机器学习理论研究导引
作业五

陈晟 MG21330006

2022 年 5 月 20 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2022/05/31 23:59:59**, 截止时间后不再接收作业, 本次作业记零分;

(2) 作业提交方式：使用此 LaTex 模板书写解答，只需提交编译生成的 pdf 文件，将 pdf 文件上传到以下 ftp 服务器的指定位置：
地址：sftp://210.28.132.67:22，用户名：mlt2022，密码：mltspring2022@nju
文件夹位置：/C:/Users/mlt2022/hw_submissions/hw5_submission/；

(3) pdf 文件命名方式：学号-姓名-作业号-v 版本号, 例 MG1900000-张三-5-v1；如果需要更改已提交的解答，请在截止时间之前提交新版本的解答，并将版本号加一；

(4) 未按照要求提交作业，或 pdf 命名方式不正确，将会被扣除部分作业分数.

1 [50pts] Consistent Surrogate Loss

考虑对率函数 $\phi(t)=\log(1+\mathrm{e}^{-t})$, 回答并证明下述问题.

1. **[15pts]** 试求解最优实值输出函数 $f_\phi^*(\boldsymbol{x})$.
2. **[10pts]** 试求解最优实值输出函数对应的最优替代泛化风险 R_ϕ^*.
3. **[25pts]** 证明对率函数针对原 0/1 目标函数具有替代一致性.

Proof.

(1) 对率替代函数:$\phi(t)=log(1+e^{-t})$

替代函数 ϕ 在数据分布 $\mathcal{D}$ 上的泛化风险定义为

$$R_\phi(f)=\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\phi(yf(x))],$$

进一步定义最优替代泛化风险为

$$R_\phi^*=\min_f\left(R_\phi(f)\right).$$

由替代函数 ϕ 在数据分布 $\mathcal{D}$ 上的泛化风险定义为

$$R_\phi(f)=\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\phi(yf(x))],$$

可得

$$\begin{aligned}R_\phi(f)&=\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\phi(yf(\boldsymbol{x}))]\\&=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}[\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x}))]\end{aligned}$$

进一步根据

$$R_\phi^*=\min_f\left(R_\phi(f)\right).$$

可得

$$R_\phi^*=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[\min_{f(\boldsymbol{x})\in\mathbb{R}}(\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x})))\right],$$

从而得到替代函数的最优实值输出函数 $f_\phi^*(x)$ 为

$$f_\phi^*(\boldsymbol{x})\in\arg\min_{f(\boldsymbol{x})\in\mathbb{R}}(\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x}))).$$

将对率替代函数:$\phi(t)=log(1+e^{-t})$ 带入上式可得：最优实值输出函数 $f_\phi^*(x)$ 为

$$f_\phi^*(\boldsymbol{x})\in\arg\min_{f(\boldsymbol{x})\in\mathbb{R}}(\eta(\boldsymbol{x})log(1+e^{-f(x)})+(1-\eta(\boldsymbol{x}))log(1+e^{f(x)})).$$

假设 $\eta(\boldsymbol{x})$ 为常值，对 $(\eta(\boldsymbol{x})log(1+e^{-f(x)})+(1-\eta(\boldsymbol{x}))log(1+e^{f(x)}))$ 中的 $f(x)$ 求导可得：

$\because ln(1+e^{-f(x)})=\frac{-e^{-f(x)}}{1+e^{-f(x)}}$

$ln(1+e^{f(x)})=\frac{e^{f(x)}}{1+e^{f(x)}}$

$\therefore(\eta(\boldsymbol{x})log(1+e^{-f(x)})+(1-\eta(\boldsymbol{x}))log(1+e^{f(x)}))'$

$=\frac{e^{-f(x)}\eta(x)}{1+e^{-f(x)}}+\frac{e^{f(x)}(1-\eta(x))}{1+e^{f(x)}}$

$=\frac{e^{-f(x)}\eta(x)}{1+e^{-f(x)}}+\frac{(1-\eta(x))}{1+e^{-f(x)}}$

$= \frac{1-\eta(x)-e^{-f(x)}\eta(x)}{1+e^{-f(x)}}$

当 $1-\eta(x)-e^{-f(x)}\eta(x)=0$ 时, $f(x)=ln\frac{\eta(x)}{1-\eta(x)}$

即 $f_\phi^*(\boldsymbol{x})=ln\frac{\eta(x)}{1-\eta(x)}$

(2)

