

深度学习中学习率策略的研究

朱梦

初稿于 2025-05-02, 修改于 2025-06-23

1. 学习率适中

学习率太大容易导致发散, 太小需要迭代的步数过多。因此, 从“节能”和“加速”的角度来看, 学习率不宜过小。如果不考虑算力和时间, 那么过小的学习率是否可取呢? 设损失函数为 $\mathcal{L}(\theta)$, 由泰勒级数有:

$$d\mathcal{L}(\theta) \approx \langle d\theta, \nabla_{\theta}\mathcal{L}(\theta) \rangle \triangleq \langle d\theta, g \rangle \quad (1)$$

如果将 θ 视作为沿着某种时间参数 t 变换的轨迹 $\theta(t)$, 那么考虑式(1)的变化率有:

$$\frac{d\mathcal{L}(\theta(t))}{dt} = \left\langle \frac{d\theta(t)}{dt}, g(t) \right\rangle = \langle \theta^{(1)}(t), g(t) \rangle \quad (2)$$

其中, $\theta^{(1)}(t)$ 表示 $\theta(t)$ 的一阶导函数。希望 $\mathcal{L}(\theta(t))$ 随着 t 的递增而递减 (损失值越小越好), 所以满足: $\frac{d\mathcal{L}(\theta(t))}{dt} \leq 0$ 。当 $\|\theta^{(1)}(t)\|_2$ 固定时, 上式右端的最小值在梯度的反方向 $-g(t)$ 取得, 故梯度的负方向为最速下降方向:

$$\theta^{(1)}(t) = -g(t) \quad (3)$$

那么求解参数 θ 转换为式(3)所示的常微分方程。可以证明式(3)最终可以收敛至一个不动点 ($\theta^{(1)}(t) = \lim_{t \rightarrow \infty} g(t) = \mathbf{0}$), 且此不动点为极小值点。

然而, 不知损失函数为 $\mathcal{L}(\theta)$ 的具体表达式, 故无法求出解析解。采用欧拉法求解:

$$\frac{\theta(t) - \theta(t - \eta)}{\eta} = -g(t) \quad (4)$$

即:

$$\theta(t) = \theta(t - \eta) - \eta g(t) \quad (5)$$

这就是最朴素的梯度下降法, 其中 η 表示学习率。利用泰勒级数展开 $\theta(t)$:

$$\theta(t) = \theta(t - \eta) + \eta \theta^{(1)}(t - \eta) + \frac{1}{2} \eta^2 \theta^{(2)}(t - \eta) + \cdots + \frac{1}{n!} \eta^n \theta^{(n)}(t - \eta) \quad (6)$$

省略剩余证明过程，可以证明：

$$\theta^{(1)}(t) = -(\mathbf{g}(t) + \frac{1}{4}\eta\|\mathbf{g}(t)\|_2^2) \quad (7)$$

由式(7)可知，离散化的迭代过程隐式地带来了梯度惩罚项，而梯度惩罚项有助于模型抵达更加平滑的区域，有利于提升泛化性能，故 η 不能趋于零。

2. 学习率衰减策略

(一) 设输入为字典 $\{t_1 : r_{\min}\}$ 表示：当 $0 < t \leq t_1$ 时，学习率衰减因子 r_t 从 1 均匀递减至 r_{\min} ；当 $t > t_1$ 时，学习率衰减因子 r_t 保持 r_{\min} 固定不变。那么有：

$$r_t = \begin{cases} 1 + \frac{r_{\min}-1}{t_1}t, & \text{如果 } 0 < t \leq t_1, \\ r_{\min}, & \text{如果 } t > t_1 \end{cases} \quad (8)$$

(二) 设输入为字典 $\{t_1 : 1, t_2 : r_{\min}\}$ 表示：当 $0 < t \leq t_1$ 时，学习率衰减因子 r_t 从 0 均匀递增至 1（即 **warmup**）；当 $t_1 < t \leq t_2$ 时，学习率衰减因子 r_t 从 1 均匀递减至 r_{\min} ；当 $t > t_2$ 时，学习率衰减因子 r_t 保持 r_{\min} 固定不变。那么有：

$$r_t = \begin{cases} \frac{t}{t_1}, & \text{如果 } 0 < t \leq t_1, \\ 1 + \frac{r_{\min}-1}{t_2-t_1}(t-t_1), & \text{如果 } t_1 < t \leq t_2, \\ r_{\min}, & \text{如果 } t > t_2 \end{cases} \quad (9)$$

(三) 设输入为字典 $\{t_1 : r_{\min}, t_2 : r_{\min}, t_3 : 1\}$ 表示：当 $0 < t \bmod t_3 \leq t_1$ 时，学习率衰减因子 r_t 从 1 均匀递减至 r_{\min} ；当 $t_1 < t \bmod t_3 \leq t_2$ 时，学习率衰减因子 r_t 保持 r_{\min} 固定不变；当 $t_2 < t \bmod t_3 \leq t_3$ 时，学习率衰减因子 r_t 从 r_{\min} 均匀递增至 1。那么有：

$$r_t = \begin{cases} 1 + \frac{r_{\min}-1}{t_1}(t \bmod t_3), & \text{如果 } 0 < t \bmod t_3 \leq t_1, \\ r_{\min}, & \text{如果 } t_1 < t \bmod t_3 \leq t_2, \\ r_{\min} + \frac{1-r_{\min}}{t_3-t_2}(t \bmod t_3 - t_2), & \text{如果 } t_2 < t \bmod t_3 \leq t_3 \end{cases} \quad (10)$$

