{1}{2t_k} \|x^{k+1} - x^k\|^2$$

至此我们解释了该线搜索准则的合理性

收敛性分析

定理 2 (非固定步长的近似点梯度法的收敛性)

从某个 $t = \hat{t} > 0$ 开始进行回溯 ($t \leftarrow \beta t$) 直到满足不等式(11), 设 $\{x^k\}$ 是由迭代格式(1) 产生的序列, 则

$$\psi(x^k) - \psi^* \leq \frac{1}{2k \min\{\hat{t}, \beta/L\}} \|x^0 - x^*\|^2$$

Proof.

由定理1的证明, 当 $0 < t \leq \frac{1}{L}$ 时, 不等式(11)成立, 故由线搜索所得的步长 t 应满足 $t \geq t_{\min} = \min\left\{\hat{t}, \frac{\beta}{L}\right\}$. 同理, 我们有 $\psi(x^i)$ 单调不增, 且

$$\psi(x^i) - \psi^* \leq \frac{1}{2t_{\min}} \left(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2 \right)$$

取 $i = 1, 2, \dots, k$ 并累加, 并利用 $\psi(x^i)$ 不增, 可得

$$\psi(x^k) - \psi^* \leq \frac{1}{2kt_{\min}} \|x^0 - x^*\|^2$$



1 近似点梯度法

2 应用

- LASSO问题
- 低秩矩阵恢复
- 小波模型求解

3 收敛性分析

4 拓展

- 非凸函数的近似点梯度法
- 镜像下降算法
- 惯性近似点梯度算法
- 条件梯度法

非凸函数的近似点梯度法

(适当闭函数的邻近算子) 设 h 是适当闭函数(可以非凸), 且具有有限的下界, 即满足 $\inf_{x \in \text{dom} h} h(x) > -\infty$, 定义 h 的邻近算子为

$$\text{prox}_h(x) = \arg \min_{u \in \text{dom} h} \left\{ h(u) + \frac{1}{2} \|u - x\|^2 \right\}.$$

- $\text{prox}_h(x)$ 良定义, 且是 \mathbb{R}^n 上的非空紧集
- 对 $u \in \text{prox}_h(x)$, 有 $x - u \in \partial h(u)$. ∂h 表示 h (包括非凸情形) 的次微分

对于复合优化问题 $\min \psi(x) = f(x) + h(x)$, f 可微, h 为适当闭函数(可非凸)。与一般的近似点梯度法类似, 有迭代格式:

$$x^{k+1} = \text{prox}_{t_k h} (x^k - t_k \nabla f(x^k))$$

迭代时往往选取 $\text{prox}_{t_k h}$ 中的一个元素, 此时算法也具有收敛性。

镜像下降算法

考虑如下的凸优化问题：

$$\min f(x), \quad \text{s.t. } x \in C,$$

f 为凸函数， C 是 $\text{dom } f$ 的凸子集，且 f 在 C 上存在次梯度。

Bregman距离：令 r 为可微凸函数，由 r 产生的 **Bregman距离** 为：

$$D_r(y, x) = r(y) - r(x) - \nabla r(x)^T (y - x)$$

镜像(非线性)次梯度方法：

- 1 取次梯度 $g^k \in \partial f(x^k)$
- 2 更新迭代格式：

$$x^{k+1} = \operatorname{argmin}_{x \in C} \left\{ (g^k)^T (x - x^k) + \frac{1}{\alpha_k} D_r(x, x^k) \right\}.$$

取 $r(x) = \frac{1}{2} \|x\|_2^2$ 时，这就是投影次梯度法。

收敛性分析

函数 r 需要满足的性质: 在范数 $\|\cdot\|$ 的意义下强凸, 即

$$r(y) \geq r(x) + \nabla r(x)^T(y - x) + \frac{1}{2}\|x - y\|^2.$$

考虑计算与任意点(包括最优点)处函数值的距离: 对于任意的 $x^* \in C$,

$$\begin{aligned} f(x^k) - f(x^*) &\leq (g^k)^T(x^k - x^*) \\ &= (g^k)^T(x^{k+1} - x^*) + (g^k)^T(x^k - x^{k+1}) \end{aligned}$$