由 *(1)* 中

$$R_\phi^*=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[\min_{f(\boldsymbol{x})\in\mathbb{R}}(\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x})))\right],$$

最优泛化风险 $R_\phi^*=\min_f\left(R_\phi(f)\right)=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[\min_{f(\boldsymbol{x})\in\mathbb{R}}(\eta(\boldsymbol{x})\phi(f_\phi^*(\boldsymbol{x})))+(1-\eta(\boldsymbol{x}))\phi(-f_\phi^*(\boldsymbol{x})))\right]$

$=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[(\eta(\boldsymbol{x})\phi(ln\frac{\eta(x)}{1-\eta(x)}))+(1-\eta(\boldsymbol{x}))\phi(-ln\frac{\eta(x)}{1-\eta(x)}))\right]$

$=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[(\eta(\boldsymbol{x})ln(1+e^{-ln\frac{\eta(x)}{1-\eta(x)}})+(1-\eta(\boldsymbol{x}))ln(1+e^{ln\frac{\eta(x)}{1-\eta(x)}})\right]$

$=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_X}\left[(\eta(\boldsymbol{x})ln(1+\frac{1-\eta(x)}{\eta(x)})+(1-\eta(\boldsymbol{x}))ln(1+\frac{\eta(x)}{1-\eta(x)})\right]$

(3) 替代函数一致性的充分条件 *[Zhang, 2004b]*: 对替代函数 ϕ, 若最优实值输出函数满足 $f_\phi^*\in\mathcal{F}^*$, 且存在常数 $c>0$ 和 $s\geqslant 1$ 使

$$|\eta(\boldsymbol{x})-1/2|^s\leqslant c^s\left(\phi(0)-\eta(\boldsymbol{x})\phi\left(f_\phi^*(\boldsymbol{x})\right)-(1-\eta(\boldsymbol{x}))\phi\left(-f_\phi^*(\boldsymbol{x})\right)\right),$$

则替代函数 ϕ 具在一致性.

上面定理的证明如下：对任意函数 f 和样本 $x\in\mathcal{X}$, 记

$$\begin{aligned}\Delta_1(\boldsymbol{x})=&\eta(\boldsymbol{x})\mathbb{I}(f(\boldsymbol{x})\leqslant 0)\\&+(1-\eta(\boldsymbol{x}))\mathbb{I}(f(\boldsymbol{x})\geqslant 0)-\min\{\eta(\boldsymbol{x}),1-\eta(\boldsymbol{x})\}\end{aligned}$$

根据 $\begin{aligned}R(f)&=\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(yf(x)\leqslant 0)]\\&=\mathbb{E}_{x\sim\mathcal{D}_x}[\eta(x)\mathbb{I}(f(x)\leqslant 0)+(1-\eta(x))\mathbb{I}(f(x)\geqslant 0)]\end{aligned}$ 有

$$R(f)-R^*=\mathbb{E}_{x\sim D_X}\left[\Delta_1(x)\right].$$

根据 $\eta(\boldsymbol{x})-1/2$ 和 $f(\boldsymbol{x})$ 的不同取值, 下面分五种情况讨论 $\Delta_1(\boldsymbol{x})$: - 当 $\eta(x)>1/2$ 且 $f(x)>0$ 时, 有 $\Delta_1(x)=0$; - 当 $\eta(x)>1/2$ 且 $f(x)\leqslant 0$ 时, 有 $\Delta_1(x)=2\eta(x)-1$; - 当 $\eta(x)<1/2$ 且 $f(x)\geqslant 0$ 时, 有 $\Delta_1(x)=1-2\eta(x)$; - 当 $\eta(\boldsymbol{x})<1/2$ 且 $f(\boldsymbol{x})<0$ 时, 有 $\Delta_1(\boldsymbol{x})=0$; -当 $\eta(x)=1/2$ 时, 无论 $f(x)$ 取何值, 都有 $\Delta_1(x)=0$. 综合上述五种$\boxed{\text{F}}$况可得

$$\Delta_1(x)=2\mathbb{I}((\eta(x)-1/2)f(x)\leqslant 0)|\eta(x)-1/2|$$

代入之前得公式有

$$R(f)-R^*=2\mathbb{E}_{(\eta(x)-1/2)f(x)\leqslant 0}[|\eta(x)-1/2|].$$

对 $s\geqslant 1$, 根据 *Jensen* 不等式 (1.11) 有 $(\mathbb{E}[X])^s\leqslant\mathbb{E}\left[X^s\right]$, 于是有

$$R(f)-R^*\leqslant 2\sqrt[3]{\mathbb{E}_{(\eta(x)-1/2)f(x)}\xi 0\left[|\eta(x)-1/2|^8\right]}.$$

