

投影次梯度下降的收敛性分析

张腾

设 $\mathcal{W}$ 是闭凸集 (closed convex set), $f(w)$ 是凸函数 (convex function), 对于如下的约束凸优化

$$\min_w f(w), \quad \text{s.t. } w \in \mathcal{W} \subseteq \mathbb{R}^n$$

投影次梯度下降 (projected subgradient descent) 的第 t 轮更新公式为

$$w_{t+1} = \text{proj}_{\mathcal{W}}(w_t - \eta_t g_t) \triangleq \underset{w \in \mathcal{W}}{\text{argmin}} \|w - (w_t - \eta_t g_t)\|_2$$

其中 $\eta_t > 0$ 是步长, $g_t \in \partial f(w_t)$ 是次梯度, 即将单步下降后的点投影到可行域 $\mathcal{W}$ 。进一步化简有

$$\begin{aligned} w_{t+1} &= \underset{w \in \mathcal{W}}{\text{argmin}} \frac{1}{2} \|w - w_t + \eta_t g_t\|_2^2 = \underset{w \in \mathcal{W}}{\text{argmin}} \{ \eta_t \langle g_t, w - w_t \rangle + \frac{1}{2} \|w - w_t\|_2^2 \} \\ &= \underset{w \in \mathcal{W}}{\text{argmin}} \{ \eta_t \underbrace{(f(w_t) + \langle g_t, w - w_t \rangle)}_{f(w) \text{ 在 } w_t \text{ 处的线性近似}} + \frac{1}{2} \underbrace{\|w - w_t\|_2^2}_{w_t \text{ 的邻域}} \} \\ &= \underset{w \in \mathcal{W}}{\text{argmin}} \{ \eta_t \langle g_t, w \rangle + \frac{1}{2} \|w - w_t\|_2^2 \} \end{aligned}$$

即 w_{t+1} 是在 w_t 的某个邻域内最小化 $f(w)$ 在 w_t 处的线性近似。

注. 若可行域 $\mathcal{W} = \mathbb{R}^n$, 则无需做投影, 若 f 可微, 则次梯度即为梯度。

引理 1. 记 $\phi(w) = \eta_t \langle g_t, w \rangle + \frac{1}{2} \|w - w_t\|_2^2$, $w_{t+1} = \underset{w \in \mathcal{W}}{\text{argmin}} \phi(w)$, 则

$$\eta_t \langle g_t, w_{t+1} - w \rangle \leq \frac{1}{2} \|w - w_t\|_2^2 - \frac{1}{2} \|w - w_{t+1}\|_2^2 - \frac{1}{2} \|w_{t+1} - w_t\|_2^2 \quad (1)$$

证明. $\phi(w)$ 是关于 w 的二次函数, 其等高线 $\phi(w_{t+1})$ 与 $\mathcal{W}$ 相切于 w_{t+1} , 梯度 $\nabla \phi(w_{t+1})$ 与切线垂直, 因此对于 $\forall w \in \mathcal{W}$ 有

$$\langle \nabla \phi(w_{t+1}), w - w_{t+1} \rangle \geq 0$$

于是

$$\begin{aligned} &\eta_t \langle g_t, w \rangle + \frac{1}{2} \|w - w_t\|_2^2 \\ &= \eta_t \langle g_t, w - w_{t+1} + w_{t+1} \rangle + \frac{1}{2} \|w - w_{t+1} + w_{t+1} - w_t\|_2^2 \\ &= \eta_t \langle g_t, w_{t+1} \rangle + \frac{1}{2} \|w_{t+1} - w_t\|_2^2 + \eta_t \langle g_t, w - w_{t+1} \rangle + \frac{1}{2} \|w - w_{t+1}\|_2^2 + \langle w - w_{t+1}, w_{t+1} - w_t \rangle \\ &= \eta_t \langle g_t, w_{t+1} \rangle + \frac{1}{2} \|w_{t+1} - w_t\|_2^2 + \langle \eta_t g_t + w_{t+1} - w_t, w - w_{t+1} \rangle + \frac{1}{2} \|w - w_{t+1}\|_2^2 \\ &= \eta_t \langle g_t, w_{t+1} \rangle + \frac{1}{2} \|w_{t+1} - w_t\|_2^2 + \langle \nabla \phi(w_{t+1}), w - w_{t+1} \rangle + \frac{1}{2} \|w - w_{t+1}\|_2^2 \\ &\geq \eta_t \langle g_t, w_{t+1} \rangle + \frac{1}{2} \|w_{t+1} - w_t\|_2^2 + \frac{1}{2} \|w - w_{t+1}\|_2^2 \end{aligned}$$