根据 x^{k+1} 点处最优性的必要条件:

$$(\alpha_k g^k + \nabla r(x^{k+1}) - \nabla r(x^k))^T(y - x^{k+1}) \geq 0, \forall y \in C$$

因此, 我们取 $y = x^*$, 得

$$(g^k)^T(x^{k+1} - x^*) \leq \frac{1}{\alpha_k}(\nabla r(x^{k+1}) - \nabla r(x^k))^T(x^* - x^{k+1})$$

收敛性分析(续)

进一步,

$$\begin{aligned} & (\nabla r(x^{k+1}) - \nabla r(x^k))^T (x^* - x^{k+1}) \\ &= D_r(x^*, x^k) - D_r(x^*, x^{k+1}) - D_r(x^k, x^{k+1}) \end{aligned}$$

综合以上各式, 对任意 $x^* \in C$,

$$\begin{aligned} f(x^k) - f(x^*) &\leq (g^k)^T (x^{k+1} - x^*) + (g^k)^T (x^k - x^{k+1}) \\ &\leq \frac{1}{\alpha_k} [D_r(x^*, x^k) - D_r(x^*, x^{k+1})] - \frac{1}{\alpha_k} D_r(x^k, x^{k+1}) \\ &\quad + (g^k)^T (x^k - x^{k+1}). \end{aligned}$$

应用Fenchel-Young不等式 $(x^T y \leq \frac{1}{2\alpha} \|x\|^2 + \frac{\alpha}{2} \|y\|_*^2)$

$$\begin{aligned} &\leq \frac{1}{\alpha_k} [D_r(x^*, x^k) - D_r(x^*, x^{k+1})] - \frac{1}{\alpha_k} D_r(x^k, x^{k+1}) \\ &\quad + \frac{\alpha_k}{2} \|g^k\|_*^2 + \frac{1}{2\alpha_k} \|x^k - x^{k+1}\|^2 \\ &\leq \frac{1}{\alpha_k} [D_r(x^*, x^k) - D_r(x^*, x^{k+1})] + \frac{\alpha_k}{2} \|g^k\|_*^2 \end{aligned}$$

保证收敛的条件

取固定步长 $\alpha_k = \alpha$, 并在上面的不等式中对 $1, \dots, k$ 求和, 得

$$\frac{1}{k} \sum_{i=1}^k f(x^i) - f(x^*) \leq \frac{1}{\alpha k} D_h(x^*, x^1) + \frac{\alpha}{2} \max_i \|g^i\|_*^2$$

一般而言,

$$f^{\text{best}, k} - f^* \leq \frac{D_h(x^*, x^1) + \frac{1}{2} \sum_{i=1}^k \alpha_i^2 \max_i \|g^i\|_*^2}{\sum_{i=1}^k \alpha_i}$$

当以下条件满足时, 算法可以保证收敛

- $D_h(x^*, x^1) < \infty$
- $\sum_k \alpha_k = \infty$ 且 $\alpha_k \rightarrow 0$ (消失步长)
- 对于任意 $g \in \partial f(x)$ 和 $x \in C$, 有 $\|g\|_* \leq G$ 恒成立, 其中 $G < \infty$