根据 $\begin{aligned}R_\phi(f)&=\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\phi(yf(\boldsymbol{x}))]\\&=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\boldsymbol{x}}}[\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x}))].\end{aligned}$ ，分别令

$$\begin{aligned}\Delta_2(\boldsymbol{x})&=\eta(\boldsymbol{x})\phi(f(\boldsymbol{x}))+(1-\eta(\boldsymbol{x}))\phi(-f(\boldsymbol{x}))\\\Delta_3(\boldsymbol{x})&=\eta(\boldsymbol{x})\phi\left(f_\phi^*(\boldsymbol{x})\right)+(1-\eta(\boldsymbol{x}))\phi\left(-f_\phi^*(\boldsymbol{x})\right)\end{aligned}$$

结合书中 (6.40) 和定理 6.1 中条件 *(6.35)* 可得

$$R(f)-R^* \leqslant 2c\sqrt[n]{\mathrm{E}_{(\eta(x)-1/2)f(x)\leqslant 0}\left[\phi(0)-\Delta_3(x)\right]}$$

另一方面, 根据书中 (6.32) $\sim$ (6.34) 有

$$R_\phi(f)-R_\phi^* \geqslant \mathbb{E}_{(\eta(x)-1/2)f(x)\leqslant 0}\left[\Delta_2(x)-\Delta_3(x)\right]$$

设函数

$$\Gamma(t)=\eta(\boldsymbol{x})\phi(t)+(1-\eta(\boldsymbol{x}))\phi(-t),$$

易知 $\Gamma(f(x))=\Delta_2(x)$ 和 $\Gamma(0)=\phi(0)$. 根据凸函数性质可知, 当 $\phi(t)$ 是凸函数时 $\Gamma(t)$ 也是凸函数, 以及当 $0\in[a,b]$ 时有

$$\Gamma(0)\leqslant\max\{\Gamma(a),\Gamma(b)\}$$

下面分三种情况讨论: - 若 $\eta(x)>1/2$, 由之前得式子可知 $f_\phi^*(x)>0$, 以及由 $(\eta(x)-1/2)f(x)\leqslant 0$ 可知 $f(x)\leqslant 0$. 因此 $0\in\left[f(x),f_\phi^*(x)\right]$, 进一步有

$$\phi(0)=\Gamma(0)\leqslant\max\left\{\Gamma(f(x)),\Gamma\left(f_\phi^*(\boldsymbol{x})\right)\right\}$$

根据 (6.33) 有 $\Gamma(f(x))\geqslant\Gamma\left(f_\phi^*(\boldsymbol{x})\right)$, 于是得到 $\phi(0)\leqslant\Gamma(f(x))=\Delta_2(\boldsymbol{x})$.

- 若 $\eta(x)<1/2$, 同理有 $f(x)\geqslant 0, f_\phi^*(x)<0$ 和 $0\in\left[f_\phi^*(x),f(x)\right]$, 以及

$$\phi(0)=\Gamma(0)\leqslant\max\left\{\Gamma(f(x)),\Gamma\left(f_\phi^*(x)\right)\right\}=\Gamma(f(x))=\Delta_2(x)$$

- 若 $\eta(x)=1/2$, 对凸函数 ϕ 有

$$\phi(0)\leqslant\phi(f(x))/2+\phi(-f(x))/2=\Delta_2(x)$$

综合上述三种情况, 我们有

$$\phi(0)\leqslant\Delta_2(x)$$

根据之前得式子, 对任何函数 f 有

$$\begin{aligned}
R(f)-R^* &\leqslant 2c\sqrt[3]{\mathbb{E}_{(\eta(x)-1/2)f(x)\leqslant 0}\left[\phi(0)-\Delta_3(x)\right]}\\
&=2c\sqrt[s]{\mathbb{E}_{(\eta(x)-1/2)f(x)\leqslant 0}\left[\phi(0)-\Delta_2(x)+\Delta_2(\boldsymbol{x})-\Delta_3(\boldsymbol{x})\right]}\\
&\leqslant 2c\sqrt[s]{\mathbb{E}_{(\eta(x)-1/2)f(x)\leqslant 0}\left[\Delta_2(x)-\Delta_3(x)\right]}\\
&\leqslant 2c\sqrt[s]{R_\phi(f)-R_\phi^*}
\end{aligned}$$

