Lecture 19: 近似点梯度算法

Lecturer: 陈士祥 *Scribes:* 陈士祥

1 问题形式

我们将考虑如下复合优化问题：

$$\min_{x\in\mathbb{R}^n} \quad \psi(x) = f(x) + h(x) \tag{19.1}$$

- 函数 f 为可微函数，其定义域 $\mathbf{dom}\ f = \mathbb{R}^n$
- 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：$\mathcal{O}(1/\sqrt{k})$

问题(19.1)可以用次梯度算法求解，但是次梯度方向并非下降方向，收敛速度是 $1/\sqrt{k}$。

是否可以设计复杂度为 $\mathcal{O}(1/k)$ 的算法？

2 近似点算法

若没有非光滑函数项 $h(x)$，回顾梯度算法，我们每次用一个二次上界函数近似 $f(x)$, 即

$$x_{k+1} = \arg\min_y f(x_k) + \nabla f(x_k)^T(y - x_k) + \frac{1}{2\alpha}\|y - x_k\|^2 = x_k - \alpha\nabla f(x_k).$$

若存在 $h(x)$, 近似点梯度算法的迭代如下：

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

对于某些 $h(x)$，上述迭代子问题是容易求解的。该子问题和邻近算子相关。

2.1 邻近算子

定义**邻近算子**：

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

直观理解：求解一个距 x 不算太远的点 u，并使函数值 $h(u)$ 也相对较小。

Definition 19.1 一个函数被称为闭函数，如果它的上方图是闭集.

Lemma 19.1 f 是闭函数当且仅当 f 的所有 α-下水平集都是闭集。其中，α-下水平集是如下集合

$$\{x : f(x) \le \alpha\}.$$

Proposition 19.1 (邻近算子是良定义的) 如果 h 为闭凸函数，则对任意 x，$\mathrm{prox}_h(x)$ **存在**且**唯一**

Proof: 首先注意到 $h(u) + \frac{1}{2}\|u-x\|_2^2$ 是关于 u 的强凸函数，则

- 存在性：强凸函数的所有 α-下水平集有界，由 h 是闭函数可知 α-下水平集是闭集。故由 Weierstrass 定理知最小值存在
- 唯一性：强凸函数最小值唯一

■

Example 19.1 (反例) 若 C 是开凸集，那么指示函数 $\mathcal{I}_C(x)$ 的邻近算子为投影点，故邻近点不存在。

2.1.1 一些典型的例子

Example 19.2 给定 $t>0$, $h(x)=\|x\|_1$, $\quad \mathrm{prox}_{th}(x) = \mathrm{sign}(x)\max\{|x|-t, 0\}$.

Proof: 邻近算子 $u=\mathrm{prox}_{th}(x)$ 的最优性条件为

$$x-u \in t\partial\|u\|_1 = \begin{cases} \{t\}, & u>0 \\ [-t,t], & u=0 \\ \{-t\}, & u<0 \end{cases}$$

当 $x>t$ 时, $u=x-t$; 当 $x<-t$ 时, $u=x+t$; 当 $x\in[-t,t]$ 时, $u=0$,
即有 $u=\mathrm{sign}(x)\max\{|x|-t,0\}$, 该映射也叫做软阈值函数 *(Soft-thresholding mapping)*。 ■

Example 19.3 给定 $t > 0$, $h(x) = \|x\|_2$, $\quad \operatorname{prox}_{th}(x) = \begin{cases} \left(1 - \frac{t}{\|x\|_2}\right)x, & \|x\|_2 \geqslant t, \\ 0, & \text{其他}. \end{cases}$

Proof: 邻近算子 $u = \operatorname{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_2 = \begin{cases} \left\{\frac{tu}{\|u\|_2}\right\}, & u \neq 0, \\ \{w : \|w\|_2 \leqslant t\}, & u = 0, \end{cases}$$

因此, 当 $\|x\|_2 > t$ 时, $u = x - \frac{tx}{\|x\|_2}$; 当 $\|x\|_2 \leqslant t$ 时, $u = 0$. ■

Example 19.4 二次函数 *(*其中 A 对称正定*)*

$$h(x) = \frac{1}{2}x^{\mathrm{T}}Ax + b^{\mathrm{T}}x + c, \quad \operatorname{prox}_{th}(x) = (I + tA)^{-1}(x - tb)$$

