Supplementary Materials for Understanding Short-Horizon Bias in Stochastic Meta-Optimization

Yuhuai Wu[∗ 1,2], **Mengye Ren**[∗ 1,2,3], **Renjie Liao**[1,2,3], **Roger B. Grosse**[1,2]

[1]University of Toronto
[2]Vector Institute for Artificial Intelligence
[3]Uber Advanced Technologies Group
{ywu, mren, rjliao, rgrosse}@cs.toronto.edu

A Proofs of Theorems

A.1 Model Dynamics

Several observations allow us to compactly model the dynamics following SGD with momentum on the noisy quadratic model. First, observe that the expected loss can be expressed in terms of the variables $\mathbb{E}[\theta_i]$ and $\mathbb{V}[\theta_i]$. Second, because the Hessians and noise covariances are diagonal, each coordinate evolves independently of the others. Third, the second-order statistics of the iterate $\theta_i^{(t)}$ and the velocity $v_i^{(t)}$ can be expressed in terms of the second-order statistics of $\theta_i^{(t-1)}$ and $v_i^{(t-1)}$. Finally, $\mathbb{E}[\theta_i^{(t)}] = \mathbb{E}[m_i^{(t)}] = 0$ for all t. Combining these observations, we derive the dynamics of SGD with momentum as a *deterministic* recurrence relation with sufficient statistics $\mathbb{V}[\theta^{(t)}]$, $\mathbb{V}[v^{(t)}]$, and $\Sigma_{\theta,v}^{(t)} = \mathrm{Cov}(\theta^{(t)}, v^{(t)})$. Note that we drop the dimension subscripts because the dimensions evolve independently. The dynamics are as follows:

Theorem 1 (Mean and variance dynamics)**.** *The expectations of the parameter θ and the velocity v are updated as,*

$$\begin{aligned}
\mathbb{E}[v^{(t+1)}] &= \mu^{(t)}\mathbb{E}[v^{(t)}] - (\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}] \\
\mathbb{E}[\theta^{(t+1)}] &= \mathbb{E}[\theta^{(t)}] + \mathbb{E}[v^{(t+1)}]
\end{aligned}$$

The variances of the parameter θ and the velocity v are updated as,

$$\begin{aligned}
\mathbb{V}[v^{(t+1)}] &= (\mu^{(t)})^2\mathbb{V}[v^{(t)}] + (\alpha^{(t)}h)^2\mathbb{V}[\theta^{(t)}] - 2\mu^{(t)}\alpha^{(t)}h\Sigma_{\theta,v}^{(t)} + (\alpha^{(t)}h\sigma)^2 \\
\mathbb{V}[\theta^{(t+1)}] &= (1 - 2\alpha^{(t)}h)\mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}] + 2\mu^{(t)}\Sigma_{\theta,v}^{(t)} \\
\Sigma_{\theta,v}^{(t+1)} &= \mu^{(t)}\Sigma_{\theta,v}^{(t)} - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}]
\end{aligned}$$

Proof. The stochastic gradient descent with momentum is defined as follows,

$$\begin{aligned}
v^{(t+1)} &= \mu^{(t)}v^{(t)} - \alpha^{(t)}(h\theta^{(t)} + h\sigma\xi), \quad \xi \sim \mathcal{N}(0,1) \\
\theta^{(t+1)} &= \theta^{(t)} + v^{(t+1)}
\end{aligned}$$

[∗]Equal contribution.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Dynamics of the expectation We calculate the mean of the velocity $v^{(t+1)}$,

$$\begin{aligned}\mathbb{E}[v^{(t+1)}] &= \mathbb{E}[\mu^{(t)}v^{(t)} - \alpha^{(t)}h\theta^{(t)}] \\ &= \mu^{(t)}\mathbb{E}[v^{(t)}] - (\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}] \quad (1)\end{aligned}$$

We calculate the mean of the parameter $\theta^{(t+1)}$,

$$\mathbb{E}[\theta^{(t+1)}] = \mathbb{E}[\theta^{(t)}] + \mathbb{E}[v^{(t+1)}] \quad (2)$$

Let's assume the following initial conditions:

$$\begin{aligned}\mathbb{E}[v_0] &= 0 \\ \mathbb{E}[\theta_0] &= E_0\end{aligned}$$

Then Eq.(1) and Eq.(2) describes how $\mathbb{E}[\theta^{(t)}]$ changes over time t.