对任何函数列 $\hat{f}_1,\hat{f}_2,\cdots,\hat{f}_m,\cdots$, 根据上式可知

$$R\left(\hat{f}_m\right)-R^*\leqslant 2c\sqrt{R_\phi\left(\hat{f}_m\right)-R_\phi^*},$$

因此当 $R_\phi\left(\hat{f}_m\right)\to R_\phi^*$ 时有 $R\left(\hat{f}_m\right)\to R^*$ 成立, 定理得证.

$f^*\in\mathcal{F}^*=\{f:$ 当 $\eta(x)=1/2$ 时 $f(x)$ 可以是任意的实数; 当 $\eta(x)\neq 1/2$ 时 $f(x)(\eta(x)-1/2)>0\}$.

显然 $f_\phi^* \in \mathcal{F}^*$ 成立，其次，看是否存在常数 $c>0$ 和 $s \geqslant 1$ 使

$$|\eta(\boldsymbol{x})-1/2|^s \leqslant c^s\left(\phi(0)-\eta(\boldsymbol{x})\phi\left(f_\phi^*(\boldsymbol{x})\right)-(1-\eta(\boldsymbol{x}))\phi\left(-f_\phi^*(\boldsymbol{x})\right)\right),$$

将 $\phi(x)=\ln\left(1+\mathrm{e}^{-x}\right)$ 以及 $f_\phi^*(\boldsymbol{x})=ln\frac{\eta(x)}{1-\eta(x)}$ 带入上式得:

$$|\eta(\boldsymbol{x})-1/2|^s \leqslant c^s\left(ln2-\eta(x)ln(ln\frac{\eta(x)}{1-\eta(x)})-(1-\eta(\boldsymbol{x}))ln(ln\frac{\eta(x)}{1-\eta(x)})\right),$$

$=c^s\left(ln2+ln(ln\frac{1-\eta(x)}{\eta(x)})\right)=c^s\left(ln2|(ln\frac{1-\eta(x)}{\eta(x)})|\right)$

可以发现后面一项当 $\eta(x)=1/2$ 时或 $\eta(x)\neq 1/2$，都存在常数 $c>0$ 和 $s \geqslant 1$ 使上式成立，即证对率函数针对原 0/1 目标函数具有替代一致性

2 [50pts] Convergence Rate

试分析下述情形下梯度下降算法 (课件第 9 页, 课本 7.2.1 节) 的收敛率.

- 可行域 $\mathcal{W}$ 是凸集.
- 目标函数 f 是 $\mathcal{W}$ 上可微的 α-强凸函数.
- 目标函数在 $\mathcal{W}$ 上是 l-Lipschitz 连续的, 即

$$\forall \boldsymbol{u} \in \mathcal{W}, \|\nabla f(\boldsymbol{u})\| \leqslant l. \tag{2.1}$$

- 梯度下降算法的步长为:

$$\forall t, \eta_t = \frac{1}{\alpha t}. \tag{2.2}$$

Proof.

(1) 当可行域 $\mathcal{W}$ 是凸集时

若函数为凸函数, 则其定义域为凸集，满足条件，其可以采用梯度下降法达到 $O(1/\sqrt{T})$ 的收敛率 *[Nesterov 2018]* 其基本流程如下：*1.* 任意初始化 $w_1 \in \mathcal{W}$

2.for $t = 1, \ldots, T$ *do*

3. 梯度下降:$w'_{t+1} = w_t - \eta_t \nabla f(w_t)$

4. 投影:$\boldsymbol{w}_{t+1} = \Pi_{\mathcal{W}}\left(\boldsymbol{w}'_{t+1}\right)$

5.end for

6. 返回 $\overline{\boldsymbol{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t$

