机器学习理论导引
作业四

DZ1833019, 欧先飞,ouxianfei@smail.nju.edu.cn

2019 年 5 月 21 日

1 [15pts] Conjugate Functions

推导下面函数的共轭函数

$$f(x) = \log(1 + \exp(-x)).$$

Proof.

首先由定义有：

$$f^*(y) = \sup_{y \in dom(f^*)} (yx - f(x)) = \sup_{y \in dom(f^*)} (yx - \ln(1 + e^{-x}))$$

令 $g(x) = yx - \ln(1 + e^{-x})$，对其求导则有 $g'(x) = y + \frac{e^{-x}}{1+e^{-x}} = y + \frac{1}{1+e^{x}}$，容易发现在整个定义域内，$g'(x)$ 呈上升趋势且 $g'(x) > y$ 恒成立，以下对 y 的取值范围进行讨论。

(1) 当 $y > 0$ 时，$g(x)$ 在 $(-\infty, +\infty)$ 内单调递增，则 $f^*(y) = \sup g(x) = \lim_{x \to +\infty}(yx - \ln(1 + e^{-x})) = +\infty$，此时 $f^*(y)$ 无定义。

(2) 当 $y = 0$ 时，$g(x)$ 在 $(-\infty, +\infty)$ 内单调递增，则 $f^*(y) = \sup g(x) = \lim_{x \to +\infty}(yx - \ln(1 + e^{-x})) = 0$。

(3) 当 $-1 < y < 0$ 时，$g'(x)$ 在定义域内存在零点 $\ln(-1 - \frac{1}{y})$，亦即 $g(x)$ 在 $x = \ln(-1 - \frac{1}{y})$ 时达到最值，故 $f^*(y) = \sup g(x) = g(\ln(-1 - \frac{1}{y})) = y\ln(-1 - \frac{1}{y}) - \ln\left(1 + e^{-\ln(-1-\frac{1}{y})}\right) = y\ln(-1 - \frac{1}{y}) - \ln\left(\frac{1}{y+1}\right)$。

(4) 当 $y = -1$ 时，$g'(x) < 0$ 在 $(-\infty, +\infty)$ 内恒成立，所以 $\sup g(x) = \lim_{x \to -\infty}(yx - \ln(1 + e^{-x})) = \lim_{x \to -\infty} \ln \frac{e^{(y+1)x}}{1+e^{x}} = 0$。

(5) 当 $y < -1$ 时，$\lim_{x \to -\infty}(yx - \ln(1 + e^{-x})) = +\infty$，所以 $\sup g(x)$ 不存在，此时 $f^*(y)$ 无定义。

综上所述

$$f^*(y) = \begin{cases} 0, & y = 0 \\ y\ln(-1 - \dfrac{1}{y}) - \ln\left(\dfrac{1}{y+1}\right), & -1 < y < 0 \\ 0, & y = -1 \end{cases}$$

$\square$

2 [15pts] Projection

对于凸集 $\mathcal{W}$，试证明投影操作 $\Pi_{\mathcal{W}}(\cdot)$ 是不扩展的，即

$$\|\Pi_{\mathcal{W}}(\mathbf{x}) - \Pi_{\mathcal{W}}(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|,\ \forall \mathbf{x}, \mathbf{y}.$$

Proof.
首先简记 $\Pi_{\mathcal{W}}(\mathbf{x})$ 为 $\mathbf{x}_{\mathcal{W}}$，由于 $\mathcal{W}$ 是凸包，所以有 $\forall \mathbf{x} \notin \mathcal{W} \forall \mathbf{y} \in \mathcal{W} (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) \leq 0$，以下对 $\mathbf{x}$ 和 $\mathbf{y}$ 的所处的位置进行讨论。
(1) $\mathbf{x}$ 和 $\mathbf{y}$ 均在凸包 $\mathcal{W}$ 范围内，则 $\mathbf{x}_{\mathcal{W}} = \mathbf{x}$、$\mathbf{y}_{\mathcal{W}} = \mathbf{y}$，所以有 $\|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| = \|\mathbf{x} - \mathbf{y}\|$
(2) 两点中有且只有一点在 $\mathcal{W}$ 中，不妨设 $\mathbf{y}$ 在 $\mathcal{W}$ 中。反设原命题不成立，即有 $\|\mathbf{x} - \mathbf{y}\| < \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| = \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}\|$，于是有