Dynamics of the variance We calculate the variance of the velocity $v^{(t+1)}$,

$$\begin{aligned}\mathbb{V}[v^{(t+1)}] &= \mathbb{V}[\mu^{(t)}v^{(t)} - \alpha^{(t)}h\theta^{(t)}] + (\alpha^{(t)}h\sigma)^2 \\ &= \mathbb{V}[\mu^{(t)}v^{(t)} - \alpha^{(t)}h\theta^{(t)}] + (\alpha^{(t)}h\sigma)^2 \\ &= (\mu^{(t)})^2\mathbb{V}[v^{(t)}] + (\alpha^{(t)}h)^2\mathbb{V}[\theta^{(t)}] - 2\mu^{(t)}\alpha^{(t)}h \cdot \mathrm{Cov}(\theta^{(t)}, v^{(t)}) + (\alpha^{(t)}h\sigma)^2 \quad (3)\end{aligned}$$

The variance of the parameter $\theta^{(t+1)}$ is given by,

$$\mathbb{V}[\theta^{(t+1)}] = \mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}] + 2(\mu^{(t)}\mathrm{Cov}(\theta^{(t)}, v^{(t)}) - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}]) \quad (4)$$

We also need to derive how the covariance of θ and v changes over time:

$$\begin{aligned}\mathrm{Cov}(\theta^{(t+1)}, v^{(t+1)}) &= \mathrm{Cov}((\theta^{(t)} + v^{(t+1)}), v^{(t+1)}) \\ &= \mathrm{Cov}(\theta^{(t)}, v^{(t+1)}) + \mathbb{V}[v^{(t+1)}] \\ &= \mu^{(t)}\mathrm{Cov}(\theta^{(t)}, v^{(t)}) - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}] \quad (5)\end{aligned}$$

Let's assume the following initial conditions:

$$\begin{aligned}\mathbb{V}[v^{(0)}] &= 0 \\ \mathbb{V}[\theta^{(0)}] &= V_0 \\ \mathrm{Cov}[\theta^{(0)}, v^{(0)}] &= 0\end{aligned}$$

Combining Eq.(3-5), we obtain the following dynamics (from $t = 0, \ldots, T-1$):

$$\begin{aligned}\mathbb{V}[v^{(t+1)}] &= (\mu^{(t)})^2\mathbb{V}[v^{(t)}] + (\alpha^{(t)}h)^2\mathbb{V}[\theta^{(t)}] - 2\mu^{(t)}\alpha^{(t)}h \cdot \mathrm{Cov}(\theta^{(t)}, v^{(t)}) + (\alpha^{(t)}h\sigma)^2 \\ \mathbb{V}[\theta^{(t+1)}] &= \mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}] + 2(\mu^{(t)}\mathrm{Cov}(\theta^{(t)}, v^{(t)}) - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}]) \\ \mathrm{Cov}(\theta^{(t+1)}, v^{(t+1)}) &= \mu^{(t)}\mathrm{Cov}(\theta^{(t)}, v^{(t)}) - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}] + \mathbb{V}[v^{(t+1)}]\end{aligned}$$