镜像下降算法的例子

- 通常的(投影)次梯度法:取 $r(x) = \frac{1}{2}\|x\|_2^2$
- 使用单纯形约束, $C = \{x \in \mathbb{R}_+^n \mid \mathbf{1}^T x = 1\}$, 并使用负熵函数

$$h(x) = \sum_{i=1}^n x_i \log x_i$$

- 1 ℓ_1 范数意义下为强凸函数
- 2 对于初始点 $x^1 = 1/n$, 有 $D_r(x^*, x^1) \leq \log n$ 对任意 $x^* \in C$ 成立
- 3 若 $G_\infty \geq \|g\|_\infty$ 对任意 $g \in \partial f(x)$, $x \in C$ 成立, 即存在有限上界, 则

$$f_{best}^k - f^* \leq \frac{\log n}{\alpha k} + \frac{\alpha}{2k} G_\infty$$

- 4 比通常的次梯度算法表现好很多

惯性近似点梯度算法

考虑复合优化问题：

$$\min_{x \in \mathbb{R}^n} \psi(x) = f(x) + h(x)$$

其中 $f(x)$ 为可微函数且 $\nabla f(x)$ 是 L -利普希茨连续的； $h(x)$ 为凸函数

- 选取初始点 x^0 ，令 $x^{-1} = x^0$ ，取 $\beta \in [0, 1]$ ，令 $\alpha < 2(1 - \beta)/L$ ，则惯性近似点梯度法的迭代格式为：

$$x^{k+1} = \text{prox}_{\alpha h} (x^k - \alpha \nabla f(x^k) + \beta (x^k - x^{k-1}))$$

- $\beta (x^k - x^{k-1})$ 表示惯性项
- 对于 $h(x) = 0$ 情形，该算法也被称为重球法(Heavy-ball)

条件梯度法: Motivation

设 C 为紧集, 考虑优化问题

$$\min_{x \in C} f(x)$$

- 如果使用近似点梯度法, 则

$$x_{k+1} = \operatorname{argmin}_{x \in C} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

这等价于投影梯度法:

$$x_{k+1} = \mathcal{P}_C(x_k - \alpha_k \nabla f(x_k))$$

- 困难: $\mathcal{P}_C(\cdot)$ 可能具有昂贵的计算代价

条件梯度法 (CndG or Frank-Wolfe Method)

- 给定 $y_0 = x_0$, 以及 $\alpha_k \in (0, 1]$. 条件梯度法的迭代格式为:

$$\begin{aligned}x_k &= \operatorname{argmin}_{x \in C} \langle \nabla f(y_{k-1}), x \rangle, \\y_k &= (1 - \alpha_k)y_{k-1} + \alpha_k x_k\end{aligned}$$

- 考虑步长参数 α_k 的选取

消失步长:

$$\alpha_k = \frac{2}{k+1}$$

或通过精确线搜索:

$$\alpha_k = \operatorname{argmin}_{\alpha \in [0,1]} f((1 - \alpha)y_{k-1} + \alpha x_k)$$

例子

考虑带某一范数 $\|\cdot\|$ 约束的凸优化问题,

$$\min_x f(x) \quad \text{s.t.} \quad \|x\| \leq t.$$

用条件梯度法求解该问题时, 需要计算子问题,

$$\begin{aligned} x_k &\in \operatorname{argmin}_{\|x\| \leq t} \langle \nabla f(y_{k-1}), x \rangle \\ &= -t \cdot \left(\operatorname{argmax}_{\|x\| \leq 1} \langle \nabla f(y_{k-1}), x \rangle \right) \\ &= -t \cdot \partial \|\nabla f(y_{k-1})\|_* . \end{aligned} \tag{12}$$

其中 $\|z\|_* = \sup\{z^T x, \|x\| \leq 1\}$ 是 $\|\cdot\|$ 的对偶范数。注意到(12) 条件梯度法的子问题相当于计算一个对偶范数的次梯度。如果计算 $\|\cdot\|$ 范数的次梯度比计算在约束集合 $C = \{x \in \mathbb{R}^n : \|x\| \leq t\}$ 上的投影要简单, 条件梯度法比投影梯度法效率更高。