$$\|\mathbf{x} - \mathbf{y}\|_2^2 < \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}\|_2^2 \tag{2.1}$$

$$(\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) \leq 0 \tag{2.2}$$

计算 $2.1 + 2 \times 2.2$ 则有：

$$\begin{aligned}
& (\mathbf{x} - \mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + 2(\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) < (\mathbf{x}_{\mathcal{W}} - \mathbf{y})^{\mathrm{T}}(\mathbf{x}_{\mathcal{W}} - \mathbf{y}) \\
\Longrightarrow\ & (\mathbf{x} - \mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(2\mathbf{x} - 2\mathbf{x}_w - \mathbf{y} + \mathbf{x}_{\mathcal{W}}) < 0 \\
\Longrightarrow\ & (\mathbf{x} - \mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{y} + \mathbf{x} - \mathbf{x}_{\mathcal{W}}) < 0 \\
\Longrightarrow\ & (\mathbf{x} - \mathbf{y})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) < 0 \\
\Longrightarrow\ & (\mathbf{x} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) < 0 \\
\Longrightarrow\ & (\mathbf{x} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}} < 0 \\
\Longrightarrow\ & \|\mathbf{x} - \mathbf{x}_{\mathcal{W}}\|_2^2 < 0
\end{aligned}$$

这与事实 $\|\mathbf{x} - \mathbf{x}_{\mathcal{W}}\|_2^2 \geq 0$ 矛盾，所以反设不成立，继而有 $\|\mathbf{x} - \mathbf{y}\| \geq \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}\|$。
(3) 两点均在 $\mathcal{W}$ 外，由凸包的性质，可以得到如下几个条件：

$$(\mathbf{y}_{\mathcal{W}} - \mathbf{x}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_{\mathcal{W}}) \leq 0 \tag{2.3}$$

$$(\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{y} - \mathbf{y}_{\mathcal{W}}) \leq 0 \tag{2.4}$$

计算 2.3+2.4 可以得到：$(\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}) < (\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{y})$，而 $(\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}})^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) = \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| \|\mathbf{x} - \mathbf{y}\| \cos\langle \mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}, \mathbf{x} - \mathbf{y} \rangle \leq \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| \|\mathbf{x} - \mathbf{y}\|$，所以有 $\|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\|_2^2 < \|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| \|\mathbf{x} - \mathbf{y}\|$，由于二范数的非负性，所以可以化简得到 $\|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| \leq \|\mathbf{x} - \mathbf{y}\|$。

结合 (1)、(2)、(3) 可知，命题 $\|\mathbf{x}_{\mathcal{W}} - \mathbf{y}_{\mathcal{W}}\| \leq \|\mathbf{x} - \mathbf{y}\|$ 成立。

□

3 [20pts] Gradient Descent with Decaying Step Size

分析采用衰减步长时梯度下降（GD）的收敛速率。具体而言，考虑

$$\eta_t = O\left(\frac{1}{\sqrt{t}}\right)$$

Analysis.
设最优值为 $\mathbf{w}$，首先参考 *slides* 中的证明有：

$$\begin{aligned}
f(\mathbf{w}_t) - f(\mathbf{w}) &\leq \nabla f(\mathbf{w}_t)^T(\mathbf{w}_t - \mathbf{w}) \\
&= \frac{1}{\eta_t}\langle \mathbf{w}_t - \mathbf{w}'_{t+1}, \mathbf{w}_t - \mathbf{w}\rangle \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] + \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] + \frac{\eta_t}{2}\|\nabla f(\mathbf{w}_t)\|_2^2 \\
&\leq \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right] + \frac{\eta_t}{2}G^2
\end{aligned}$$

将上面得到的不等式从 $t=1$ 加到 T 则有：

$$\begin{aligned}
\sum_{t=1}^{T}[f(\mathbf{w}_t) - f(\mathbf{w})] &\leq \sum_{t=1}^{T}\frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \sum_{t=1}^{T}\frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t G^2 \\
&= \frac{1}{2\eta_1}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right)\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{1}{2\eta_T}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t G^2
\end{aligned}$$