□

A.2 Proof of the Greedy learning rate and momentum Theorem

Univariate Case The loss at time step t is,

$$\begin{aligned}\mathcal{L}^{(t+1)} &= \frac{1}{2}h(\mathbb{E}[\theta^{(t+1)}]^2 + \mathbb{V}[\theta^{(t+1)}]) \\ &= \frac{1}{2}h\Big[(\mathbb{E}[\theta^{(t)}] + \mu^{(t)}\mathbb{E}[v^{(t)}] - (\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}])^2 + \mathbb{V}[\theta^{(t)}] + (\mu^{(t)})^2\mathbb{V}[v^{(t)}] + (\alpha^{(t)}h)^2\mathbb{V}[\theta^{(t)}] \\ &\quad - 2\mu^{(t)}\alpha^{(t)}h \cdot \mathrm{Cov}(\theta^{(t)}, v^{(t)}) + (\alpha^{(t)}h\sigma)^2 + 2(\mu^{(t)}\mathrm{Cov}(\theta^{(t)}, v^{(t)}) - \alpha^{(t)}h\mathbb{V}[\theta^{(t)}])\Big] \\ &= \frac{1}{2}h\Big[((1-\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}] + \mu^{(t)}\mathbb{E}[v^{(t)}])^2 + (1-\alpha^{(t)}h)^2\mathbb{V}[\theta^{(t)}] + (\mu^{(t)})^2\mathbb{V}[v^{(t)}] \\ &\quad + 2\mu^{(t)}(1-\alpha^{(t)}h)\mathrm{Cov}(\theta^{(t)}, v^{(t)}) + (\alpha^{(t)}h\sigma)^2\Big] \\ &= \frac{1}{2}h\Big[(1-\alpha^{(t)}h)^2\left(\mathbb{E}[\theta^{(t)}]^2 + \mathbb{V}[\theta^{(t)}]\right) + (\mu^{(t)})^2\left(\mathbb{E}[v^{(t)}]^2 + \mathbb{V}[v^{(t)}]\right) \\ &\quad + 2\mu^{(t)}(1-\alpha^{(t)}h)\left(\mathbb{E}[\theta^{(t)}]\mathbb{E}[v^{(t)}] + \mathrm{Cov}(\theta^{(t)}, v^{(t)})\right) + (\alpha^{(t)}h\sigma)^2\Big]\end{aligned}$$

For simplicity, we denote $A(\cdot) = \mathbb{E}[\cdot]^2 + \mathbb{V}[\cdot]$, and notice that $\mathbb{E}[\theta^{(t)}v^{(t)}] = \mathbb{E}[\theta^{(t)}]\mathbb{E}[v^{(t)}] + \mathrm{Cov}(\theta^{(t)}, v^{(t)})$, hence,

$$\mathcal{L}^{(t+1)} = \frac{1}{2}h\Big[(1-\alpha^{(t)}h)^2A(\theta^{(t)})+(\mu^{(t)})^2A(v^{(t)})+2\mu^{(t)}(1-\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}v^{(t)}]+(\alpha^{(t)}h\sigma)^2\Big] \quad (6)$$

In order to find the optimal learning rate and momentum for minimizing $\mathcal{L}^{(t+1)}$, we take the derivative with respect to $\alpha^{(t)}$ and $\mu^{(t)}$, and set it to 0:

$$\begin{aligned}
\nabla_{\alpha^{(t)}}\mathcal{L}^{(t+1)} &= (1-\alpha^{(t)}h)A(\theta^{(t)})\cdot(-h) - \mu^{(t)}h\mathbb{E}[\theta^{(t)}v^{(t)}] + \alpha^{(t)}(h\sigma)^2 = 0\\
\alpha^{(t)}h(A(\theta^{(t)})+\sigma^2) &= A(\theta^{(t)}) + \mu^{(t)}\mathbb{E}[\theta^{(t)}v^{(t)}]\\
\nabla_{\mu^{(t)}}\mathcal{L}^{(t+1)} &= \mu^{(t)}A(v^{(t)}) + (1-\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}v^{(t)}] = 0\\
\mu^{(t)*} &= -\frac{(1-\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}v^{(t)}]}{A(v^{(t)})}\\
\alpha^{(t)}h(A(\theta^{(t)})+\sigma^2) &= A(\theta^{(t)}) - \frac{(1-\alpha^{(t)}h)\mathbb{E}[\theta^{(t)}v^{(t)}]}{A(v^{(t)})}\mathbb{E}[\theta^{(t)}v^{(t)}]\\
\alpha^{(t)}h\left(A(v^{(t)})(A(\theta^{(t)})+\sigma^2) - \mathbb{E}[\theta^{(t)}v^{(t)}]^2\right) &= A(\theta^{(t)})A(v^{(t)}) - \mathbb{E}[\theta^{(t)}v^{(t)}]^2\\
\alpha^{(t)*} &= \frac{A(\theta^{(t)})A(v^{(t)}) - \mathbb{E}[\theta^{(t)}v^{(t)}]^2}{h\left(A(v^{(t)})(A(\theta^{(t)})+\sigma^2) - \mathbb{E}[\theta^{(t)}v^{(t)}]^2\right)}
\end{aligned}$$