例子: l_1 范数约束问题

由于 l_1 范数的对偶范数是 l_∞ 范数, 因此用条件梯度法求解该问题时子问题为:

$$x_k \in -t \cdot \partial \|\nabla f(y_{k-1})\|_\infty.$$

考虑到 l_∞ 范数的次梯度为 $\partial \|x\|_\infty = \{v : \langle v, x \rangle = \|x\|_\infty, \|v\|_1 \leq 1\}$, 子问题等价于,

$$\begin{aligned} i_k &\in \operatorname{argmax}_{i=1, \dots, n} |\nabla_i f(y_{k-1})| \\ x_k &= -t \cdot \operatorname{sgn} [\nabla_{i_k} f(y_{k-1})] \cdot e_{i_k}. \end{aligned}$$

其中 $\nabla_i f(y_{k-1})$ 表示向量 $\nabla f(y_{k-1})$ 的第 i 个元素, e_i 表示第 i 个元素为 1 的单位向量。可以看到计算 $\|\cdot\|_\infty$ 的次梯度和计算集

合 $C := \{x \in \mathbb{R}^n : \|x\|_1 \leq t\}$ 上的投影都需要 $\mathcal{O}(n)$ 的计算复杂度, 但是条件梯度法子问题计算明显要更简单直接。

例子: l_p 范数约束问题, $1 \leq p \leq \infty$

由于 l_p 范数的对偶范数是 l_q 范数, 其中 $1/p + 1/q = 1$, 因此用条件梯度法求解该问题时子问题为,

$$x_k \in -t \cdot \partial \|\nabla f(y_{k-1})\|_q.$$

注意到 l_q 范数的次梯度为 $\partial \|x\|_q = \{v : \langle v, x \rangle = \|x\|_q, \|v\|_p \leq 1\}$, 子问题等价于,

$$x_k^i = -\beta \cdot \text{sgn}[\nabla f(y_{k-1})] \cdot |\nabla f(y_{k-1})|^{p/q}.$$

其中 β 是使得 $\|x_k\|_q = t$ 的归一化常数。可以看到, 除过 $p = 1, 2, \infty$ 这些特殊情形, 条件梯度法的子问题计算复杂度比直接计算点在集合 $C = \{x \in \mathbb{R}^n : \|x\|_p \leq t\}$ 上的投影要简单, 后者投影计算需要单独解一个优化问题。

例子: 矩阵核范数约束优化问题

矩阵核范数 $\|\cdot\|_*$ 的对偶范数是其谱范数 $\|\cdot\|_2$:

$$\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X), \quad \|X\|_2 = \max_{i=1,\dots,\min\{m,n\}} \sigma_i(X).$$

因此条件梯度法的子问题为 $X_k \in -t \cdot \partial \|\nabla f(Y_{k-1})\|_2$. 对矩阵范数的次梯度: $\partial \|X\| = \{Y : \langle Y, X \rangle = \|X\|, \|Y\|_* \leq 1\}$, 设 u, v 分别是矩阵 $\nabla f(Y_{k-1})$ 最大奇异值对应的左、右奇异向量, 注意到,

$$\langle uv^T, \nabla f(Y_{k-1}) \rangle = u^T \nabla f(Y_{k-1}) v = \sigma_{\max}(\nabla f(Y_{k-1})) = \|\nabla f(Y_{k-1})\|_2.$$

且 $\|uv^T\|_* = 1$, 因此矩阵 $uv^T \in \partial \|\nabla f(Y_{k-1})\|_2$. 则条件梯度法子问题等价于,

$$X_k \in -t \cdot uv^T. \quad (13)$$

可以看到, 条件梯度法计算子问题时只需要计算矩阵最大的奇异值对应的左、右奇异向量。如果采用投影梯度法, 其子问题是计算 X 到集合 $\{X \in \mathbb{R}^{m \times n} : \|X\|_* \leq t\}$ 的投影, 需要对矩阵做全奇异值分解, 计算量比条件梯度法复杂很多。

