# A Appendix / supplemental material

| Notations | |
|---|---|
| Symbol | Meaning |
| $m$ | Number of neurons in hidden-layer of score network |
| $C_{w_y,u_y}$ | Upper bound on $\|w_{y,i}\|_1, \|u_{y,i}\|_1$ |
| $F_T^2$ | Upper bound on $\mathbb{E}_{X_0}\mathbb{E}_{\xi_j}\left[\|\sigma(Wx(t)+Ue(t))\|_2^2\right], 0 \le t \le T$ |
| $L_y$ | $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ is $\frac{L_y}{m^2}$ smooth w.r.t. $\theta_y$ |
| $\phi_y$ | Lipschitz constant of $\bar{\mathcal{L}}_{mut}^{n_y}(\theta_y, \theta_{-y})$ w.r.t. $\theta_y$ |
| $\sigma_y$ | $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ is $\sigma_y$ Lipschitz in $\theta_y$ |
| $B$ | Upper Bound of the Frobenius norm of $A_y$ |

# B Sampling

Denote the backward time schedule as $\{t_j^{\leftarrow}\}_{0\le j\le N}$ such that $0 = t_0^{\leftarrow} < t_1^{\leftarrow} < \cdots, t_N^{\leftarrow} = T - \alpha$. Lower case $p_t$ represents the density of $P_t$. We consider the exponential integrator scheme for simulating the backward SDE with

The generation algorithm can be expressed as a piecewise continuous-time SDE: for any $t \in [t_j^{\leftarrow}, t_{j+1}^{\leftarrow})$.

$$d\bar{Y}_t = (\bar{Y}_t + 2s_{T-t_j^{\leftarrow},\theta_y}(\bar{Y}_{t_j^{\leftarrow}}))dt + \sqrt{2}d\bar{W}_t \tag{23}$$

Denote $q_t :=$ Law $(\bar{Y}_t), \forall t \in [0, T-\delta]$.

**Theorem 2.** *[3, Theorem 1] Let Assumption 1 hold. Then there exists a numerical constant $C_0 > 0$, such that*

$$D_{KL}(p_\alpha(.|y)||q_{T-\alpha}(.|y)) \le C_0(E_S + E_D + E_F) \tag{24}$$

*where $E_D \le \kappa^2 N u_2^2 + \kappa T u_2^2$ is the discretization error due to the reverse SDE, $E_F \le \exp(-2T)u_2^2$ is the error due to the convergence of the forward SDE and $E_S$ is the score estimation error*

$$E_S(\theta_y) = \sum_{j=0}^{N-1} \gamma_j \mathbb{E}_{x\sim p_{T-t_j^{\leftarrow}}}\left[\left\|\nabla \log p_{T-t_j^{\leftarrow}}(x|y) - s_{T-t_j^{\leftarrow},\theta_y}(x)\right\|_2^2\right] \tag{25}$$

*where $\gamma_j := t_{j+1}^{\leftarrow} - t_j^{\leftarrow}, \forall j = 0, 1, \cdots N-1$ is the step-size of the generation algorithm.*

When the training is done over the forward discretization given by $(t_{N-j} = T - t_j^{\leftarrow})_{j=0}^{N-1}$, we have

$$\begin{aligned} E_S =& \sum_{j=0}^{N-1} \frac{\bar{\sigma}_{t_{N-j}}\lambda(t_{N-j})}{\lambda(t_{N-j})\bar{\sigma}_{t_{N-j}}}(t_{N-j} - t_{N-j-1})\mathbb{E}_{X_0}\mathbb{E}_{X_{t_{N-j}}|X_0}\left\|\bar{\sigma}_{t_{N-j}} s_{t_{N-j},\theta_y}(X_{t_{N-j}}) + \xi\right\|^2 \\ &+ \sum_{j=0}^{N-1} \frac{\bar{\sigma}_{t_{N-j}}}{\lambda(t_{N-j})}\lambda(t_{N-j})(t_{N-j} - t_{N-j-1})C_{t_{N-j}} \\ \le& 2\max_j \frac{\bar{\sigma}_{t_{N-j}}}{\lambda(t_{N-j})}\mathcal{L}^y(\theta_y) \end{aligned}$$

where

$$\begin{aligned} \mathcal{L}^y(\theta_y) =& \frac{1}{2}\sum_{j=0}^{N-1} \mathbb{E}_{X_0}\mathbb{E}_{X_{t_{N-j}}|X_0}\Big[\lambda(t_{N-j})(t_{N-j} - t_{N-j-1}) \\ &\left\|\nabla_{x(t_{N-j})} \log p_{t_{N-j}}(x(t_{N-j})|x_0) - s_{t_{N-j},\theta_y}(x(t_{N-j}))\right\|_2^2\Big] \\ &+ \frac{1}{2}\sum_{j=0}^{N-1} \lambda(t_{N-j})(t_{N-j} - t_{N-j-1})C_{t_{N-j}}(y) \end{aligned} \tag{26}$$

**Theorem 3.** *(Appendix B and [3, Theorem 1]) Let Assumption 1 hold. Then there exists a numerical constant $C_0 > 0$, such that*

$$D_{KL}(p_\alpha(.|y)||q_{T-\alpha}(.|y)) \leq C_0(\mathcal{L}^y(\theta_y) + E_D + E_F) \tag{27}$$

*where $E_D \leq \kappa^2 N u_2^2 + \kappa T u_2^2$ is the discretization error due to the reverse SDE, $E_F \leq \exp(-2T)u_2^2$ is the error due to the convergence of the forward SDE.*

### B.1 Decomposition of $\mathcal{L}^y(\theta_y)$

Let $\theta_y^* = \text{argmin}_{\theta_y} \mathcal{L}^y(\theta_y)$. We further decompose $\mathcal{L}^y(\theta_y)$ as

$$\begin{aligned}
\max_{y\in Y}\Big(\mathcal{L}^y(\theta_y) - \mathcal{L}^y(\theta_y^*)\Big) \leq & \max_{y\in Y}\Big(\mathcal{L}^y(\theta_y) + \beta\mathcal{L}^y_{mut}(\theta_y,\theta_{-y}) - (\mathcal{L}^y(\theta_y^*) + \beta\mathcal{L}^y_{mut}(\theta_y^*,\theta_{-y})))\\
& + \beta(\mathcal{L}^y_{mut}(\theta_y^*,\theta_{-y}) - \mathcal{L}^y_{mut}(\theta_y,\theta_{-y}))\Big)\\
\overset{(a)}{\leq} & \max_{y\in Y}\Big(\mathcal{L}^y(\theta_y) + \beta\mathcal{L}^y_{mut}(\theta_y,\theta_{-y}) - (\mathcal{L}^y(B(\theta_{-y}))\\
& + \beta\mathcal{L}^y_{mut}(B(\theta_{-y}),\theta_{-y}))\Big) + \beta\max_{y\in Y}\mathcal{L}^y_{mut}(B(\theta_{-y}),\theta_{-y})\\
\leq & \max_{y\in Y}\Big(\mathcal{L}^y_{reg}(\theta_y,\theta_{-y}) - \mathcal{L}^y_{reg}(B(\theta_{-y}),\theta_{-y})\Big)\\
& + \beta\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})}\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})
\end{aligned}$$

where $(a)$ follows from the fact that $\mathcal{L}^y_{reg}(B(\theta_{-y}),\theta_{-y}) \leq \mathcal{L}^y_{reg}(\theta_y,\theta_{-y})$. We further decompose this to obtain an upper bound on $\max_{y\in Y}\min_t \mathcal{L}^y(\theta_y^t)$