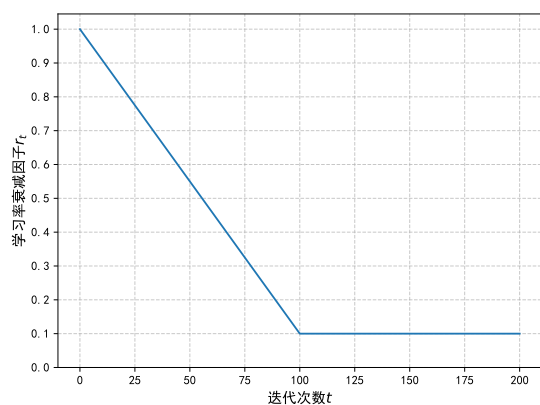
(四) 设输入为字典 $\{t_1 : 1, t_2 : r_{\min}, t_3 : r_{\min}, t_4 : 1\}$ 表示：当 $0 < t \leq t_1$ 时，学习率衰减因子 r_t 从 0 均匀递增至 1（即 **warmup**）；当 $t_1 < t \leq t_2$ 时，学习率衰减因子 r_t 从 1 均匀递减至 r_{\min} ；当 $t_2 < t \leq t_3$ 时，学习率衰减因子 r_t 保持

r_{\min} 固定不变；当 $t_3 < t \leq t_4$ 时，学习率衰减因子 r_t 从 r_{\min} 均匀递增至 1；对于 $t > t_4$ ，周期重复 $t_1 \sim t_4$ 。那么有：

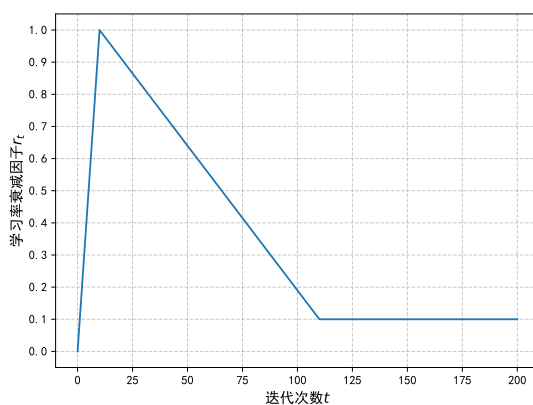
$$r(t) = \begin{cases} \frac{t}{t_1}, & 0 < t \leq t_1 \\ 1 + \frac{r_{\min}-1}{t_2-t_1} \hat{t}, & t > t_1 \text{ 且 } \hat{t} \leq t_2 - t_1 \\ r_{\min}, & t > t_1 \text{ 且 } t_2 - t_1 < \hat{t} \leq t_3 - t_1 \\ r_{\min} + \frac{1-r_{\min}}{t_4-t_3} (\hat{t} - (t_3 - t_1)), & t > t_1 \text{ 且 } t_3 - t_1 < \hat{t} < t_4 - t_1 \end{cases} \quad (11)$$

$$\hat{t} = (t - t_1) \bmod (t_4 - t_1)$$

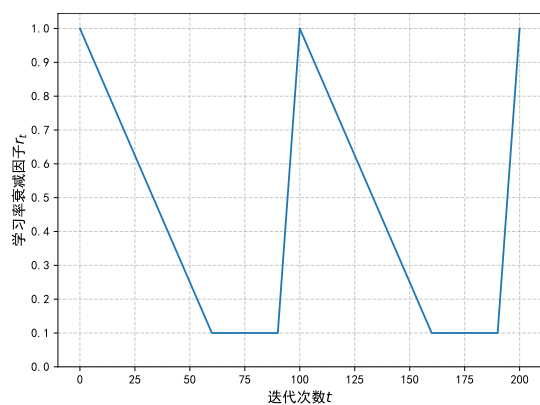
（五）总结。图1可视化了上述四种学习率衰减策略。对于 **post-norm** 结构或深层模型，建议启用 **warmup**。对于希望探索更多可能性，建议启用周期。结合博客《优化算法的分析及改进（一）：基于动量累积或不同参数元素更新步长相同的优化算法》可得，对于 **SGD/SGDM** 类优化算法， r_{\min} 可设置为 0.1 或者 0.01；对于 **Adam** 类优化算法， r_{\min} 可设置为 0.1 或者 0.01；对于 **Lion/Tiger** 类优化算法， r_{\min} 可设置为 $1e-4$ 或者 $1e-5$ 。



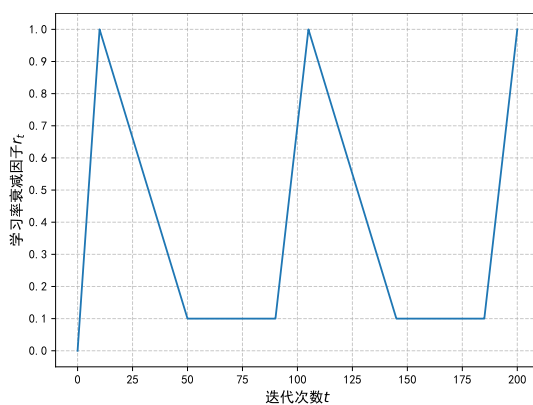
(a) 无 warmup, 无周期



(b) 有 warmup, 无周期



(c) 无 warmup, 有周期



(d) 有 warmup, 有周期

图 1 学习率衰减策略