(2) 目标函数 f 是 $\mathcal{W}$ 上可微的 α-强凸函数

考虑目标函数 f:$\mathcal{W} \mapsto \mathbb{R}$ 是 α-强凸函数, 对 α-强凸函数 f 有下面性质，若 $\boldsymbol{w}^*$ 为其最优解, 对于任意 $\boldsymbol{w} \in \mathcal{W}$ 有：

$f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geqslant \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$

此外，若梯度有上界 l，则 $\|w - \boldsymbol{w}^*\| \leqslant \frac{2l}{\lambda}$

$f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \leqslant \frac{2l^2}{\lambda}$

如果考虑强凸且光滑的函数 f，即 α-强凸函数可微时，还满足以下的条件：

$f(\boldsymbol{w}') \leqslant f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w}\rangle + \frac{\gamma}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2, \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$

对可微的 α-强凸函数梯度下降时，其本值任然时进行梯度下降更新后再投影到可行域，对于其梯度下降算法有如下定理 *[Nesterov,2013]:*

若目标函数满足可微的 α-强凸函数，或 α-强凸函数且光滑，梯度下降取得了线性收敛率 $O\left(\frac{1}{\beta^T}\right)$，其中 $\beta > 1$

证明：根据目标函数的性质以及更新公式

$f(\boldsymbol{w}_{t+1}) \leqslant f(\boldsymbol{w}_t) + \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle + \frac{\gamma}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2$

$= \min_{\boldsymbol{w}\in\mathcal{W}}\left(f(\boldsymbol{w}_t) + \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t\rangle + \frac{\gamma}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2\right)$

$\leqslant \min_{\boldsymbol{w}\in\mathcal{W}}\left(f(\boldsymbol{w}) - \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \frac{\gamma}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2\right)$

$\leqslant \min_{\substack{w=\alpha w^* + (1-\alpha)\boldsymbol{w}_t \\ \alpha\in[0,1]}}\left(f(\boldsymbol{w}) + \frac{\gamma-\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2\right)$

$= \min_{\alpha\in[0,1]}\left(f\left(\alpha \boldsymbol{w}^*+(1-\alpha)\boldsymbol{w}_t\right)+\frac{\gamma-\lambda}{2}\left\|\alpha \boldsymbol{w}^*+(1-\alpha)\boldsymbol{w}_t-\boldsymbol{w}_t\right\|^2\right)$
$\leqslant \min_{\alpha\in[0,1]}\left(\alpha f\left(\boldsymbol{w}^*\right)+(1-\alpha) f\left(\boldsymbol{w}_t\right)+\frac{\gamma-\lambda}{2}\alpha^2\left\|\boldsymbol{w}^*-\boldsymbol{w}_t\right\|^2\right)$
$= \min_{\alpha\in[0,1]}\left(f\left(\boldsymbol{w}_t\right)-\alpha\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)+\frac{\gamma-\lambda}{2}\alpha^2\left\|\boldsymbol{w}^*-\boldsymbol{w}_t\right\|^2\right)$
$\leqslant \min_{\alpha\in[0,1]}\left(f\left(\boldsymbol{w}_t\right)-\alpha\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)+\frac{\gamma-\lambda}{2}\frac{2}{\lambda}\alpha^2\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)\right)$
$= \min_{\alpha\in[0,1]}\left(f\left(\boldsymbol{w}_t\right)+\left(\frac{\gamma-\lambda}{\lambda}\alpha^2-\alpha\right)\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)\right)$

如果 $\frac{\lambda}{2(\gamma-\lambda)}\geqslant 1$, 令 $\alpha=1$, 则有

$$f\left(\boldsymbol{w}_{t+1}\right)-f\left(\boldsymbol{w}^*\right)\leqslant\frac{\gamma-\lambda}{\lambda}\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)\leqslant\frac{1}{2}\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)$$

如果 $\frac{\lambda}{2(\gamma-\lambda)}<1$, 令 $\alpha=\frac{\lambda}{2(\gamma-\lambda)}$, 则有

$$\begin{aligned}f\left(\boldsymbol{w}_{t+1}\right)-f\left(\boldsymbol{w}^*\right)&\leqslant\left(1-\frac{\lambda}{4(\gamma-\lambda)}\right)\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)\\&=\frac{4\gamma-5\lambda}{4(\gamma-\lambda)}\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)\end{aligned}$$