$$\begin{aligned}
\max_{y\in Y}\mathcal{L}^y(\theta_y) \leq & \max_{y\in Y}\mathcal{L}^y(\theta_y^*) + \max_{y\in Y}\Big(\mathcal{L}^y_{reg}(\theta_y,\theta_{-y}) - \mathcal{L}^y_{reg}(B(\theta_{-y}),\theta_{-y})\Big)\\
& + \beta\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})}\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})\\
\overset{(a)}{\leq} & \max_{y\in Y}\mathcal{L}^y(\theta_y^*) + \max_{y\in Y} \mid \mathcal{L}^y_{reg}(\theta_y,\theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y}) \mid\\
& + \max_{y\in Y} \mid \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(B(\theta_{-y}),\theta_{-y}) \mid\\
& + \max_{y\in Y} \mid \mathcal{L}^y_{reg}(B(\theta_{-y}),\theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(B(\theta_{-y}),\theta_{-y}) \mid\\
& + \beta\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})}\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})\\
\overset{(b)}{\leq} & \max_{y\in Y}\mathcal{L}^y(\theta_y^*) + 2\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})} \mid \mathcal{L}^y_{reg}(\theta_y,\theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y}) \mid\\
& + \text{NE-gap}(\theta_y,\theta_{-y}) + \beta\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})}\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})\\
\implies \min_{\tau\in[T_{train}]}\max_{y\in\mathcal{Y}}\mathcal{L}^y(\theta_y^\tau) \overset{(c)}{\leq} & \max_{y\in Y}\mathcal{L}^y(\theta_y^*) +\\
& + 2\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})} \mid \mathcal{L}^y_{reg}(\theta_y,\theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y}) \mid\\
& + \min_{\tau\in[T_{train}]}\text{NE-gap}(\theta_y^\tau,\theta_{-y}^\tau) + \beta\max_{y\in Y}\sup_{(\theta_y,\theta_{-y})}\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})
\end{aligned}$$

where $(a)$ follows from adding and subtracting the empirical losses $\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})$ and $\bar{\mathcal{L}}^{n_y}_{reg}(B_y(\theta_{-y}),\theta_{-y})$ and using triangle inequality of the max norm, $(b)$ follows from the gradi-

ent domination property for strongly convex functions, $(c)$ follows from taking the minimum over the iterates of the algorithm.

### B.2 Boundedness of Forward Dynamics

**Lemma 1.** *Consider the forward diffusion process with linear drift coefficients. For any $\delta > 0, \delta \ll 1$, w.p. (with probability) of atleast $1-\delta$. we have*

$$\|x(t)\|_\infty \leq C_T\left(\|x(0)\|_\infty + \sqrt{\log\frac{2}{\pi\delta^2}}\right) \tag{28}$$

*where $C_T := \max_{t\in[0,T]} r(t), r(t)v(t)$.*

*Proof:* The proof is similar to [15, Lemma 1] When the drift coefficient $f(.,t) : \mathbb{R}^d \to \mathbb{R}^d$ is linear in $x$ i.e. $f(x,t) = -f(t)x$, the transition kernel $p_{t|0}$ has a closed form

$$p_{t|0}(x(t)|x(0)) = \mathcal{N}(x(t); \mu(t)x(0), \bar{\sigma}^2(t)I_d) \tag{29}$$

where $\mu(t) := \exp\left(\int_0^t f(\xi)d\xi\right), \bar{\sigma}^2(t) := 2\int_0^t \exp\left(2\mu_s - 2\mu_t\right)\sigma_s^2 ds$. Together we get,

$$x(t) = \mu(t)x(0) + \bar{\sigma}(t)z, z \sim \mathcal{N}(0, I_d) \tag{30}$$

For any $\epsilon \sim \mathcal{N}(0,1), c > 1$, we have

$$\mathbb{P}\{\epsilon : |\epsilon| > c\} = 2\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \leq \frac{1}{\sqrt{2\pi}}\int_c^\infty 2xe^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}}\int_{c^2}^\infty e^{-\frac{x}{2}} dx = \sqrt{\frac{2}{\pi}} e^{-\frac{c^2}{2}} \tag{31}$$

Let $\delta = \sqrt{\frac{2}{\pi}} e^{-\frac{c^2}{2}}$, then

$$\mathbb{P}\{\epsilon : |\epsilon| \leq \sqrt{\log\frac{2}{\pi\delta^2}}\} \geq 1 - \delta \tag{32}$$

Hence, for any $\delta \in (0,1)$ with $\delta \ll 1$, w.p. at least $1-\delta$, we have

$$\|x(t)\|_\infty \leq C_T\left(\|x(0)\|_\infty + \sqrt{\log\frac{2}{\pi\delta^2}}\right) \tag{33}$$

where $C_T := \max_{t\in[0,T]}\{\mu(t), \bar{\sigma}(t)\}$. Let $C_{T,\delta} = C_T(K + \sqrt{\log\frac{2}{\pi\delta^2}})$

### B.3 Boundedness of Loss function $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$

In this section, study some properties of the game defined by $\langle \mathcal{Y}, (\bar{\mathcal{L}}_{reg}^{n_y})_{y\in\mathcal{Y}}, (\Theta_y)_{y\in\mathcal{Y}}\rangle$. From Eq. 10, we have

$$\mathcal{L}_{conti,reg}^y(\theta_y, \theta_{-y}) = \mathcal{L}_{conti}^y(\theta_y) + \beta\mathcal{L}_{conti,mut}^y(\theta_y, \theta_{-y}) \tag{34}$$

where

$$\mathcal{L}_{conti,mut}^y(\theta_y, \theta_{-y}, \omega(.)) = \frac{1}{2}\int_{t_0}^T \omega(t)\mathbb{E}_{x(t)\sim p_t}\mathbb{E}_{y'\sim Q}\left[\left\|s_{t,\theta_y}(x(t)) - s_{t,\theta_{y'}}(x(t))\right\|_2^2\right]dt$$

and

$$\mathcal{L}_{conti}^y(\theta_y, \theta_{-y}) = \frac{1}{2}\int_{t_0}^T \lambda(t)\mathbb{E}_{(x(t),y)}\left[\left\|\nabla_{x(t)}\log p_t(x(t)|y) - s_{t,\theta}(x(t),y)\right\|_2^2\right]dt$$

Conditioning on $X_0$ and using law of iterated expectation, we can write [29, Appendix A], we get

$$\begin{aligned}\mathcal{L}_{conti,reg}^y(\theta_y, \theta_{-y}) =& \frac{1}{2}\int_{t_0}^T \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0,y}\Big[\lambda(t)\left\|s_{t,\theta_y}(x(t)) - \nabla_{x(t)}\log p_t(x(t)|x_0)\right\|_2^2 \\ &+ \beta\omega(t)\mathbb{E}_{y'\sim Q}\left[\left\|s_{t,\theta_y}(x(t)) - s_{t,\theta_{y'}}(x(t))\right\|_2^2\right]\Big]dt + \frac{1}{2}\int_{t_0}^T\left[\lambda(t)C_t(y)\right]dt\end{aligned}$$

where $C_t(y) = \mathbb{E}_{X_t} \|\nabla \log p_t(X_t|y)\|^2 - \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0} \|\nabla \log p_t(X_t|X_0, y)\|^2$ to learn the score $\nabla_{x(t)} \log p_t(x(t)|x_0, y)$.