2.2 闭凸集上的投影

Example 19.5 设 C 为闭凸集, 则示性函数 I_C 的邻近算子为点 x 到 C 的投影 $\mathcal{P}_C(x)$:

$$\begin{aligned} \operatorname{prox}_{I_C}(x) &= \arg\min_u \left\{ I_C(u) + \frac{1}{2}\|u - x\|^2 \right\} \\ &= \underset{u \in C}{\operatorname{argmin}} \|u - x\|^2 = \mathcal{P}_C(x) \end{aligned}$$

这个等式具有几何意义:

$$u = \mathcal{P}_C(x) \Leftrightarrow (x - u)^{\mathrm{T}}(z - u) \leqslant 0, \quad \forall z \in C$$

- **超平面** $C = \{x | a^T x = b\}$ ($a \neq 0$)

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2} a$$

- **仿射集** $C = \{x | Ax = b\}$ ($A \in \mathbb{R}^{p \times n}$, 且 $\operatorname{rank}(A) = p$)

$$P_C(x) = x + A^T(AA^T)^{-1}(b - Ax)$$

  当 $p \ll n$, 或$AA^T = I, \ldots$ 时, 计算成本较低.

- **半平面** $C = \{x | a^T x \leq b\}$ ($a \neq 0$)

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2} a \quad if \quad a^T x > b,$$

$$P_C(x) = x \quad if \quad a^T x \leq b$$

- **矩形**：$C=[l,u]=\{x \mid l\le x\le u\}$

$$P_C(x)_i=\begin{cases} l_i & x_i\le l_i \\ x_i & l_i\le x_i\le u_i \\ u_i & x_i\ge u_i \end{cases}$$

- **非负象限**：$C=\mathbf{R}^n_+$

$$P_C(x)=x_+ \qquad (x_+ \text{ 表示各分量取 } \max\{0,x\})$$

- **概率单纯形**：$C=\{x|1^Tx=1,x\ge 0\}$

$$P_C(x)=(x-\lambda 1)_+$$

其中，λ 是下面方程的解：

$$1^T(x-\lambda 1)_+=\sum_{i=1}^n \max\{0,x_k-\lambda\}=1.$$

- **（一般的）概率单纯形**：$C=\{x|a^Tx=b,l\le x\le u\}$

$$P_c(x)=P_{[l,u]}(x-\lambda a)$$

其中, λ 是下面方程的解：

$$a^TP_{[l,u]}(x-\lambda a)=b$$

- **Euclidean 球**：$C=\{x|\|x\|_2\le 1\}$

$$P_C(x)=\frac{1}{\|x\|_2}x \quad if \quad \|x\|_2>1,$$

$$P_C(x)=x \quad if \quad \|x\|_2\le 1.$$

- ℓ_1 **范数球**：$C=\{x|\|x\|_1\le 1\}$

$$P_c(x)_k=\begin{cases} x_k-\lambda & x_k>\lambda \\ 0 & -\lambda\le\ x_k\le\lambda \\ x_k+\lambda & x_k<-\lambda \end{cases}$$

若 $\|x\|_1\le 1$，则 $\lambda=0$; 其他情形，λ 是下面方程的解

$$\sum_{k=1}^n \max\{|x_k|-\lambda,0\}=1.$$

作业 19.1 证明 ℓ_1 范数球的投影算子如上，并给出求解 λ 的一个算法。

3 近似点梯度算法

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

$$x^{k+1} = \operatorname{prox}_{t_k h}\left(x^k - t_k \nabla f\left(x^k\right)\right) \tag{19.2}$$

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: $\quad x^{k+1} = \operatorname{prox}_{t_k h}\left(x^k - t_k \nabla f\left(x^k\right)\right)$.
4: **end while**

根据定义，(19.2)式等价于

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

根据邻近算子与次梯度的关系, 又可以形式地写成

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

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.

3.1 投影梯度法

考虑问题

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

集合 C 是给定的闭凸集，定义 $\mathcal{I}_C(x)$ 表示指示函数，若 $h(x) = \mathcal{I}_C(x)$。那么(19.2)可以写为

$$x^{k+1} = \mathcal{P}_C\left(x^k - t_k \nabla f\left(x^k\right)\right). \tag{19.3}$$

这便是投影梯度法，即每次先沿着负梯度方向更新以减少函数值，再投影回到可行域 C 上保证迭代点可行性。所以投影梯度法可以看成近似点梯度法的一个特例。

3.2 收敛性分析

收敛性分析依赖于下面基本的假设。

假设 19.1 • f 在 $\mathbb{R}^n$ 上是凸的; ∇f 为 L-利普希茨连续，即

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

- 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).