结合书中 (7.20) 和 *(7.21)*, 令

$$\beta=\begin{cases}\frac{\lambda}{\gamma-\lambda}, & \frac{\lambda}{2(\gamma-\lambda)}\geqslant 1\\ \frac{4(\gamma-\lambda)}{4\gamma-5\lambda}, & \frac{\lambda}{2(\gamma-\lambda)}<1\end{cases}$$

那么下式总是成立

$$f\left(\boldsymbol{w}_{t+1}\right)-f\left(\boldsymbol{w}^*\right)\leqslant\frac{1}{\beta}\left(f\left(\boldsymbol{w}_t\right)-f\left(\boldsymbol{w}^*\right)\right)$$

将上式扩展可得

$$f\left(\boldsymbol{w}_T\right)-f\left(\boldsymbol{w}^*\right)\leqslant\frac{1}{\beta^{T-1}}\left(f\left(\boldsymbol{w}_1\right)-f\left(\boldsymbol{w}^*\right)\right)=O\left(\frac{1}{\beta^T}\right)$$

即证可微的 α-强凸函数，或 α-强凸函数且光滑，梯度下降取得了线性收敛率 $O\left(\frac{1}{\beta^T}\right)$，其中 $\beta>1$

(3) 目标函数在 $\mathcal{W}$ 上是 l-*Lipschitz* 连续的, 即 $\forall\boldsymbol{u}\in\mathcal{W},\|\nabla f(\boldsymbol{u})\|\leqslant l$

梯度下降收玫率若目标函数是 l-*Lipschitz* 连续函数, 且可行域有界, 则采用固定步长梯度下降的收玫率为 $O\left(\frac{1}{\sqrt{T}}\right)$. 证明假设可行域 $\mathcal{W}$ 直径为 Γ, 并且目标函数满足 l-*Lipschitz* 连续, 即对于任意 $\boldsymbol{u},\boldsymbol{v}\in\mathcal{W}$,

$$\|\boldsymbol{u}-\boldsymbol{v}\|\leqslant\Gamma,\|\nabla f(\boldsymbol{u})\|\leqslant l.$$

为了简化分析, 考虑固定的步长 $\eta_t=\eta$. 对于任意的 $\boldsymbol{w}\in\mathcal{W}$,

$$f\left(\boldsymbol{w}_t\right)-f(\boldsymbol{w})\leqslant\left\langle\nabla f\left(\boldsymbol{w}_t\right),\boldsymbol{w}_t-\boldsymbol{w}\right\rangle=\frac{1}{\eta}\left\langle\boldsymbol{w}_t-\boldsymbol{w}_{t+1}',\boldsymbol{w}_t-\boldsymbol{w}\right\rangle$$

$$\begin{aligned}&=\frac{1}{2\eta}\left(\left\|\boldsymbol{w}_t-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{t+1}'-\boldsymbol{w}\right\|^2+\left\|\boldsymbol{w}_t-\boldsymbol{w}_{t+1}'\right\|^2\right)\\&=\frac{1}{2\eta}\left(\left\|\boldsymbol{w}_t-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{t+1}'-\boldsymbol{w}\right\|^2\right)+\frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{w}_t\right)\right\|^2\\&\leqslant\frac{1}{2\eta}\left(\left\|\boldsymbol{w}_t-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\right\|^2\right)+\frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{w}_t\right)\right\|^2\end{aligned}$$

最后一个不等号利用了凸集合投影操作的非扩展性质 *[Nemirovski et al., 2009]:*

$$\|\Pi_{\mathcal{W}}(x)-\Pi_{\mathcal{W}}(z)\| \leqslant \|\boldsymbol{x}-\boldsymbol{z}\|, \quad \forall x, z.$$

注意到目标函数满足 *l-Lipschitz* 连续, 由上面的公式可得

$$f\left(\boldsymbol{w}_t\right)-f(\boldsymbol{w}) \leqslant \frac{1}{2\eta}\left(\left\|\boldsymbol{w}_t-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\right\|^2\right)+\frac{\eta}{2} l^2.$$

对上式从 $t=1$ 到 T 求和, 有

$$\begin{aligned}
\sum_{t=1}^T f\left(\boldsymbol{w}_t\right)-T f(\boldsymbol{w}) & \leqslant \frac{1}{2\eta}\left(\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{T+1}-\boldsymbol{w}\right\|^2\right)+\frac{\eta T}{2} l^2 \\
& \leqslant \frac{1}{2\eta}\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2+\frac{\eta T}{2} l^2 \leqslant \frac{1}{2\eta} \Gamma^2+\frac{\eta T}{2} l^2
\end{aligned}$$