移项整理即可。



1 非平滑函数

假设:

- f 是 L -李普希茨连续函数, 即对 $\forall w \in \mathcal{W}$ 及 $\forall g \in \partial f(w)$ 有 $\|g\| \leq L$ 。
- f 是 α -强凸函数, 对一般凸函数可令 $\alpha = 0$ 。

引理 2. 对 $\forall w \in \mathcal{W}$ 有

$$f(w_t) - f(w) \leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 + \frac{L^2\eta_t}{2} \quad (2)$$

证明. 分别根据 α -强凸、式(1)、柯西不等式、 L -李普希茨连续、均值不等式有

$$\begin{aligned} f(w_t) - f(w) &\leq \langle g_t, w_t - w \rangle - \frac{\alpha}{2} \|w - w_t\|_2^2 \\ &= \langle g_t, w_t - w_{t+1} \rangle + \langle g_t, w_{t+1} - w \rangle - \frac{\alpha}{2} \|w - w_t\|_2^2 \\ &\stackrel{(1)}{\leq} \langle g_t, w_t - w_{t+1} \rangle + \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 - \frac{1}{2\eta_t} \|w_{t+1} - w_t\|_2^2 \\ &\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 + \|g_t\|_2 \|w_t - w_{t+1}\|_2 - \frac{1}{2\eta_t} \|w_{t+1} - w_t\|_2^2 \\ &\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 + L \|w_t - w_{t+1}\|_2 - \frac{1}{2\eta_t} \|w_{t+1} - w_t\|_2^2 \\ &\leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 + \frac{L^2\eta_t}{2} \end{aligned}$$

♣

1.1 f 非强凸

此时 $\alpha = 0$, 根据式(2), 对 $\forall w \in \mathcal{W}$ 有

$$f(w_t) - f(w) \leq \frac{1}{2\eta_t} \|w - w_t\|_2^2 - \frac{1}{2\eta_t} \|w - w_{t+1}\|_2^2 + \frac{L^2\eta_t}{2}$$

累加可得

$$\begin{aligned} \sum_{t=s}^r \eta_t f(w_t) - \sum_{t=s}^r \eta_t f(w) &\leq \frac{1}{2} \|w - w_s\|_2^2 - \frac{1}{2} \|w - w_{r+1}\|_2^2 + \frac{L^2}{2} \sum_{t=s}^r \eta_t^2 \\ &\leq \frac{1}{2} \|w - w_s\|_2^2 + \frac{L^2}{2} \sum_{t=s}^r \eta_t^2 \\ &\leq R^2 + \frac{L^2}{2} \sum_{t=s}^r \eta_t^2 \end{aligned} \quad (3)$$

其中 $R^2 \triangleq \max_{u,v \in \mathcal{W}} \|u - v\|_2^2/2$ 。

1.1.1 固定步长

取 $\eta_t = \eta$, $s = 1$, $r = T$, 由式(3)有

$$\frac{1}{T} \sum_{t \in [T]} f(w_t) - f(w) \leq \frac{R^2}{\eta T} + \frac{L^2 \eta}{2}$$

欲使上界最紧, 可令

$$\eta = \sqrt{\frac{2R^2}{TL^2}} \implies \frac{1}{T} \sum_{t \in [T]} f(w_t) - f(w) \leq \sqrt{\frac{2R^2 L^2}{T}}$$

令 $\bar{w} = \sum_{t \in [T]} w_t / T$, $\hat{w} = \operatorname{argmin}_{w_t: t \in [T]} f(w_t)$, 由 f 的凸性有

$$f(\bar{w}) - f(w) \leq \frac{\sqrt{2}RL}{\sqrt{T}}, \quad f(\hat{w}) - f(w) \leq \frac{\sqrt{2}RL}{\sqrt{T}}$$