Furthermore, we discretize the time points $0 = t_0 < t_1 < \cdots < t_N = T$ to the objective function

$$\begin{aligned}
\mathcal{L}^y_{reg}(\theta_y, \theta_{-y}) &= \mathcal{L}^y(\theta_y) + \beta \mathcal{L}^y_{mut}(\theta_y, \theta_{-y}) \\
&= \frac{1}{2}\sum_{j=1}^{N} \lambda(t_j)(t_j - t_{j-1}) \mathbb{E}_{X_0}\mathbb{E}_{X_{t_j}|X_0}\left[ \left\| \nabla_{x(t_j)} \log p_t(x_i(t_j)|x_0) - s_{t_j,\theta_y}(x_i(t_j)) \right\|_2^2 \right] + \\
&+ \bar{C}(y) + \beta \frac{1}{2}\sum_{j=1}^{N} \omega(t_j)(t_j - t_{j-1}) \mathbb{E}_{X_0}\mathbb{E}_{X_{t_j}|X_0}\mathbb{E}_{y' \sim Q}\left[ \left\| s_{t_j,\theta_y}(x_i(t_j)) - s_{t,\theta_{y'}}(x_i(t_j)) \right\|_2^2 \right]
\end{aligned} \tag{35}$$

where $\bar{C}(y) = \frac{1}{2}\sum_{j=1}^{N} \lambda(t_j)(t_j - t_{j-1}) C_{t_j}(y)$ From [29, Appendix A], we have $X_t|X_0 \sim \mathcal{N}(e^{-\mu_t} X_0, \bar{\sigma}_t^2 I)$ and its density function is

$$p_t(x|x_0) = (2\pi\bar{\sigma}_t^2)^{-\frac{d}{2}} \exp(-\frac{\|x - e^{-\mu_t} x_0\|}{2\bar{\sigma}_t^2})$$

Then,

$$\begin{aligned}
\Delta =& \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0} \left\| s_{t_j,\theta_y}(x_i(t_j)) - \nabla_{x(t_j)} \log p_t(x(t_j)|x_0) \right\| \\
=& \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0} \left\| s_{t_j,\theta_y}(x_i(t_j)) - \nabla_x \left( - \frac{\|X_t - e^{-\mu_t} X_0\|}{2\bar{\sigma}_t^2} \right) \right\|^2 \\
=& \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0} \left\| s_{t_j,\theta_y}(x_i(t_j)) + \frac{X_t - e^{\mu_t} X_0}{\bar{\sigma}_t^2} \right\|^2 \\
=& \mathbb{E}_{X_0}\mathbb{E}_{\epsilon_t} \left\| s_{t_j,\theta_y}(x_i(t_j)) + \frac{\epsilon_t}{\bar{\sigma}_t^2} \right\|^2
\end{aligned}$$

Let $\xi = \frac{\epsilon_t}{\bar{\sigma}_t} \sim \mathcal{N}(0, I)$

$$\Delta = \frac{1}{\bar{\sigma}_t} \mathbb{E}_{X_0}\mathbb{E}_{\xi} \left\| \bar{\sigma}_t s_{t_j,\theta_y}(x_i(t_j)) + \xi \right\|^2 \tag{36}$$

Finally putting all of it together, we get the empirical loss function

$$\begin{aligned}
\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y}) =& \frac{1}{2n_y} \sum_{i=1}^{n_y} \sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} \left[ \left\| \bar{\sigma}_{t_j} s_{t_j,\theta_y}(x_i(t_j)) + \xi_{ij} \right\|_2^2 \right. \\
&\left. + \beta\omega(t_j)(t_j - t_{j-1}) \mathbb{E}_{y' \sim Q}\left[ \left\| s_{t_j,\theta_y}(x_i(t_j)) - s_{t,\theta_{y'}}(x_i(t_j)) \right\|_2^2 \right] \right]
\end{aligned} \tag{37}$$

We will show that the empirical loss function for the label $y \in \mathcal{Y}$, $\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y})$ that is optimized is convex and smooth in $\theta_y$ with high probability.

**Lemma 2.** *For $\delta > 0, \delta \ll 1, wp.1 - n_y N\delta$, the empirical loss function*

$$\begin{aligned}
\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y}) =& \frac{1}{2n_y} \sum_{i=1}^{n_y} \sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} \left[ \left\| \bar{\sigma}_{t_j} s_{t_j,\theta_y}(x_i(t_j)) + \xi_{ij} \right\|_2^2 \right. \\
&\left. + \beta\omega(t_j)(t_j - t_{j-1}) \mathbb{E}_{y' \sim Q}\left[ \left\| s_{t_j,\theta_y}(x_i(t_j)) - s_{t,\theta_{y'}}(x_i(t_j)) \right\|_2^2 \right] \right]
\end{aligned} \tag{38}$$

*is bounded i.e.*

$$\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y}) = \mathcal{O}(\sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} + \beta\omega(t_j)(t_j - t_{j-1}))$$

**Proof:** From Lemma 1, we have $\delta > 0, \delta \ll 1$

$$\mathbb{P}\{|\xi_{ij}| > \sqrt{\frac{2}{\pi\delta^2}}\} \leq \delta \tag{39}$$

Thus, $w.p.\ 1-n_yN\delta$, we have $|\xi_{ij}| \leq \sqrt{\frac{2}{\pi\delta^2}}$ and hence we have $\|x(t_j)\|_\infty \leq C_{t_N,\delta}, \forall i = 1, \cdots, n_y$ and $j = 1, , \cdots, N$

Thus, $w.p.\ 1 - n_yN\delta$

$$\begin{aligned}
\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y}) =& \frac{1}{2n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}}\Big[\left\|\bar{\sigma}_{t_j}s_{t_j,\theta_y}(x_i(t_j)) + \xi_{ij}\right\|_2^2 \\
&+ \beta\omega(t_j)(t_j - t_{j-1})\mathbb{E}_{y'\sim Q}\left[\left\|s_{t_j,\theta_y}(x_i(t_j)) - s_{t,\theta_{y'}}(x_i(t_j))\right\|_2^2\right]\Big] \\
&\leq \frac{1}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}}(\bar{\sigma}_{t_j}^2\left\|s_{t_j,\theta_y}(x_i(t_j))\right\|_2^2 + \|\xi_{ij}\|^2) \\
&+ \beta\omega(t_j)(t_j - t_{j-1})(\left\|s_{t_j,\theta_y}(x_i(t_j))\right\|_2^2 + \max_{y'\in\mathcal{Y}}\left\|s_{t_j,\theta_{y'}}(x_i(t_j))\right\|_2^2)
\end{aligned}$$

For a bound on $\left\|s_{t_j,\theta_y}(x(t_j))\right\|_2$

$$\begin{aligned}
\left\|s_{t_j,\theta_y}(x(t_j))\right\|_2 &= \left\|\frac{1}{m}\sum_{i=1}^{m}a_{y,i}\sigma(w_{y,i}^Tx(t_j) + u_{y,i}^Te(t_j))\right\|_2 && (40)\\
&\overset{(a)}{\leq} \frac{1}{m}\sum_{i=1}^{m}\|a_{y,i}\|_2\,|\sigma(w_{y,i}^Tx(t_j) + u_{y,i}^Te(t))| && (41)\\
&\overset{(b)}{\leq} \frac{1}{m}\sum_{i=1}^{m}\|a_{y,i}\|_2\,(\|w_{y,i}\|_1\,\|x(t_j)\|_\infty + \|u_{y,i}\|_1\,\|e(t))\|_\infty) && (42)\\
&\leq \frac{1}{m}\sum_{i=1}^{m}\|a_{y,i}\|_2\,(C_{t_N,\delta}\,\|w_{y,i}\|_1 + \max_j\|e(t_j)\|_\infty\,\|u_{y,i}\|_1) && (43)\\
&\overset{(c)}{\leq} (C_{t_n,\delta} + C_{t_N,e})C_{w_y,u_y}B && (44)
\end{aligned}$$

where $(a)$ follows from triangle inequality for norms, $(b)$ follows from the fact that the ReLu function satisfies $|\sigma(x)| \leq |x|$ and Holder inequality and $(c)$ follows from the bounds on the embeddings and $x(t_j)$ with $\|w_{y,i}\|_1, \|u_{y,i}\|_1 \leq C_{w_y,u_y}, \forall i \in [m]$. Thus, for $\delta > 0, \delta \ll 1$, we have $w.p.\ 1 - n_yN\delta$

$$\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y}) \leq C_1\sum_{j=1}^{N}\frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} + \beta\omega(t_j)(t_j - t_{j-1}) \tag{45}$$

where $C_1 = (\bar{\sigma}_{t_N}^2 + 2)(C_{t_n,\delta} + C_{t_N,e})^2C_{w_y,u_y}^2B^2 + \frac{2}{\pi\delta^2}$. Since $\bar{\sigma}_{t_j}$ is non-decreasing in $j$, so $\max_j\bar{\sigma}_{t_j} = \bar{\sigma}_{t_N}$.