最后, 依据 *Jensen* 不等式可得

$$\begin{aligned}
f\left(\overline{\boldsymbol{w}}_T\right)-f(\boldsymbol{w}) & =f\left(\frac{1}{T} \sum_{t=1}^T \boldsymbol{w}_t\right)-f(\boldsymbol{w}) \\
& \leqslant \frac{1}{T} \sum_{t=1}^T f\left(\boldsymbol{w}_t\right)-f(\boldsymbol{w}) \leqslant \frac{\Gamma^2}{2\eta T}+\frac{\eta l^2}{2}
\end{aligned}$$

因此,

$$f\left(\overline{\boldsymbol{w}}_T\right)-\min_{w \in \mathcal{W}} f(\boldsymbol{w}) \leqslant \frac{\Gamma^2}{2\eta T}+\frac{\eta l^2}{2}=\frac{l \Gamma}{\sqrt{T}}=O\left(\frac{1}{\sqrt{T}}\right)$$

其中步长设置为 $\eta=\Gamma /(l \sqrt{T})$ 定理得证

(4) 当步长设置为变长 $\forall t, \eta_t=\frac{1}{\alpha t}$ 时的收敛率

以 *(3)* 中的条件更改收敛率进行推算

对 *(3)* 中间步骤, 从 $t=1$ 到 T 求和, 有

$$\begin{aligned}
\sum_{t=1}^T f\left(\boldsymbol{w}_t\right)-T f(\boldsymbol{w}) & \leqslant \frac{1}{2\eta}\left(\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{T+1}-\boldsymbol{w}\right\|^2\right)+\frac{\eta T}{2} l^2 \\
& \leqslant \frac{1}{2\eta}\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2+\frac{\eta T}{2} l^2 \leqslant \frac{1}{2\eta} \Gamma^2+\frac{\eta T}{2} l^2
\end{aligned}$$

将 $\eta_t=\frac{1}{\alpha t}$ 带入上式可得:

$f\left(\boldsymbol{w}_t\right)-f(\boldsymbol{w}) \leqslant \frac{1}{2\alpha t}\left(\left\|\boldsymbol{w}_t-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\right\|^2\right)+\frac{\alpha t}{2} l^2$

$$\sum_{t=1}^T f\left(\boldsymbol{w}_t\right)-T f(\boldsymbol{w}) \leqslant \frac{1}{2\alpha}\left(\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2-\left\|\boldsymbol{w}_{T+1}-\boldsymbol{w}\right\|^2\right)+\frac{\alpha T^2}{2} l^2$$

$$\leqslant \frac{1}{2\alpha}\left\|\boldsymbol{w}_1-\boldsymbol{w}\right\|^2+\frac{\alpha T^2}{2} l^2 \leqslant \frac{1}{2\alpha} \Gamma^2+\frac{\alpha T^2}{2} l^2$$

最后, 依据 *Jensen* 不等式可得

$$\begin{aligned}
f\left(\overline{\boldsymbol{w}}_T\right)-f(\boldsymbol{w}) & =f\left(\frac{1}{T} \sum_{t=1}^T \boldsymbol{w}_t\right)-f(\boldsymbol{w}) \\
& \leqslant \frac{1}{T} \sum_{t=1}^T f\left(\boldsymbol{w}_t\right)-f(\boldsymbol{w}) \leqslant \frac{\Gamma^2}{2\alpha T}+\frac{\alpha T l^2}{2}
\end{aligned}$$

因此，

$$f\left(\overline{\boldsymbol{w}}_T\right)-\min_{\boldsymbol{w} \in \mathcal{W}} f(\boldsymbol{w}) \leqslant \frac{\Gamma^2}{2 \alpha T}+\frac{\alpha T l^2}{2}$$

当 $\frac{\alpha T l^2}{2}=\frac{\Gamma^2}{2 alpha T}$ 时，$\alpha=\frac{\Gamma}{T l}$

$f\left(\overline{\boldsymbol{w}}_T\right)-\min_{\boldsymbol{w} \in \mathcal{W}} f(\boldsymbol{w}) \leqslant l \Gamma=O(1)$

即其收敛率变为常数级别了