High Dimension Case The loss is the sum of losses on each direction:

$$\mathcal{L}^{(t+1)} = \sum_i \frac{1}{2}h_i\Big[(1-\alpha^{(t)}h_i)^2A(\theta_i^{(t)})+(\mu^{(t)})^2A(v_i^{(t)})+2\mu^{(t)}(1-\alpha^{(t)}h_i)\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]+(\alpha^{(t)}h_i\sigma_i)^2\Big]$$

Now we obtain optimal learning rate and momentum by setting the derivative to 0,

$$\begin{aligned}
&\nabla_{\alpha^{(t)}}\mathcal{L}^{(t+1)} = \sum_i h_i\Big[(1-\alpha^{(t)}h_i)A(\theta_i^{(t)})\cdot(-h_i) - \mu^{(t)}h_i\mathbb{E}[\theta_i^{(t)}v_i^{(t)}] + \alpha^{(t)}(h_i\sigma_i)^2\Big] = 0\\
&\alpha^{(t)}\sum_i\Big((h_i)^3(A(\theta_i^{(t)})+(\sigma_i)^2)\Big) = \sum_i\Big((h_i)^2A(\theta_i^{(t)}) + \mu^{(t)}(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\Big)\\
&\nabla_{\mu^{(t)}}\mathcal{L}^{(t+1)} = \sum_i h_i\mu^{(t)}A(v_i^{(t)}) + h_i(1-\alpha^{(t)}h_i)\mathbb{E}[\theta_i^{(t)}v_i^{(t)}] = 0\\
&\mu^{(t)*} = -\frac{\sum_i h_i(1-\alpha^{(t)}h_i)\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]}{\sum_i h_iA(v_i^{(t)})}\\
&\alpha^{(t)}\sum_i\Big((h_i)^3(A(\theta_i^{(t)})+(\sigma_i)^2)\Big)\cdot\Big(\sum_j h_jA(v_j^{(t)})\Big) =\\
&\qquad\left(\sum_i(h_i)^2A(\theta_i^{(t)})\Big(\sum_j h_jA(v_j^{(t)})\Big)\right) - \left(\sum_i\Big(\sum_j h_j(1-\alpha^{(t)}h_j)\mathbb{E}[\theta_j^{(t)}v_j^{(t)}]\Big)(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\right)
\end{aligned}$$

$$\begin{aligned}
&\alpha^{(t)}\sum_i\left(\Big((h_i)^3(A(\theta_i^{(t)})+(\sigma_i)^2)\Big)\Big(\sum_j h_jA(v_j^{(t)})\Big) - \Big(\sum_j(h_j)^2\mathbb{E}[\theta_j^{(t)}v_j^{(t)}]\Big)(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\right) =\\
&\qquad\sum_i\left((h_i)^2A(\theta_i^{(t)})\Big(\sum_j h_jA(v_j^{(t)})\Big) - \Big(\sum_j h_j\mathbb{E}[\theta_j^{(t)}v_j^{(t)}]\Big)(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\right)\\
&\alpha^{(t)*} = \frac{\sum_i\Big((h_i)^2A(\theta_i^{(t)})\Big(\sum_j h_jA(v_j^{(t)})\Big) - \Big(\sum_j h_j\mathbb{E}[\theta_j^{(t)}v_j^{(t)}]\Big)(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\Big)}{\sum_i\Big(\Big((h_i)^3(A(\theta_i^{(t)})+(\sigma_i)^2)\Big)\Big(\sum_j h_jA(v_j^{(t)})\Big) - \Big(\sum_j(h_j)^2\mathbb{E}[\theta_j^{(t)}v_j^{(t)}]\Big)(h_i)^2\mathbb{E}[\theta_i^{(t)}v_i^{(t)}]\Big)}
\end{aligned}$$

A.3 Coordinate-specific Optimality in SGD

Now we consider solving the noisy quadratic problem using coordinate-specific hyperparameters. This reduces to a one dimensional problem as each coordinate evolves independently of others. In the vanilla SGD case, the greedy optimal learning rate is indeed optimal, in the sense that it achieves the minimum possible loss value of *any sequence* of learning rates.