### B.4 Boundedness of Gradient of Loss function $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$

$$
\begin{aligned}
&\left\|\nabla_{A_y}\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})\right\|_F^2 = \sum_{k=1}^{d}\left\|\nabla_{(A_y)_k}\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})\right\|^2 \\
&= \sum_{k=1}^{d}\Bigg|\Bigg|\frac{1}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\lambda(t_j)(t_j-t_{j-1})(\bar{\sigma}_{t_j}s_{t_j,\theta_y}(x(t_j))+\xi_{ij})_k\sigma(W_yx(t_j)+U_ye(t_j)) \\
&\quad +\beta\omega(t_j)(t_j-t_{j-1})\mathbb{E}_{y'}[(s_{t_j,\theta_y}(x(t_j))-s_{t_j,\theta_{y'}}(x(t_j)))_k\sigma(W_yx(t_j)+U_ye(t_j))]\Bigg|\Bigg|^2 \\
&\le 2\sum_{k=1}^{d}\left\|\frac{1}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\lambda(t_j)(t_j-t_{j-1})(\bar{\sigma}_{t_j}s_{t_j,\theta_y}(x(t_j))+\xi_{ij})_k\sigma(W_yx(t_j)+U_ye(t_j))\right\|^2 \\
&\qquad +2\beta^2\sum_{k=1}^{d}\Bigg\|\frac{1}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\omega(t_j)(t_j-t_{j-1})\mathbb{E}_{y'}[(s_{t_j,\theta_y}(x(t_j))-s_{t_j,\theta_{y'}}(x(t_j)))_k\cdot \\
&\qquad\qquad \sigma(W_yx(t_j)+U_ye(t_j))]\Bigg\|^2 \\
&\le 2\frac{N}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\lambda(t_j)^2(t_j-t_{j-1})^2\sum_{k=1}^{d}\left\|\bar{\sigma}s_{t_j,\theta_y}(x(t_j))+\xi_{ij}\right\|^2\left\|\sigma(W_yx(t_j)+U_ye(t_j))\right\|^2 \\
&\qquad\qquad +2\beta^2\frac{N}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\omega(t_j)^2(t_j-t_{j-1})^2\cdot \\
&\qquad\qquad \sum_{k=1}^{d}\mathbb{E}_{y'}[\left\|s_{t_j,\theta_y}(x(t_j))-s_{t_j,\theta_{y'}}(x(t_j))\right\|^2]\left\|\sigma(W_yx(t_j)+U_ye(t_j))\right\|^2 \\
&\le 4Nd\left\|\sigma(W_yx(t_j)+U_ye(t_j))\right\|_2^2\max_j\{\lambda(t_j)(t_j-t_{j-1})\bar{\sigma}_{t_j},\beta\omega(t_j)(t_j-t_{j-1})\}\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y}) \\
&\le 4Nd^2(C_{t_N,\delta}+C_{t_N,e})^2C_{w_y,u_y}^2\max_j\{\lambda(t_j)(t_j-t_{j-1})\bar{\sigma}_{t_j},\beta\omega(t_j)(t_j-t_{j-1})\}\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})
\end{aligned}
$$

Since $w.p.1-n_yN\delta$ the empirical loss function $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})$ is bounded, $\left\|\nabla_{A_y}\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})\right\|_F^2$ is bounded with the same probability.

This also shows that for fixed $\theta_{-y}$, $(W_y,U_y)_{y\in\mathcal{Y}}$, $w.p.1-n_yN\delta$, $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})$ is a Lipschitz function in $\theta_y$ with Lipschitz constant $\sigma_y$ such that $\sigma_y^2 = 4C_1Nd^2(C_{t_N,\delta}+C_{t_N,e})^2C_{w_y,u_y}^2\max_j\{\lambda(t_j)(t_j-t_{j-1})\bar{\sigma}_{t_j},\beta\omega(t_j)(t_j-t_{j-1})\}\left(\sum_{j=1}^{N}\frac{\lambda(t_j)(t_j-t_{j-1})}{\bar{\sigma}_{t_j}}+\beta\omega(t_j)(t_j-t_{j-1})\right)$

### B.5 Smoothness of Loss Function $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})$

**Lemma 3.** *Let* $(W_y,U_y)_{y\in\mathcal{Y}},\theta_{-y},\{t_j\}_{j=1}^N$ *be fixed. Let* $L_y = d(C_{t_N,\delta}+C_{t_N,e})^2C_{w_y,u_y}^2\sum_{j=1}^{N}\left(\lambda(t_j)(t_j-t_{j-1})\bar{\sigma}_{t_j}+\beta\omega(t_j)(t_j-t_{j-1})\right)$. *Then for* $\delta>0,\delta\ll 1$, $w.p.\ 1-n_yN\delta$, $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})$ *is* $\frac{L_y}{m^2}$ *smooth and convex in* $\theta_y$.

**Proof** We have,

$$
\begin{aligned}
\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y}) =& \frac{1}{2n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\frac{\lambda(t_j)(t_j-t_{j-1})}{\bar{\sigma}_{t_j}}\Bigg[\left\|\bar{\sigma}_{t_j}s_{t_j,\theta_y}(x(t_j))+\xi_{ij}\right\|_2^2 \\
&+\beta\omega(t_j)(t_j-t_{j-1})\mathbb{E}_{y'\sim Q}\left[\left\|s_{t_j,\theta_y}(x(t_j))-s_{t,\theta_{y'}}(x(t_j))\right\|_2^2\right]\Bigg]
\end{aligned}
\tag{46}
$$

