SGD收敛速率

计算机学院+20020129+王悟信+SGD收敛速率

证明目标：

$$\mathbb{E}_{V_{1:T}}[\frac{1}{T}\sum_{i=1}^{T}(f(w^{(t)})-f(w^\star)] \le \mathbb{E}_{V_{1:T}}[\frac{1}{T}\sum_{i=1}^{T}\langle w^{(t)}-w^\star, v_t\rangle]$$

证明：

$$\underset{v_{1:T}}{\mathbb{E}}[\frac{1}{T}\sum_{i=1}^{T}\langle w^(t)-w^\star, v_t\rangle] = \frac{1}{T}\sum_{i=1}^{T}\underset{v_{1:T}}{\mathbb{E}}\langle w^{(t)}-w^\star, v_t\rangle \tag{1}$$

则(1)右侧中取出一项，由于$w^{(t)}=w^{(t-1)}-\eta v_{t-1}$，则对于当前的$t$，$\mathbb{E}$中$i>t$的部分均不用考虑，根据全期望公式：

$$\begin{aligned}&\forall 变量\,\alpha,\beta 和某个函数\, g\\ &\mathbb{E}_\alpha[g(\alpha)] = \mathbb{E}_\beta[\mathbb{E}_\alpha[g(\alpha)|\beta]]\end{aligned} \tag{2}$$

有：

$$\begin{aligned}\underset{v_{1:T}}{\mathbb{E}}[\langle w^{(t)}-w^\star, v_t\rangle] &= \underset{v_{1:t}}{\mathbb{E}}[\langle w^{(t)}-w^\star, v_t\rangle]\\ &= \underset{1:t-1}{\mathbb{E}}[\underset{1:t}{\mathbb{E}}[\langle w^{(t)}-w^\star, v_t\rangle|v_{1:t-1}]]\end{aligned} \tag{3}$$

当$v_{1:t-1}$确定时，$w^{(t)}$也就确定了，所以：

$$\underset{1:t-1}{\mathbb{E}}\underset{1:t}{\mathbb{E}}[\langle w^{(t)}-w^\star, v_t\rangle|v_{1:t-1}] = \underset{v_{1:t-1}}{\mathbb{E}}\langle w^{(t)}-w^\star, \underset{v_t}{\mathbb{E}}[v_t|v_{t-1}]\rangle \tag{4}$$

由于SGD算法要求$\underset{v_t}{\mathbb{E}}[v_t|w^{(t)}]\in\partial f(w^{(t)})$，所以：

$$\underset{v_{1:t-1}}{\mathbb{E}}\langle w^{(t)}-w^\star, \underset{v_t}{\mathbb{E}}[v_t|v_{t-1}]\rangle \ge \underset{v_{1:t-1}}{\mathbb{E}}[f(w^{(t)})-f(w^\star)] \tag{5}$$

所以

$$\begin{aligned}\underset{v_{1:T}}{\mathbb{E}}[\langle w^{(t)-w^\star}, v_t\rangle] &\ge \underset{v_{1:t-1}}{\mathbb{E}}[fw^{(t)}-f(w^\star)]\\ &= \underset{v_{1:T}}{\mathbb{E}}[fw^{(t)}-f(w^\star)]\end{aligned} \tag{6}$$

故要证的式子成立