Theorem 2 (Optimal learning rate)**.** *For all $T \in \mathbb{N}$, the sequence of learning rate $\{\alpha^{(t)*}\}_{t=1}^{T-1}$ that minimizes the loss at time step T, $\mathcal{L}(\theta^T)$ is given by,*

$$\alpha^{(t)*} = \frac{A(\theta^{(t)})}{h(A(\theta^{(t)}) + \sigma^2)} \tag{7}$$

Moreover, this agrees with the greedy optimal learning rate.

Proof. Note that when all the coordinates have equal curvature and noise, the optimal learning rate is shared across all dimensions, so that applying the greedy optimal learning rate is optimal. Second-order methods try to precondition the problem so that the curvature is spherical. The noise variance becomes spherical when the Fisher is a good approximation to the Hessian, so that the noise variance matches the curvature. Therefore, one-step lookahead will work well with a sufficiently good natural gradient method. We now consider a dynamic programming approach to solve the problem. We formalize the optimization problem of $\{\alpha_i\}$ as follows. We first denote $f^{(t)}$ as the minimum expected loss at times step T (i.e., under the optimal learning rate) as a function of the mean $\mathbb{E}[\theta^{(t)}]$ and variance $\mathbb{V}[\theta^{(t)}]$ of $\theta^{(t)}$ at step t.

$$f^{(t)}(\mathbb{E}[\theta^{(t)}], \mathbb{V}[\theta^{(t)}]) = \min_{\alpha^{(t)}, \alpha^{(t+1)}, \dots, \alpha^{(T-1)}} \mathbb{E}_{\xi^{(t)}, \xi^{(t+1)}, \dots, \xi^{(T-1)}}[\mathcal{L}(\theta^{(T)})]$$

Because we can express the $\mathbb{E}[\theta^{(t)}]$ and $\mathbb{V}[\theta^{(t)}]$ in terms of the $\mathbb{E}[\theta^{(t-1)}]$ and $\mathbb{V}[\theta^{(t-1)}]$ of $\theta^{(t-1)}$ and the learning rate $\alpha^{(t-1)}$ at the previous time step,

$$\theta^{(t)} = (1 - \alpha^{(t-1)}h)\theta^{(t-1)} + \alpha^{(t-1)}h\sigma\xi^{(t-1)}$$

$$\Rightarrow (\mathbb{E}[\theta^{(t)}], (\mathbb{V}[\theta^{(t)}])^2) = ((1 - \alpha^{(t-1)}h)\mathbb{E}[\theta^{(t-1)}], ((1 - \alpha^{(t-1)}h)\mathbb{V}[\theta^{(t-1)}]^2 + (\alpha^{(t-1)}h\sigma)^2),$$

we can solve the minimization problem recursively by,

$$f^{(t)}(\mathbb{E}[\theta^{(t-1)}], \mathbb{V}[\theta^{(t-1)}] = \min_{\alpha^{(t-1)}} f^{(t)}(\mathbb{E}[\theta^{(t-1)}], \mathbb{V}[\theta^{(t-1)}], \alpha^{(t-1)}).$$

Now, let's first derive the form of f for $T, T-1, T-2$ for illustration. Let $A^{(t)} = (\mathbb{E}[\theta^{(t)}])^2 + (\mathbb{V}[\theta^{(t)}])^2$. We can find a recurrence relation in terms of A:

$$A^{(t)} = (1-\alpha^{(t-1)}h)^2((\mathbb{E}[\theta^{(t-1)}])^2 + \mathbb{V}[\theta^{(t-1)}]^2) + (\alpha^{(t-1)}h\sigma)^2 = (1-\alpha^{(t-1)}h)^2 A^{(t-1)} + (\alpha^{(t-1)}h\sigma)^2$$