To show smoothness, we will show that the function $f(\theta_y) = \left\|\bar{\sigma}_{t_j} s_{t_j,\theta_y}(x(t_j)) + \xi_{ij}\right\|_2^2$ and $g(\theta_y) = \left\|s_{t_j,\theta_y}(x_{t_j}) - s_{t_j,\theta_{y'}}(x(t_j))\right\|_2^2$ are individually smooth. Once we prove this, it is easy to show $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ is smooth as the linear combination of smooth functions is again smooth. To show smoothness, we need to show that $\left\|\nabla^2_{\theta_y} f(\theta_y)\right\|$ and $\left\|\nabla^2_{\theta_y} g(\theta_y)\right\|$ have a bounded norm. Recall that $s_{t,\theta_y}(x) = \frac{1}{m} A_y \sigma(W_y x(t) + U_y e(t))$. Let $h_1(x,t) := \sigma(W_y x + U_y e(t)), h_2(x,t) := s_{t,\theta_{y'}}(x), h_3(i,j) = \xi_{ij}$, we have

$$
\begin{aligned}
f(\theta_y) =& \left\|\bar{\sigma}_{t_j} s_{t_j,\theta_y}(x(t_j)) + \xi_{ij}\right\|^2 && (47)\\
=& \frac{\bar{\sigma}_{t_j}^2}{m^2} h_1^T(x(t_j),t_j) A_y^T A_y h_1(x(t_j),t_j) - 2\bar{\sigma}_{t_j} h_3^T(i,j)\left(\frac{A_y}{m}\right) h_1(x(t_j),t_j) && (48)\\
&+ h_3^T(i,j) h_3(i,j) && (49)\\
\overset{a}{=}& \frac{\bar{\sigma}_{t_j}^2}{m^2}\text{trace}(A_y^T A_y B_1) - \frac{2\bar{\sigma}_{t_j}}{m}\text{trace}(A_y B_3) + \text{constant} && (50)\\
\overset{b}{=}& \frac{\bar{\sigma}_{t_j}^2}{m^2} vec(A_y)^T (B_1 \bigotimes I) vec(A_y) - \frac{2\bar{\sigma}_{t_j}}{m} vec(B_3^T)^T vec(A_y) + \text{constant} && (51)
\end{aligned}
$$

where $(a)$ follows from the identity $x^T A y = \text{trace}(Byx^T)$, $(b)$ follows from the following identities

$$
\begin{aligned}
\text{trace}(A^T A B) =& \text{trace}(ABA^T) = vec(A)^T (B \bigotimes I) vec(A)\\
\text{trace}(AB) =& vec(A)^T vec(B^T)
\end{aligned}
$$

and $B_3 = h_1(x(t_j),t_j) h_3^T(i,j)$.

Similarly, we have for $g(\theta_y)$

$$
\begin{aligned}
g(\theta_y) =& \left\|s_{t,\theta_y}(x(t_j)) - s_{t,\theta_{y'}}(x(t_j))\right\|^2\\
=& \frac{1}{m^2} h_1^T(x(t_j),t_j) A_y^T A_y h_1(x(t_j),t_j) - 2 h_2^T(x(t_j),t_j)\left(\frac{A_y}{m}\right) h_1(x(t_j),t_j)\\
&+ h_2^T(x(t_j),t_j) h_2(x(t_j),t_j)\\
\overset{a}{=}& \frac{1}{m}\text{trace}(A_y^T A_y B_1) - \frac{2}{m}\text{trace}(A_y B_2) + \text{constant}\\
\overset{b}{=}& \frac{1}{m^2} vec(A_y)^T (B_1 \bigotimes I) vec(A_y) - \frac{2}{m} vec(B_2^T)^T vec(A) + \text{constant}
\end{aligned}
$$

where $B_1 := h_1(x(t_j),t_j) h_1^T(x(t_j),t_j)$ and $B_2 := h_1(x(t_j),t_j) h_2^T(x(t_j),t_j)$. Thus,

$$
\frac{1}{\bar{\sigma}_{t_j}^2}\nabla^2_{\theta_y} f(\theta_y) = \nabla^2_{\theta_y} g(\theta_y) = \nabla^2_{vec(A_y)} g(\theta_y) = \frac{2}{m^2}(B_1 \bigotimes I) \tag{52}
$$

The eigenvalues of $(B_1 \bigotimes I)$ is the same as $B_1$ with multiplicity. Thus, to show smoothness, we need to bound the maximum eigenvalues of $B_1$. For any $v \in \mathbb{R}^m$

$$
0 \le v^T B_1 v = (v^T h_1(x(t_j),t_j))^2 \le d \left\|h_1(x(t_j),t_j)\right\|_\infty^2 v^T v \tag{53}
$$

Now,

$$
\begin{aligned}
\left\|\sigma(W_y x(t_j) + U_y e(t_j))\right\|_\infty =& \max_{i=1,\cdots,m} \sigma(w_{y,i}^T x(t_j) + u_{y,i}^T e(t_j)) && (54)\\
\le& \max_{i=1,\cdots,m} |w_{y,i}^T x(t_j) + u_{y,i}^T e(t_j)| && (55)\\
\le& \max_{i=1,\cdots,m} \left\|w_{y,i}\right\|_1 \left\|x(t_j)\right\|_\infty + \left\|u_{y,i}\right\|_1 \left\|e(t_j)\right\|_\infty && (56)\\
\le& (C_{t_N,\delta} + C_{t_N,e}) C_{w_y,u_y} && (57)
\end{aligned}
$$

Thus, we have for any $v \in \mathbb{R}^m$

$$0 \leq v^T B_1 v \leq d(C_{t_N,\delta} + C_{t_N,e})^2 C_{w_y,u_y}^2 v^T v \tag{58}$$

Since $w.p.\ 1 - n_y N\delta$ we have $\{\|x_{ij}\|_\infty \leq C_{t_N,\delta}\}_{i=1,j=1}^{n_y,N}$, we have with the same probability $f(\theta_y)$ and $g(\theta_y)$ are smooth in $\theta_y$ for every $W_y, U_y, x(t_j), \theta_{-y}$.

Thus, $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ is $L_y \frac{1}{m^2} = \frac{1}{m^2} d(C_{t_N,\delta} + C_{t_N,e})^2 C_{w_y,u_y}^2 \sum_{j=1}^{N} \Bigg( \lambda(t_j)(t_j - t_{j-1})\bar{\sigma}_{t_j} + \beta\omega(t_j)(t_j - t_{j-1}) \Bigg)$ smooth.

### B.6 Proof: First order convergence of the algorithm

**Proof** Our proof follows closely along the lines of [11]. Let $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ be the empirical version of $\mathcal{L}_{reg}^{y}(\theta_y, \theta_{-y})$ with $n_y$ samples. By $L_y$ smoothness of $\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y, \theta_{-y})$ we have, for any $y \in \mathcal{Y}$,

$$\begin{aligned} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau+1}, \theta_{-y}^{\tau}) \leq \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) + \langle \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}), \theta_y^{\tau+1} - \theta_y^{\tau} \rangle \\ + \frac{L_y}{2} \left\| \theta_y^{\tau+1} - \theta_y^{\tau} \right\|^2 \end{aligned} \tag{59}$$

$$\begin{aligned} \implies \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau+1}, \theta_{-y}^{\tau}) \leq \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) - \eta_\tau \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \\ + \frac{L_y}{2} \eta_\tau^2 \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \end{aligned} \tag{60}$$

$$\begin{aligned} \implies \eta_\tau \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \leq \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) - \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau+1}, \theta_{-y}^{\tau}) \\ + \frac{L_y}{2} \eta_\tau^2 \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{t}, \theta_{-y}^{t}) \right\|^2 \end{aligned} \tag{61}$$

$$\begin{aligned} \implies \sum_{\tau=1}^{T_{train}} \eta_\tau \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \leq \bar{\mathcal{L}}^{n_y}(\theta_y^{t}) - \bar{\mathcal{L}}^{n_y}(\theta_y^{\tau+1}) + \beta \sum_{\tau=1}^{T_{train}} \psi(\theta_y^{\tau+1}, \theta_y^{\tau}, \theta_{-y}^{\tau}) \\ + \sum_{\tau=1}^{T_{train}} \frac{L_y}{2} \eta_\tau^2 \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \end{aligned} \tag{62}$$

$$\begin{aligned} \implies \sum_{\tau=1}^{T_{train}} \eta_\tau \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \leq \bar{\mathcal{L}}^{n_y}(\theta_y^{1}) - \bar{\mathcal{L}}^{n_y}(\theta_y^{\tau+1}) + \sum_{\tau=1}^{T_{train}} \frac{L_y}{2} \eta_\tau^2 \sigma_y^2 \\ + \beta \sum_{\tau=1}^{T_{train}} \psi(\theta_y^{\tau+1}, \theta_y^{\tau}, \theta_{-y}^{\tau}) \end{aligned} \tag{63}$$