由题意可设 $\eta_t = \frac{a}{\sqrt{t}}$，而 $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} < 1 + \int_1^T \frac{1}{\sqrt{x}}dx = 1 + (2\sqrt{x})|_1^T = 2\sqrt{T} - 1$，代入上面的不等式则有：

$$\begin{aligned}
\sum_{t=1}^{T}[f(\mathbf{w}_t) - f(\mathbf{w})] &\leq \frac{1}{2\eta_1}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}\right)\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{1}{2\eta_T}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_t G^2 \\
&= \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2a} + \sum_{t=1}^{T-1}\left(\frac{\sqrt{t+1}}{2a} - \frac{\sqrt{t}}{2a}\right)\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{\sqrt{T}}{2a}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{2\sqrt{T}-1}{2}aG^2 \\
&= \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2a} + \sum_{t=1}^{T-1}\frac{1}{2a(\sqrt{t+1}+\sqrt{t})}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{\sqrt{T}}{2a}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{2\sqrt{T}-1}{2}aG^2 \\
&< \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2a} + \sum_{t=1}^{T-1}\frac{1}{4a\sqrt{t}}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{\sqrt{T}}{2a}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{2\sqrt{T}-1}{2}aG^2 \\
&= \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2a} + \frac{2\sqrt{T-1}-1}{4a}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 - \frac{\sqrt{T}}{2a}\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + \frac{2\sqrt{T}-1}{2}aG^2
\end{aligned}$$

参考 *slides* 中对定义域直径的假设，上述不等式可以继续化简为：

$$\begin{aligned}
\sum_{t=1}^{T}[f(\mathbf{w}_t)-f(\mathbf{w})] &\le \frac{\|\mathbf{w}_1-\mathbf{w}\|_2^2}{2a}+\frac{2\sqrt{T-1}-1}{4a}\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2-\frac{\sqrt{T}}{2a}\|\mathbf{w}_{T+1}-\mathbf{w}\|_2^2+\frac{2\sqrt{T}-1}{2}aG^2\\
&\le \frac{D^2}{2a}+\frac{2\sqrt{T-1}-1}{4a}D^2+\frac{2\sqrt{T}-1}{2}aG^2\\
&= \frac{D^2}{2a}\left(1+\frac{2\sqrt{T-1}-1}{2}\right)+\frac{2\sqrt{T}-1}{2}G^2\\
&< \frac{D^2}{2a}\left(1+\sqrt{T}\right)+\sqrt{T}aG^2
\end{aligned}$$

再由 *Jensen* 不等式便有 $f(\overline{\mathbf{w}}_T)-f(\mathbf{w})<\left(\frac{1}{T}\sum_{t=1}^{T}f(\mathbf{w}_t)\right)-f(\mathbf{w})<\frac{D^2}{2a}(\frac{1}{T}+\frac{1}{\sqrt{T}})+\frac{G^2}{\sqrt{T}}=O\left(\frac{1}{\sqrt{T}}\right)$ 成立。

4 [50pts] Stochastic Optimization

考虑随机优化问题

$$\min_{\mathbf{w}\in\mathcal{W}} \; F(\mathbf{w}) = \mathbb{E}_\xi\left[f(\mathbf{w},\xi)\right]$$

其中目标函数是 λ 强凸的，也就是

$$F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\lambda}{2}\|\mathbf{w}'-\mathbf{w}\|_2^2 \le F(\mathbf{w}'),\; \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{4.1}$$

试分析采用 $\eta_t = O(1/[\lambda t])$ 的随机梯度下降算法的额外风险。

(1) **[25pts]** 证明期望意义上的额外风险为 $O(\log T/T)$。

提示：该问题非常简单，可以将步长设置为 $\eta_t = 1/[\lambda t]$。然后，参考 [1] 中定理 1 的证明，得到

$$\sum_{t=1}^{T} F(\mathbf{w}_t) - TF(\mathbf{w}) \le \frac{G^2}{2\lambda}(\log T + 1) + \sum_{t=1}^{T}\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}\rangle.$$

接下来只需要求期望，化简即可。

(2) **[25pts]** 证明 $O(\log T/T)$ 的收敛速率同样以大概率成立。

第一种途径：可以将步长设置为 $\eta_t = 2/[\lambda t]$)。然后，参考 [1] 中定义 1 的证明，得到

$$\sum_{t=1}^{T} F(\mathbf{w}_t) - TF(\mathbf{w}) \le \frac{G^2}{\lambda}(\log T + 1) + \sum_{t=1}^{T}\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}\rangle - \frac{\lambda}{4}\sum_{t=1}^{T}\|\mathbf{w}_t - \mathbf{w}\|^2.$$