1.1.2 递减步长

固定步长 η 依赖总迭代轮数 T , 可取递减步长

$$\eta_t = \sqrt{\frac{2R^2}{tL^2}}$$

易知

$$\begin{aligned} \sum_{t=\lceil T/2 \rceil}^T \eta_t &= \sqrt{\frac{2R^2}{L^2}} \sum_{t=\lceil T/2 \rceil}^T \frac{1}{\sqrt{t}} \geq \sqrt{\frac{2R^2}{L^2}} \int_{(T+1)/2}^{T+1} \frac{1}{\sqrt{t}} dt = 2\sqrt{\frac{R^2(T+1)}{L^2}}(\sqrt{2}-1) \\ \sum_{t=\lceil T/2 \rceil}^T \eta_t^2 &= \frac{2R^2}{L^2} \sum_{t=\lceil T/2 \rceil}^T \frac{1}{t} \leq \frac{2R^2}{L^2} \int_{\lceil T/2 \rceil - 1}^T \frac{1}{t} dt = \frac{2R^2}{L^2} \ln \frac{T}{\lceil T/2 \rceil - 1} \leq \frac{2R^2}{L^2} \ln 3 \end{aligned} \quad (4)$$

由 f 的凸性、式(3)、式(4)有

$$\begin{aligned} f\left(\sum_{t=\lceil T/2 \rceil}^T \frac{\eta_t w_t}{\sum_{t'=\lceil T/2 \rceil}^T \eta_{t'}}\right) - f(w) &\leq \sum_{t=\lceil T/2 \rceil}^T \frac{\eta_t}{\sum_{t'=\lceil T/2 \rceil}^T \eta_{t'}} f(w_t) - f(w) \\ &\leq \frac{1}{\sum_{t'=\lceil T/2 \rceil}^T \eta_{t'}} \left(R^2 + \frac{L^2}{2} \sum_{t=\lceil T/2 \rceil}^T \eta_t^2 \right) \\ &\leq \frac{L}{2(\sqrt{2}-1)R\sqrt{T+1}} R^2(1+\ln 3) = \frac{(1+\ln 3)RL}{2(\sqrt{2}-1)\sqrt{T+1}} \end{aligned}$$

注. • 所有界都是 $\mathcal{O}(RL/\sqrt{T})$, 要想误差不超过 ϵ , 需做 $\mathcal{O}(R^2 L^2 / \epsilon^2)$ 轮迭代。

- 所有界都是对 $\forall w \in \mathcal{W}$ 成立, 不仅仅是 w^* 。
- 界不是关于最后的 w_T 的, 固定步长是平均和最好的 w , 递减步长是后一半 w 的加权平均。

1.2 f 强凸

此时 $\alpha > 0$, 累加式(2)可得

$$\begin{aligned} \sum_{t \in [T]} a_t f(w_t) - \sum_{t \in [T]} a_t f(w) &\leq \sum_{t \in [T]} a_t \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \sum_{t \in [T]} \frac{a_t}{2\eta_t} \|w - w_{t+1}\|_2^2 + \frac{L^2}{2} \sum_{t \in [T]} a_t \eta_t \\ &= \sum_{t \in [T]} a_t \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w - w_t\|_2^2 - \sum_{t=2}^{T+1} \frac{a_{t-1}}{2\eta_{t-1}} \|w - w_t\|_2^2 + \frac{L^2}{2} \sum_{t \in [T]} a_t \eta_t \\ &= \frac{a_1(1 - \alpha\eta_1)}{2\eta_1} \|w - w_1\|_2^2 + \sum_{t=2}^{T+1} \left(\frac{a_t(1 - \alpha\eta_t)}{2\eta_t} - \frac{a_{t-1}}{2\eta_{t-1}} \right) \|w - w_t\|_2^2 + \frac{L^2}{2} \sum_{t \in [T]} a_t \eta_t \end{aligned}$$

取 $\eta_t = 2/\alpha t$ 、 $a_t = t$, 则

$$\frac{a_t(1 - \alpha\eta_t)}{\eta_t} = \frac{\alpha t^2(1 - 2/t)}{2} = \frac{\alpha t(t-2)}{2} \leq \frac{\alpha(t-1)^2}{2} = \frac{a_{t-1}}{\eta_{t-1}}$$