$$\begin{aligned} \implies \min_{\tau \in [T_{train}]} \left\| \nabla_{\theta_y} \bar{\mathcal{L}}_{reg}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) \right\|^2 \leq \frac{\bar{\mathcal{L}}^{n_y}(\theta_y^{1}) - \bar{\mathcal{L}}^{n_y}(\theta_y^{*}) + \frac{L_y}{2} \sigma_y^2 \sum_{\tau=1}^{T_{train}} \eta_\tau^2}{\sum_{\tau=1}^{T_{train}} \eta_\tau} \\ + \beta \frac{\sum_{t=1}^{T_{train}} \psi(\theta_y^{\tau+1}, \theta_y^{\tau}, \theta_{-y}^{\tau})}{\sum_{\tau=1}^{T} \eta_\tau} \end{aligned} \tag{64}$$

where $\psi(\theta_y^{\tau+1}, \theta_y^{\tau}, \theta_{-y}^{\tau}) = \bar{\mathcal{L}}_{mut}^{n_y}(\theta_y^{\tau}, \theta_{-y}^{\tau}) - \bar{\mathcal{L}}_{mut}^{n_y}(\theta_y^{\tau+1}, \theta_{-y}^{\tau})$.

#### B.6.1 Analyzing the Bias Term

**Lemma 4.** *Suppose $\theta_{-y}, (W_y, U_y)_{y \in \mathcal{Y}}$ are fixed. Let $\phi_y = d^{1.5} N (C_{t_N,\delta} + C_{t_N,e})^2 C_{w_y,u_y}^2 B \max_j \omega(t_j)(t_j - t_{j-1})$. Then for $\delta > 0, \delta \ll 1, w.p. 1 - n_y N\delta$, we have*

$$\bar{\mathcal{L}}_{mut}^{n_y}(\theta_y, \theta_{-y}) = \frac{1}{2n_y} \sum_{i=1}^{n_y} \sum_{j=1}^{N} \omega(t_j)(t_j - t_{j-1}) \mathbb{E}_{y' \sim Q} \left[ \left\| s_{t_j,\theta_y}(x(t_j)) - s_{t,\theta_{y'}}(x(t_j)) \right\|_2^2 \right] \tag{65}$$

*is $\phi_y$ Lipschitz in $\theta_y$.*

**Proof:**

$$\left\|\nabla_{A_y}\bar{\mathcal{L}}^{n_y}_{reg,mut}(\theta_y,\theta_{-y})\right\|_F^2=\sum_{k=1}^{d}\left\|\nabla_{(A_y)_k}\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})\right\|^2$$

$$=\sum_{k=1}^{d}\left\|\frac{1}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\omega(t_j)(t_j-t_{j-1})\mathbb{E}_{y'}[(s_{t_j,\theta_y}(x(t_j))-s_{t_j,\theta_{y'}}(x(t_j)))_k\sigma(W_yx(t_j)+U_ye(t_j))]\right\|^2$$

$$\leq\frac{N}{n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\omega(t_j)^2(t_j-t_{j-1})^2.$$

$$\sum_{k=1}^{d}\mathbb{E}_{y'}[\left\|s_{t_j,\theta_y}(x(t_j))-s_{t_j,\theta_{y'}}(x(t_j))\right\|^2]\left\|\sigma(W_yx(t_j)+U_ye(t_j))\right\|^2$$
$$\leq\left\|\sigma(W_yx(t_j)+U_ye(t_j))\right\|^2Nd\max_j\omega(t_j)(t_j-t_{j-1})\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})$$
$$\leq 4d^2N(C_{t_N,\delta}+C_{t_N,e})^2C^2_{w_y,u_y}\max_j\omega(t_j)(t_j-t_{j-1})\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})$$
$$\leq d^3N(C_{t_N,\delta}+C_{t_N,e})^4C^4_{w_y,u_y}B^2\max_j\omega(t_j)(t_j-t_{j-1})\sum_{j=1}^{N}\omega(t_j)(t_j-t_{j-1})$$

Since $\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})\leq\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})$ and $w.p.1-n_yN\delta$, $\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})$ is bounded. Thus, $\left\|\nabla_{A_y}\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})\right\|_F^2$ is bounded and hence $\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})$ is Lipschitz in $\theta_y$ with $\phi_y=d^{1.5}N(C_{t_N,\delta}+C_{t_N,e})^2C^2_{w_y,u_y}B\max_j\omega(t_j)(t_j-t_{j-1})$ Here,

$$\psi(\theta_y^{\tau+1},\theta_y^{\tau},\theta_{-y}^{\tau})=\left|\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y^{\tau},\theta_{-y}^{t})-\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y^{\tau+1},\theta_{-y}^{\tau})\right| \tag{66}$$
$$\leq\phi_y\left\|\theta_y^{\tau}-\theta_y^{\tau+1}\right\| \tag{67}$$
$$\leq\phi_y\eta_t\left\|\nabla_{A_y}\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y^{\tau},\theta_{-y}^{\tau})\right\| \tag{68}$$
$$\leq\phi_y\eta_\tau\sigma_y \tag{69}$$

By taking $\eta_\tau\leq\frac{m^2}{\max_{y\in\mathcal{Y}}L_y\sqrt{T_{train}}},\forall y\in\mathcal{Y}$

$$\max_{y\in\mathcal{Y}}\min_{\tau\in[T_{train}]}\left\|\nabla_{\theta_y}\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y^{\tau},\theta_{-y}^{\tau})\right\|^2=\max_{y\in\mathcal{Y}}\mathcal{O}\left(\frac{2(\bar{\mathcal{L}}^{n_y}(\theta_y^0))-\bar{\mathcal{L}}^{n_y}(\theta_y^*))}{\max_{y\in\mathcal{Y}}L_y\sqrt{T_{train}}}+\frac{\sigma_y^2}{\sqrt{T_{train}}}+\beta\phi_y\sigma_y\right) \tag{70}$$
$$=\mathcal{O}\left(\frac{m^2}{\sqrt{T_{train}}}+\beta\right) \tag{71}$$

For $(\theta_y,\theta_{-y})$, we have

$$\text{NE-gap}(\theta_y,\theta_{-y})=\max_{y\in Y}\mid\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})-\bar{\mathcal{L}}^{n_y}_{reg}(B(\theta_{-y}),\theta_{-y})\mid \tag{72}$$
$$\leq\max_{y\in\mathcal{Y}}\left\|\nabla_{\theta_y}\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})\right\|^2\left\|\theta_y-B(\theta_{-y})\right\|_2^2 \tag{73}$$

Since the strategy space for $\theta_y$ is bounded in norm. We have

$$\text{NE-gap}(\theta_y,\theta_{-y})\lesssim\max_{y\in\mathcal{Y}}\left\|\nabla_{\theta_y}\bar{\mathcal{L}}^{n_y}_{reg}(\theta_y,\theta_{-y})\right\|^2 \tag{74}$$
$$\implies\min_{\tau\in[T_{train}]}\text{NE-gap}(\theta_y^{\tau},\theta_{-y}^{\tau})=\mathcal{O}\left(\frac{m^2}{\sqrt{T_{train}}}+\beta\right) \tag{75}$$