参考讲义公式 (21) 的推导过程，我们知道以至少 $1-\delta$ 的概率

$$\sum_{t=1}^{T}\langle \nabla F(\mathbf{w}_t) - \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}\rangle \le 2\sqrt{4G^2\log\frac{m}{\delta}\left(\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|^2\right)} + \frac{8G^2}{3\lambda}\log\frac{m}{\delta} + \frac{4G^2}{\lambda}.$$

最后对结果化简即可。

第二种途径：参考论文 [2]。

Proof.

(1) 设最优值为 $\mathbf{w}$，由 λ 强凸的性质可以知道：

$$\begin{aligned}
F(\mathbf{w}_t) - F(\mathbf{w}) &\le \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}\rangle - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 \\
&= \langle \nabla f(\mathbf{w}_t,\xi_t), \mathbf{w}_t - \mathbf{w}\rangle + \langle \nabla F(\mathbf{w}_t) - f(\mathbf{w}_t,\xi_t), \mathbf{w}_t - \mathbf{w}\rangle - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2
\end{aligned}$$

记 $\epsilon_t = \langle \nabla F(\mathbf{w}_t) - f(\mathbf{w}_t,\xi_t), \mathbf{w}_t - \mathbf{w}\rangle$ 进而有：

$$\begin{aligned}
F(\mathbf{w}_t) - F(\mathbf{w}) &\le \langle \nabla f(\mathbf{w}_t,\xi_t), \mathbf{w}_t - \mathbf{w}\rangle - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t \\
&= \frac{1}{\eta_t}\langle \mathbf{w}_t - \mathbf{w}'_{t+1}, \mathbf{w}_t - \mathbf{w}\rangle - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] + \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t \\
&= \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2\right] + \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t \\
&\le \frac{1}{2\eta_t}\left[\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right] + \frac{\eta_t}{2}G^2 - \frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \epsilon_t
\end{aligned}$$

将上述不等式从 $t=1$ 加到 T 有：

$$\begin{aligned}\sum_{t=1}^{T}F(\mathbf{w}_t)-TF(\mathbf{w}) &\le \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}\left[\|\mathbf{w}_t-\mathbf{w}\|_2^2-\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2\right]+\frac{\eta_t}{2}G^2-\frac{\lambda}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2+\epsilon_t\right)\\ &= \sum_{t=2}^{T}\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}-\frac{\lambda}{2}\right)\|\mathbf{w}_t-\mathbf{w}\|_2^2+\left(\frac{1}{2\eta_1}-\frac{\lambda}{2}\right)\|\mathbf{w}_1-\mathbf{w}\|_2^2\\ &\quad -\frac{1}{2\eta_{T+1}}\|\mathbf{w}_{T+1}-\mathbf{w}\|_2^2+\sum_{t=1}^{T}\frac{\eta_t}{2}G^2+\sum_{t=1}^{T}\epsilon_t\\ &= \sum_{t=2}^{T}\left(\frac{\lambda t}{2}-\frac{\lambda(t-1)}{2}-\frac{\lambda}{2}\right)\|\mathbf{w}_t-\mathbf{w}\|_2^2+\left(\frac{\lambda}{2}-\frac{\lambda}{2}\right)\|\mathbf{w}_1-\mathbf{w}\|_2^2\\ &\quad -\frac{1}{2\eta_{T+1}}\|\mathbf{w}_{T+1}-\mathbf{w}\|_2^2+\sum_{t=1}^{T}\frac{1}{2\lambda t}G^2+\sum_{t=1}^{T}\epsilon_t\\ &\le \sum_{t=1}^{T}\frac{1}{2\lambda t}G^2+\sum_{t=1}^{T}\epsilon_t<\frac{1+\ln T}{2\lambda}G^2+\sum_{t=1}^{T}\epsilon_t\end{aligned}$$