在基本假设的基础上，我们定义**梯度映射**:

$$G_t(x)=\frac{1}{t}\left(x-\operatorname{prox}_{th}(x-t\nabla f(x))\right) \quad (t>0) \tag{19.4}$$

不难推出梯度映射具有以下的性质:

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

$$G_t(x)-\nabla f(x) \in \partial h\left(x-tG_t(x)\right) \tag{19.5}$$

- 与算法的收敛性的关系:

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

Theorem 19.1 (固定步长近似点梯度法的收敛性) 取定步长为 $t_k=t\in\left(0,\frac{1}{L}\right]$, 设 $\left\{x^k\right\}$ 由迭代格式(19.2)产生，则

$$\psi\left(x^k\right)-\psi^* \leqslant \frac{1}{2kt}\left\|x^0-x^*\right\|^2.$$

Proof: 根据利普希茨连续“二次上界”的性质，得到

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

令 $x^+=x-tG_t(x)$, 有

$$\begin{aligned} f\left(x^+\right) &\leqslant f(x)-t\nabla f(x)^{\mathrm{T}}G_t(x)+\frac{t^2L}{2}\left\|G_t(x)\right\|^2 \\ &\leqslant f(x)-t\nabla f(x)^{\mathrm{T}}G_t(x)+\frac{t}{2}\left\|G_t(x)\right\|^2. \end{aligned} \tag{19.6}$$

此外，由 $f(x), h(x)$ 为凸函数，结合(19.4)式我们有

$$h\left(x^+\right) \leqslant h(z)-\left(G_t(x)-\nabla f(x)\right)^{\mathrm{T}}\left(z-x^+\right) \tag{19.7}$$

$$f(x) \leqslant f(z)-\nabla f(x)^{\mathrm{T}}(z-x) \tag{19.8}$$

将(19.6)(19.7)(19.8)式相加可得对任意 $z \in \mathbf{dom}\,\psi$ 有

$$\psi\left(x^{+}\right) \leqslant \psi(z)+G_t(x)^{\mathrm{T}}(x-z)-\frac{t}{2}\left\|G_t(x)\right\|^2 \tag{19.9}$$

由 $x^i=x^{i-1}-tG_t(x^{i-1})$，在不等式 (19.9) 中，取 $z=x^*, x=x^{i-1}$ 得到

$$\begin{aligned}
\psi(x^i)-\psi^* &\leqslant G_t(x^{i-1})^{\mathrm{T}}\left(x^{i-1}-x^*\right)-\frac{t}{2}\left\|G_t(x^{i-1})\right\|^2 \\
&=\frac{1}{2t}\left(\left\|x^{i-1}-x^*\right\|^2-\left\|x^{i-1}-x^*-tG_t(x^{i-1})\right\|^2\right) \\
&=\frac{1}{2t}\left(\left\|x^{i-1}-x^*\right\|^2-\left\|x^{i}-x^*\right\|^2\right)
\end{aligned} \tag{19.10}$$

取 $i=1,2,\cdots,k$ 并累加，得

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

注意到在不等式 (19.9)中，取 $z=x^{i-1}$ 即得:

$$\psi(x^i) \leqslant \psi(x^{i-1})-\frac{t}{2}\left\|G_t(x^{i-1})\right\|^2.$$

即 $\psi\left(x^i\right)$ 不增，因此

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

■

4 应用

考虑低秩矩阵恢复模型:

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

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

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

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
$$f(X)=\frac{1}{2}\|P\odot(X-M)\|_F^2$$
进一步可以得到
$$\begin{aligned}\nabla f(X)&=P\odot(X-M),\\ \operatorname{prox}_{t_kh}(X)&=U\operatorname{Diag}\left(\max\{|d|-t_k\mu,0\}\right)V^{\mathrm{T}},\end{aligned}$$
其中 $X=U\,\operatorname{Diag}(d)V^{\mathrm{T}}$ 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:

$$\begin{aligned}Y^k&=X^k-t_kP\odot\left(X^k-M\right)\\ X^{k+1}&=\operatorname{prox}_{t_kh}\left(Y^k\right)\end{aligned}$$