故不等号右边第二项 ≤ 0 , 将其放掉有

$$\sum_{t \in [T]} t f(w_t) - \frac{T(T+1)}{2} f(w) \leq -\frac{\alpha}{4} \|w - w_1\|_2^2 + \frac{L^2 T}{\alpha}$$

不等号右边第一项也 ≤ 0 , 将其放掉, 两边同除以 $T(T+1)/2$ 可得

$$f\left(\sum_{t=1}^T \frac{2t}{T(T+1)} w_t\right) - f(w) \leq \frac{2L^2}{\alpha(T+1)}$$

注. 误差界是 $\mathcal{O}(L^2/\alpha T)$, 要想误差不超过 ϵ , 需做 $\mathcal{O}(L^2/\alpha\epsilon)$ 轮迭代。

2 β -平滑函数

假设

- f 是 β -平滑函数, 即 f 可微且梯度 β -李普希茨连续, 即对于 $\forall w, u \in S$ 有

$$\|\nabla f(w) - \nabla f(u)\|_2 \leq \beta \|w - u\|_2 \implies f(w) \leq f(u) + \langle \nabla f(u), w - u \rangle + \frac{\beta}{2} \|w - u\|_2^2$$

- f 是 α -强凸函数, 对一般凸函数可令 $\alpha = 0$ 。

引理 3 (单调性). 若步长 $\eta_t \leq 2/\beta$, 则 $f(w_{t+1}) \leq f(w_t)$ 。

证明. 此时 $g_t = \nabla f(w_t)$ 就是梯度, 式(1)变成

$$\eta_t \langle \nabla f(w_t), w_{t+1} - w \rangle \leq \frac{1}{2} \|w - w_t\|_2^2 - \frac{1}{2} \|w - w_{t+1}\|_2^2 - \frac{1}{2} \|w_{t+1} - w_t\|_2^2, \quad w \in \mathcal{W} \quad (5)$$

令 $w = w_t$ 可得 $\eta_t \langle \nabla f(w_t), w_{t+1} - w_t \rangle \leq -\|w_t - w_{t+1}\|_2^2$, 于是

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{\beta}{2} \|w_{t+1} - w_t\|_2^2$$

$$\begin{aligned}
&\leq f(\mathbf{w}_t) - \left(\frac{1}{\eta_t} - \frac{\beta}{2} \right) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq f(\mathbf{w}_t)
\end{aligned}$$

♣

引理 4. 若步长 $\eta_t \leq 1/\beta$, 则

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}) + \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2, \quad \forall \mathbf{w} \in \mathcal{W} \quad (6)$$

证明. 对 $\forall \mathbf{w} \in \mathcal{W}$, 分别根据 β -平滑、 α -强凸、式(5)、 $\beta \leq 1/\eta_t$ 有

$$\begin{aligned}
f(\mathbf{w}_{t+1}) &\leq f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{\beta}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= \underbrace{f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + \frac{\alpha}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2}_{\leq f(\mathbf{w})} \\
&\quad + \underbrace{\langle \nabla f(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w} \rangle}_{\text{用式(5)来放}} + \frac{\beta}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 - \frac{\alpha}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \\
&\leq f(\mathbf{w}) + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\quad + \frac{\beta}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 - \frac{\alpha}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \\
&\leq f(\mathbf{w}) + \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2
\end{aligned}$$

♣

2.1 f 非强凸

此时 $\alpha = 0$, 式(6)变成

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}) \leq \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2$$

累加可得

$$\sum_{t \in [T]} f(\mathbf{w}_{t+1}) - Tf(\mathbf{w}) \leq \frac{1}{2\eta_1} \|\mathbf{w} - \mathbf{w}_1\|_2^2 - \frac{1}{2\eta_T} \|\mathbf{w} - \mathbf{w}_{T+1}\|_2^2$$

取 $\eta_t = 1/\beta$, 根据单调性有

$$f(\mathbf{w}_{T+1}) - f(\mathbf{w}) \leq \frac{\beta}{2T} \|\mathbf{w} - \mathbf{w}_1\|_2^2 - \frac{\beta}{2T} \|\mathbf{w} - \mathbf{w}_{T+1}\|_2^2 \leq \frac{\beta R^2}{T}$$

