B.4 APPROXIMATION OF $\tilde{\mathcal{L}}(\hat{W}, t)$ IN EQ. 22.

Lemma B.1. *Let $\mathcal{L}_T(\hat{W})$ be the training loss function on which we apply gradient descent. The t^{th} iterate of gradient descent matches the minimum of $\tilde{\mathcal{L}}_T(\hat{W}, t)$ defined as,*

$$\tilde{\mathcal{L}}_T(\hat{W}, t) := \frac{1}{2n} \sum \Big[\hat{y}^\mu - y^\mu\Big]^2 + \frac{1}{\eta t}||\hat{W}||_2^2. \tag{92}$$

Proof. The goal is to show,

$$\hat{W}_t = \arg\min_{\hat{W}} \tilde{\mathcal{L}}_T(\hat{W}, t), \qquad \text{where,} \qquad \hat{W}_t := \hat{W}_{t-1} - \eta \nabla_{\hat{W}_{t-1}} \mathcal{L}(\hat{W}_{t-1}). \tag{93}$$

For brevity of derivations, here we only consider the case where $\lambda = \sigma_\epsilon^2 = 0$. Recall the closed-form derivation of $\hat{W}_t$ in Eq. 18,

$$\hat{W}_t = \Big(I - \big[I - \eta X^T X\big]^t\Big)(X^T X)^{-1} X^T y, \tag{94}$$

$$= \arg\min_{\hat{W}} \Big[X\hat{W} - X\Big(I - \big[I - \eta X^T X\big]^t\Big)(X^T X)^{-1} X^T y\Big]^2, \tag{95}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T}\Big(I - \big[I - \eta X^T X\big]^t\Big) \underbrace{(X^T X)^{-1} X^T y}_{=W, \text{ assuming } \sigma_\epsilon^2 = 0}\Big]^2, \tag{96}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - \underbrace{x^{\mu^T}\Big(I - \big[I - \eta X^T X\big]^t\Big) W}_{\text{a dynamic target (function of } t)}\Big]^2, \tag{97}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \big[I - \eta\Lambda\big]^t\Big) V^T W\Big]^2, \qquad (X^T X = V\Lambda V^T) \tag{98}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(t \log[I - \eta\Lambda]\big)\Big) V^T W\Big]^2, \tag{99}$$

$$\approx \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\eta\Lambda t\big)\Big) V^T W\Big]^2, \qquad (\log(1+x) \approx x) \tag{100}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\frac{\Lambda}{1/\eta t}\big)\Big) V^T W\Big]^2, \tag{101}$$

$$\approx \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\log(\frac{\Lambda}{1/\eta t} + I)\big)\Big) V^T W\Big]^2, \quad (\log(1+x) \approx x) \tag{102}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \big[\Lambda + \frac{1}{\eta t} I\big]^{-1} \frac{1}{\eta t}\Big) V^T W\Big]^2, \tag{103}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big((\Lambda + \frac{1}{\eta t} I)^{-1} \Lambda\Big) V^T W\Big]^2, \tag{104}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} (X^T X + \frac{1}{\eta t} I)^{-1} X^T X W\Big]^2, \tag{105}$$

$$= \arg\min_{\hat{W}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} \underbrace{(X^T X + \frac{1}{\eta t} I)^{-1} X^T y}_{\text{the normal equation}}\Big]^2, \tag{106}$$

$$= \arg\min_{\hat{W}} \underbrace{\frac{1}{2n} \sum \Big[\hat{y}^\mu - y^\mu\Big]^2 + \frac{1}{\eta t}||\hat{W}||_2^2}_{\tilde{\mathcal{L}}_T(\hat{W}, t)}, \tag{107}$$

which concludes the proof. Consistent with (Ali et al., 2019), this approximation implies that "L2 regularization and early stopping can be seen as equivalent (at least under the quadratic approximation of the objective function)." (Goodfellow et al., 2016).