Since f depends on $\mathbb{E}[\theta]$ and σ only through A, so we instead write f as a function of A:

$$f^{(t)}(A^{(T)}) = \frac{1}{2}hA^{(T)} + \sigma^2$$

Now since $A^{(T)}$ is a function of $\alpha^{(T-1)}$, if we take the derivative of $f^{(t)}$ w.r.t. $\alpha^{(T-1)}$ and setting it to zero, we get:

$$\begin{aligned} \frac{df^{(t)}}{d\alpha^{(T-1)}} &= \frac{1}{2}h\frac{dA^{(T-1)}}{d\alpha^{(T-2)}} = 0 \\ \Rightarrow \quad \frac{dA^{(T)}}{d\alpha^{(T-1)}} &= 0 \\ \Rightarrow \quad \alpha^{(T-1)*} &= \frac{A^{(T-1)}}{h(A^{(T-1)} + \sigma^2)} \end{aligned}$$

Thus we can write $A^{(T)}$ in terms of $A^{(T-1)}$ and the optimal $\alpha^{(T-1)*}$:

$$\begin{aligned} A^{(T)}(A^{(T-1)}, \alpha^{(T-1)*}) &= (1 - \frac{A^{(T-1)}}{A^{(T-1)} + \sigma^2})^2 A^{(T-1)} + (\frac{A^{(T-1)}}{A^{(T-1)} + \sigma^2}\sigma)^2 \\ &= (\frac{\sigma^2}{A^{(T-1)} + \sigma^2})^2 A^{(T-1)} + (\frac{A^{(T-1)}}{A^{(T-1)} + \sigma^2}\sigma)^2 \\ &= \frac{A^{(T-1)}\sigma^2}{A^{(T-1)} + \sigma^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} f^{(T-1)}(A^{(T-1)}) &= \frac{1}{2}hA^{(T)} + \sigma^2 \\ &= \frac{1}{2}h(\frac{A^{(T-1)}\sigma^2}{A^{(T-1)} + \sigma^2}) + \sigma^2 \end{aligned}$$

Now since $A^{(T-1)}$ is a function of $\alpha^{(T-2)}$, if we take the derivative of $f^{(T-1)}$ w.r.t. $\alpha^{(T-2)}$ and setting it to zero, we get:

$$\begin{aligned} \frac{df^{(T-1)}}{d\alpha^{(T-2)}} &= \frac{1}{2}\frac{h\sigma^4 \frac{dA^{(T-1)}}{d\alpha^{(T-2)}}}{(A^{(T-1)} + \sigma^2)^2} = 0 \\ \Rightarrow \quad & \frac{dA^{(T-1)}}{d\alpha^{(T-2)}} = 0 \\ \Rightarrow \quad & \alpha^{(T-2)*} = \frac{A^{(T-2)}}{h(A^{(T-2)} + \sigma^2)} \\ \Rightarrow \quad & f^{(T-2)}(A^{(T-2)}) = \frac{1}{2}h(\frac{A^{(T-2)}\sigma^2}{2A^{(T-2)} + \sigma^2}) + \sigma^2 \end{aligned}$$

We then observe the form of f_{T-k} can be easily derived by induction on k, and use the identity that,

$$\frac{(\frac{ab}{a+b})b}{k(\frac{ab}{a+b}) + b} = \frac{ab}{(k+1)a + b}.$$

The learning rate then follows immediately by taking the derivative of f_{T-k} w.r.t. $\alpha^{(T-k-1)}$ and setting it to zero. Hence we conclude, for all $T \in \mathbb{N}$, and $k \in \mathbb{N}$, $1 \leq k \leq T$, we have,

$$f^{(t)}(A^{(T-k)}) = \frac{1}{2}h(\frac{A^{(T-k)}\sigma^2}{kA^{(T-k)} + \sigma^2}) + \sigma^2. \tag{8}$$

Therefore, the optimal learning $\alpha^{(t)}$ at timestep t is given as,

$$\alpha^{(t)*} = \frac{A^{(t)}}{h(A^{(t)} + \sigma^2)}, \tag{9}$$

which agrees with the greedy optimal learning rate. □