注. 误差界是 $\mathcal{O}(\beta R^2/T)$, 要想误差不超过 ϵ , 需做 $\mathcal{O}(\beta R^2/\epsilon)$ 轮迭代。

注. 有了平滑性, 就有了 $\{f(\mathbf{w}_t)\}$ 序列的单调性, 从而误差界也只跟最后的 $\mathbf{w}$ 相关了。

2.2 f 强凸

此时 $\alpha > 0$, 对式(6)令 $w = w^*$, 结合 $f(w_{t+1}) \geq f(w^*)$ 可得

$$0 \leq \left(\frac{1}{2\eta_t} - \frac{\alpha}{2} \right) \|w^* - w_t\|_2^2 - \frac{1}{2\eta_t} \|w^* - w_{t+1}\|_2^2$$

即 $\|w^* - w_{t+1}\|_2^2 \leq (1 - \alpha\eta_t) \|w^* - w_t\|_2^2$, 取 $\eta_t = 1/\beta$ 累乘可得

$$\|w^* - w_{T+1}\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^T \|w^* - w_1\|_2^2 = \left(1 - \frac{1}{\kappa}\right)^T \|w^* - w_1\|_2^2 \leq \exp\left(-\frac{T}{\kappa}\right) 2R^2$$

其中 $\kappa = \beta/\alpha$ 为条件数 (condition number)。

注. 误差界是 $\mathcal{O}(R^2 \exp(-T/\kappa))$, 要想误差不超过 ϵ , 需做 $\mathcal{O}(\kappa \ln(R^2/\epsilon))$ 轮迭代。

若想得到目标函数值的误差界, 对式(6)令 $w = w^*$, $\eta_t = 1/\beta$ 可得

$$\begin{aligned} f(w_{T+1}) - f(w^*) &\leq \left(\frac{\beta}{2} - \frac{\alpha}{2} \right) \|w^* - w_T\|_2^2 - \frac{1}{2\eta_t} \|w^* - w_{T+1}\|_2^2 \\ &\leq \frac{\beta - \alpha}{2} \|w^* - w_T\|_2^2 \\ &\leq \frac{\beta - \alpha}{2} \left(1 - \frac{\alpha}{\beta}\right)^{T-1} \|w^* - w_1\|_2^2 \\ &= \frac{\beta}{2} \left(1 - \frac{\alpha}{\beta}\right)^T \|w^* - w_1\|_2^2 \\ &\leq \exp\left(-\frac{T}{\kappa}\right) \beta R^2 \end{aligned}$$

注. 误差界是 $\mathcal{O}(\beta R^2 \exp(-T/\kappa))$, 要想误差不超过 ϵ , 需做 $\mathcal{O}(\kappa \ln(\beta R^2/\epsilon))$ 轮迭代。

注. Nesterov 在 Introductory Lectures on Convex Optimization 一书中采用步长 $\eta = \frac{2}{\beta + \alpha}$, 得到

$$f(w_{T+1}) - f(w^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\kappa + 1}\right) \|w_1 - w^*\|_2^2 \leq \exp\left(-\frac{4T}{\kappa + 1}\right) \beta R^2$$

要想误差不超过 ϵ , 需做 $\mathcal{O}((\kappa + 1) \ln(\beta R^2/\epsilon)/4)$ 轮迭代, 在常数上有所改进。

本文的所有结论总结如下:

表 1: 4 种不同的情况下投影次梯度下降的收敛率

	L-李普希茨连续	α -强凸、L-李普希茨连续	β -平滑	β -平滑、 α -强凸
T 轮误差	$\mathcal{O}(RL/\sqrt{T})$	$\mathcal{O}(L^2/\alpha T)$	$\mathcal{O}(\beta R^2/T)$	$\mathcal{O}(\beta R^2 \exp(-T/\kappa))$
迭代轮数	$\mathcal{O}(R^2 L^2/\epsilon^2)$	$\mathcal{O}(L^2/\alpha \epsilon)$	$\mathcal{O}(\beta R^2/\epsilon)$	$\mathcal{O}(\kappa \ln(\beta R^2/\epsilon))$