### B.7 Monte Carlo Error of the Finite Neural Network

Observe that

$$\mathcal{L}^y(\theta_y) = \frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})\mathbb{E}_{X_0}\mathbb{E}_{X_{t_j}|X_0}\left[\left\|\nabla_{x(t_j)}\log p_t(x_i(t_j)|x_0) - s_{t_j,\theta_y}(x_i(t_j))\right\|_2^2\right] \tag{76}$$

$$+ \frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})C_{t_j} \tag{77}$$

$$= \frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})\mathbb{E}_{x(t_j)\sim p_{t_j}}\left[\left\|s_{t_j,\theta_y}(x(t_j)) - \nabla_x \log p_{t_j}(x(t_j))\right\|_2^2\right]\Bigg] \tag{78}$$

For each $y \in \mathcal{Y}$, $\mathcal{L}^y(\theta_y^*)$ is the optimal loss function for the unregularized version under the current hypothesis class. Let $\mathcal{L}^y(\bar{\theta}_y^*)$ be the optimal unregularized loss function under the continuous version of the random feature model. Then,

$$\mathcal{L}^y(\theta_y^*) = \frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})\mathbb{E}_{x(t_j)\sim p_{t_j}}\left[\left\|s_{t_j,\theta_y^*}(x(t_j)) - \nabla_x \log p_{t_j}(x(t_j))\right\|_2^2\right]\Bigg] \tag{79}$$

$$\leq 2\Bigg(\frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})\mathbb{E}_{x(t_j)\sim p_{t_j}}\left[\left\|\bar{s}_{t,\bar{\theta}_y^*}(x(t)) - \nabla_x \log p_t(x(t))\right\|_2^2\right] \tag{80}$$

$$+ \frac{1}{2}\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1})\mathbb{E}_{x(t_j)\sim p_{t_j}}\left[\left\|\bar{s}_{t_j,\bar{\theta}_y^*}(x(t_j)) - s_{t_j,\theta_y^*}(x(t_j))\right\|_2^2\right]\Bigg) \tag{81}$$

$$\leq 2\mathcal{L}^y(\bar{\theta}_y^*) + Err_{MC}(\theta_y^*, \bar{\theta}_y^*; \{t_j\}_{j=1}^N, \{\lambda(t_j)\}_{j=1}^N) \tag{82}$$

**Proposition 2.** ***Monte Carlo estimates***. *Define the Monte Carlo error*

$$Err_{MC}(\theta, \bar{\theta}, \{t_j\}_{j=1}^N, \{\lambda(t_j)\}_{j=1}^N) := \sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1}). \\ \mathbb{E}_{x(t_j)\sim p_{t_j}}\left[\left\|\bar{s}_{t_j,\bar{\theta}_y^*}(x(t_j)) - s_{t_j,\theta_y^*}(x(t_j))\right\|_2^2\right] \tag{83}$$

*Suppose that* $\|X(0)\|_\infty \leq K$ *and the trainable parameter* $a$ *and embedding functions* $W, U, e(.)$ *are both bounded. Then. given any* $\bar{\theta}$*. for any* $\delta > 0, \delta \ll 1$*, with probability of at least* $1 - 2N\delta$*, there exists* $\theta$ *such that*

$$Err_{MC}(\theta, \bar{\theta}, \{t_j\}_{j=1}^N, \{\lambda(t_j)\}_{j=1}^N) \leq \frac{2C_{w,u}^2 B^2 (C_{t_N,\delta} + C_{t_N.e})^2 d^2}{m}\log\left(\frac{2}{\delta}\right)\sum_{j=1}^{N}\lambda(t_j)(t_j - t_{j-1}) \tag{84}$$

**Proof.** The proof closely along the line of [15]. Fix any $\bar{\theta}$. For notational convenience, we will drop $y$ from $\theta_y$ and $\bar{\theta}_y$. For $k = 1, 2, \cdots d$, define

$$Z_{t,k}(W, U) := \left\|s_{t,\theta,k}(x) - \bar{s}_{t,\bar{\theta},k}(x)\right\|_{L^2(p_t)} = \mathbb{E}_{x\sim p_t}^{1/2}\left[|s_{t,\theta,k}(x) - \bar{s}_{t,\bar{\theta},k}(x)|^2\right] \tag{85}$$

$$= \mathbb{E}_{x\sim p_t}\left[\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,k}\sigma(w_i^T x + u_i^T e(t)) - \mathbb{E}_{(w,u)}[a_k(w,u)\sigma(w^T x + u^T e(t))]\right|^2\right] \tag{86}$$

Then, we have

$$\mathbb{E}_{x\sim p_t}\left[\left\|s_{t,\theta_y}(x)-\bar{s}_{t,\bar{\theta}_y}(x)\right\|_2^2\right]=\sum_{k=1}^{d}\mathbb{E}_{x\sim p_t}\left[\left|s_{t,\theta_y,k}(x)-\bar{s}_{t,\bar{\theta}_y,k}\right|^2\right] \tag{87}$$

$$=\sum_{k=1}^{d}Z_{t,k}^2(W,U) \tag{88}$$

$$\leq\sum_{k=1}^{d}\left(\left|Z_{t,k}(W,U)-\mathbb{E}_{W,U}[Z_{t,k}]\right|+\left|\mathbb{E}_{W,U}[Z_{t,k}(W,U)]\right|\right)^2 \tag{89}$$

$$\overset{(a)}{\leq}2\sum_{k=1}^{d}\Bigg(\left|Z_{t,k}(W,U)-\mathbb{E}_{W,U}[Z_{t,k}(W,U)]\right|^2 \tag{90}$$

$$+\mathbb{E}_{W,U}[Z_{t,k}^2(W,U)]\Bigg) \tag{91}$$

where $(a)$ follows from the fact that $(a+b)^2\leq 2(a^2+b^2)$ and Jensen's Inequality $\mathbb{E}^2[Z_{t,k}(W,U)]\leq \mathbb{E}_{W,U}[Z_{t,k}^2(W,U)]$. According to Lemma 1. for any $\delta>0,\delta\ll 1$, w.p. atleast $1-\delta$, we have

$$\left\|x(t)\right\|_\infty\leq C_{t_N,\delta} \tag{92}$$

If $(\tilde{W},\tilde{U})$ is different from $(W,U)$ at only one component indexed by $i$, we have w.p. $1-\delta$

$$\left|Z_{t,k}(W,U)-Z_{t,k}(\tilde{W},\tilde{U})\right| \tag{93}$$

$$=\left|\left\|s_{t,\theta,k}(x)-\bar{s}_{t,\bar{\theta},k}(x)\right\|_{L^2(p_t)}-\left\|s_{t,\tilde{\theta},k}(x)-\bar{s}_{t,\bar{\theta},k}(x)\right\|_{L^2(p_t)}\right| \tag{94}$$

$$\overset{(a)}{\leq}\left\|s_{t,\tilde{\theta},k}(x)-s_{t,\theta,k}(x)\right\|_{L^2(p_t)} \tag{95}$$

$$=\frac{1}{m}\left\|a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\tilde{a}_{i,k}\sigma(\tilde{w}_i^Tx+\tilde{u}_i^Te(t))\right\|_{L^2(p_t)} \tag{96}$$

$$\overset{(b)}{\leq}\frac{1}{m}\left(\left|a_{i,k}\right|\left\|\sigma(w_i^Tx+u_i^Te(t))\right\|_{L^2(p_t)}+\left|\tilde{a}_{i,k}\right|\left\|\sigma(\tilde{w}_i^Tx+\tilde{u}_i^Te(t))\right\|_{L^2(p_t)}\right) \tag{97}$$

$$\overset{(c)}{\leq}\frac{1}{m}\left(\left|a_{i,k}\right|\left\|w_i^Tx+u_i^Te(t))\right\|_{L^2(p_t)}+\left|\tilde{a}_{i,k}\right|\left\|(\tilde{w}_i^Tx+\tilde{u}_i^Te(t))\right\|_{L^2(p_t)}\right) \tag{98}$$