收敛性分析: 引理

令 $\gamma_t \in (0, 1]$, $t = 1, 2, \dots$, 构造序列

$$\Gamma_t = \begin{cases} 1 & t = 1 \\ (1 - \gamma_t)\Gamma_{t-1} & t \geq 2 \end{cases}.$$

如果序列 $\{\Delta_t\}_{t \geq 0}$ 满足

$$\Delta_t \leq (1 - \gamma_t)\Delta_{t-1} + B_t \quad t = 1, 2, \dots$$

则对任意的 k 我们对 Δ_k 有估计

$$\Delta_k \leq \Gamma_k(1 - \gamma_1)\Delta_0 + \Gamma_k \sum_{t=1}^k \frac{B_t}{\Gamma_t}.$$

收敛性分析

令 $f(x)$ 是凸函数, $\nabla f(x)$ 是 L -利普希茨的, $D_C = \sup_{x,y \in C} \|x - y\|$. 则

$$f(y_k) - f(x^*) \leq \frac{2L}{k(k+1)} \sum_{i=1}^k \|x_i - y_{i-1}\|^2 \leq \frac{2L}{k+1} D_C^2.$$

证明: 令 $\gamma_k = \frac{2}{k+1}$, 记 $\bar{y}_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_k$, 则不管

$$\alpha_k = \frac{2}{k+1} \quad \text{或} \quad \alpha_k = \operatorname{argmin}_{\alpha \in [0,1]} f((1 - \alpha)y_{k-1} + \alpha x_k).$$

对 $y_k = (1 - \alpha_k)y_{k-1} + \alpha_k x_k$, 我们都有 $f(y_k) \leq f(\bar{y}_k)$ 。注意到 $\bar{y}_k - y_{k-1} = \gamma_k(x_k - y_{k-1})$, 由 $f(x) \in C_L^{1,1}(C)$ 有

$$f(y_k) \leq f(\bar{y}_k) \leq f(y_{k-1}) + \langle \nabla f(y_{k-1}), \bar{y}_k - y_{k-1} \rangle + \frac{L}{2} \|\bar{y}_k - y_{k-1}\|^2 \quad (14)$$

$$\leq (1 - \gamma_k)[f(y_{k-1}) + \gamma_k[f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle]] + \frac{L\gamma_k^2}{2} \|x_k - y_{k-1}\|^2 \quad (15)$$

$$\leq (1 - \gamma_k)f(y_{k-1}) + \gamma_k f(x) + \frac{L\gamma_k^2}{2} \|x_k - y_{k-1}\|^2, \quad \text{对任意 } x \in C. \quad (16)$$

收敛性分析

其中不等式(15)是因为 $x_k \in \min_{x \in C} \langle \nabla f(y_{k-1}), x \rangle$, 由最优性条件我们可以得到对任意 $x \in C$ 有 $\langle x - x_k, \nabla f(y_{k-1}) \rangle \geq 0$ 。将不等式(16)稍做变换, 对任意 $x \in C$,

$$f(y_k) - f(x) \leq (1 - \gamma_k)[f(y_{k-1}) - f(x)] + \frac{L}{2} \gamma_k^2 \|x_k - y_{k-1}\|^2. \quad (17)$$

由引理可知,

$$f(y_k) - f(x) \leq \Gamma_k(1 - \gamma_1)[f(y_0) - f(x)] + \frac{\Gamma_k L}{2} \sum_{i=1}^k \frac{\gamma_i^2}{\Gamma_i} \|x_i - y_{i-1}\|^2.$$

由 $\gamma_k = \frac{2}{k+1}$, $\gamma_1 = 1$ 得到 $\Gamma_k = \frac{2}{k(k+1)}$, 我们可以得到收敛性不等式,

$$f(y_k) - f^* \leq \frac{2L}{k(k+1)} \sum_{i=1}^k \|x_i - y_{i-1}\|^2 \leq \frac{2L}{k+1} D_C^2.$$

令 $\frac{2L}{k+1} D_C^2 \leq \epsilon$, 可以得到分析复杂度结论。