再由 *Jensen* 不等式可以得到 $F(\overline{\mathbf{w}}_T)-F(\mathbf{w})\le\frac{1}{T}\sum_{t=1}^{T}F(\mathbf{w}_t)-F(\mathbf{w})\le\frac{1+\ln T}{2\lambda T}G^2+\frac{1}{T}\sum_{t=1}^{T}\epsilon_t$，又因为 $\mathbb{E}[f(\mathbf{w}_t,\xi_t)]=F(\mathbf{w}_t)$，所以 $\mathbb{E}[\epsilon_t]=\mathbb{E}[\langle\nabla F(\mathbf{w}_t)-f(\mathbf{w}_t,\xi_t),\mathbf{w}_t-\mathbf{w}\rangle]=0$。最后对上述不等式两边求期望即可得到 $\mathbb{E}[F(\overline{\mathbf{w}}_T)]-F(\mathbf{w})\le\frac{1+\ln T}{2\lambda T}G^2=O(\frac{\ln T}{T})$。

(2) 令 $\eta_t=\frac{2}{\lambda t}$，则由 (1) 的推导可得

$$\begin{aligned}\sum_{t=1}^{T}F(\mathbf{w}_t)-TF(\mathbf{w}) &\le \sum_{t=1}^{T}\left(\frac{1}{2\eta_t}-\frac{1}{2\eta_{t-1}}-\frac{\lambda}{2}\right)\|\mathbf{w}_t-\mathbf{w}\|_2^2+\left(\frac{1}{2\eta_1}-\frac{\lambda}{2}\right)\|\mathbf{w}_1-\mathbf{w}\|_2^2\\ &\quad -\frac{1}{2\eta_{T+1}}\|\mathbf{w}_{T+1}-\mathbf{w}\|_2^2+\sum_{t=1}^{T}\frac{\eta_t}{2}G^2+\sum_{t=1}^{T}\epsilon_t\\ &= \sum_{t=2}^{T}\left(\frac{\lambda t}{4}-\frac{\lambda(t-1)}{4}-\frac{\lambda}{2}\right)\|\mathbf{w}_t-\mathbf{w}\|_2^2+\left(\frac{\lambda}{4}-\frac{\lambda}{2}\right)\|\mathbf{w}_1-\mathbf{w}\|_2^2\\ &\quad -\frac{1}{2\eta_{T+1}}\|\mathbf{w}_{T+1}-\mathbf{w}\|_2^2+\sum_{t=1}^{T}\frac{1}{\lambda t}G^2+\sum_{t=1}^{T}\epsilon_t\\ &\le \frac{1+\ln T}{\lambda}G^2+\sum_{t=1}^{T}\epsilon_t-\frac{\lambda}{4}\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|_2^2-\frac{1}{2\eta_{T+1}}\|\mathbf{w}_{T+1}\|_2^2\\ &\le \frac{1+\ln T}{\lambda}G^2+\sum_{t=1}^{T}\epsilon_t\\ \Longrightarrow \frac{1}{T}\sum_{t=1}^{T}F(\mathbf{w}_t)-F(\mathbf{w}) &\le \frac{1+\ln T}{\lambda T}G^2+\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\end{aligned}$$

记 $M(\lambda,\frac{1}{\delta},m)=2\sqrt{4G^2\log\frac{m}{\delta}\left(\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}\|^2\right)}+\frac{8G^2}{3\lambda}\log\frac{m}{\delta}+\frac{4G^2}{\lambda}$，由题目中 *slides* 的结论知 $\sum_{t=1}^{T}\epsilon_t\le M(\lambda,\frac{1}{\delta},m)$ 以至少 $1-\delta$ 的概率成立，所以 $\frac{1}{T}\sum_{t=1}^{T}F(\mathbf{w}_t)-F(\mathbf{w})\le\frac{1+\ln T}{\lambda T}G^2+\frac{1}{T}M(\lambda,\frac{1}{\delta},m)=O(\frac{\ln T}{T})$ 也以至少 $1-\delta$ 的概率成立，再代一步 *Jensen* 不等式即可得到 $\Pr[F(\overline{\mathbf{w}}_t)-F(\mathbf{w})\le\frac{1+\ln T}{\lambda T}G^2+\frac{1}{T}M(\lambda,\frac{1}{\delta},m)]\ge1-\delta$，所以 $O(\frac{\ln T}{T})$ 的收敛速率同样以大概率成立。

□

Reference

[1] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169-192, 2007.

[2] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, pages 801-808, 2009.