$$\overset{(d)}{\leq}\frac{1}{m}\left(\left|a_{i,k}\right|(\left\|w_i\right\|_1C_{t_N,\delta}+\left\|u_i\right\|_1\left\|e(t)\right\|_\infty)+\left|\tilde{a}_{i,k}\right|(\left\|\tilde{w}_i\right\|_1C_{t_N,\delta}+\left\|\tilde{u}_i\right\|_1\left\|e(t)\right\|_\infty)\right) \tag{99}$$

$$\overset{(e)}{\leq}\frac{2}{m}BC_{w,u}(C_{t_N,\delta}+C_{t_N,e}) \tag{100}$$

where $(a)$ and $(b)$ follows from triangle inequality $|\,\|a\|-\|b\|\,|\leq\|a-b\|$ and $\|a-b\|\leq\|a\|+\|b\|$, $(c)$ follows from the fact that $|\sigma(y)|\leq|y|$, $(d)$ follows from Lemma 1 and Holder Inequality, $(e)$ follows from the bounds on $\left\|w_i\right\|_1,\left\|u_i\right\|_1,x,|a_{i,k}|,e(t_j)$.

Thus, $w,p.1-\delta$, $Z_{t,k}(W,U)$ has bounded increment property. Using McDiarmid's inequality, $w.p.1-2\delta$, we have

$$\left|Z_{t,k}(W,U)-\mathbb{E}_{W,U}[Z_{t,k}(W,U)]\right|\leq\frac{B}{m}C_{w,u}(C_{t_N,\delta}+C_{t_N.e})\sqrt{d\log(\frac{2}{\delta})} \tag{101}$$

Now we compute

$$
\begin{aligned}
&\mathbb{E}_{W,U}[Z_{t,k}^2(W,U)] \\
=&\mathbb{E}_{W,U}\left[\mathbb{E}_{x\sim p_t}[|s_{t,\theta,k}(x)-\bar{s}_{t,\bar{\theta},k}(x)|^2]\right] \\
\overset{(b)}{=}&\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{W,U}[|s_{t,\theta,k}(x)-\bar{s}_{t,\bar{\theta},k}(x)|^2]\right] \\
=&\frac{1}{m^2}\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{W,U}\left[\left|\sum_{i=1}^m\left(a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+u^Te(t))]\right)\right|^2\right]\right] \\
&+\frac{1}{m^2}\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{W,U}\left[\sum_{i\neq j}(a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+u^Te(t))]\right.\right. \\
&\left.\left.\times(a_{j,k}\sigma(w_j^Tx+u_j^Te(t))-\mathbb{E}_{w,u}[a_k\sigma(w^Tx+u^Te(t))])\right]\right] \\
\overset{(c)}{=}&\frac{1}{m^2}\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{W,U}\left[\sum_{i=1}^m\left(a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+u^Te(t))]\right)^2\right]\right] \\
&+\frac{1}{m^2}\mathbb{E}_{x\sim p_t}\left[\sum_{i\neq j}\mathbb{E}_{w_i,u_i}\left[(a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+Ue(t))]\right.\right. \\
&\left.\left.\times\mathbb{E}_{w_j,u_j}\left[(a_{j,k}\sigma(w_j^Tx+u_j^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+Ue(t))]\right]\right] \\
\overset{(d)}{=}&\frac{1}{m^2}\mathbb{E}_{x\sim p_t}\left[\sum_{i=1}^m\mathbb{E}_{W,U}\left[\left(a_{i,k}\sigma(w_i^Tx+u_i^Te(t))-\mathbb{E}_{w,u}[a_k(w,u)\sigma(w^Tx+u^Te(t))]\right)^2\right]\right] \\
\overset{(e)}{\leq}&\frac{1}{m}\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{w,u}\left[(a_k(w,u)\sigma(w^Tx+u^Te(t)))^2\right]\right] \\
\overset{(f)}{\leq}&\frac{1}{m}\mathbb{E}_{x\sim p_t}\left[\mathbb{E}_{w,u}\left[(|a_{y,k}(w,u)|(\|w\|_1C_{t_N,\delta}+\|u\|_1\|e(t)\|_\infty))^2\right]\right] \\
\leq&\frac{1}{m}(C_{t_N,\delta}+C_{t_N,e})^2C_{w,u}^2B^2
\end{aligned}
$$

where $(b)$ is due to Fubini's theorem, $(c)$ is due to independence of sampling $(w_i,u_i)$ and $(w_j,u_j)$, $(d)$ is due to $a_{j,k}\sigma(w_j^Tx+u_j^Te(t))$ being an unbiased estimator of the continuous version of score network, $(e)$ follows from $Var(X)\leq\mathbb{E}[X^2]$, $(f)$ follows from $|\sigma(y)|\leq|y|$ and Holder's inequality. Thus. $w.p.1-2\delta$,

$$
\mathbb{E}_{x\sim p_t}\left[\left\|s_{t,\theta_y}(x)-\bar{s}_{t,\bar{\theta}_y}(x)\right\|_2^2\right]\leq\frac{2C_{w,u}^2B^2(C_{t_N,\delta}+C_{t_N.e})^2d^2}{m}\log(\frac{2}{\delta}) \tag{102}
$$

Finally, we have $w.p.1-2N\delta$

$$
Err_{MC}(\theta,\bar{\theta},\{t_j\}_{j=1}^N,\{\lambda(t_j)\}_{j=1}^N)\leq\frac{2C_{w,u}^2B^2(C_{t_N,\delta}+C_{t_N.e})^2d^2}{m}\log(\frac{2}{\delta})\sum_{j=1}^N\lambda(t_j)(t_j-t_{j-1}) \tag{103}
$$

### B.8 Radamacher Complexity

In this section, we will bound the term related to the generalization bound

$$
\sup_{(\theta_y,\theta_{-y})}\mid\mathcal{L}_{reg}^y(\theta_y,\theta_{-y})-\bar{\mathcal{L}}_{reg}^{n_y}(\theta_y,\theta_{-y})\mid \tag{104}
$$

The Rademacher complexity of a real valued function class $\mathcal{F}$ is defined as:

$$
\mathcal{R}_n(\mathcal{F}):=\mathbb{E}_{x_1,\cdots,x_n}\mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_if(x_i)\right], \tag{105}
$$

The variables $\sigma_1, \cdots, \sigma_m$ are iid Bernoulli random variables that take values $\{+1, -1\}$ with equal probability and are independent of $x_1, \cdots, x_m$. However, for our random feature model, we have a vector valued function class

$$\hat{\mathcal{F}}_{W,U} := \left\{ f(x) = \frac{A}{m}\Phi(x, W, U) = \frac{1}{m}\sum_{k=1}^{m} \alpha_k \phi(x, w_k, u_k) \middle| \|A\|_F \leq B \right\} \tag{106}$$

**Theorem 4.** *[17, Theorem 3] Let $X$ be nontrivial, symmetric and subgaussian. Then there exists a constant $C < \infty$, depending only on the distribution of $X$, such that for any countable set $S$ and functions $\psi_i : S \to \mathbb{R}, \phi_i : S \to l_2, 1 \leq i \leq n$ satisfying*

$$\forall s, s' \in S, \psi_i(s) - \psi(s_i') \leq \|\phi_i(s) - \phi_i(s')\| \tag{107}$$

*we have*

$$\mathbb{E}\sup_{s\in S}\sum_i \epsilon_i \psi_i(s) \leq C\mathbb{E}\sup_{s\in S}\sum_{i,k} X_{ik}\phi_i(s)_k \tag{108}$$

*where the $X_{ik}$ are independent copies of $X$ for $1 \leq i \leq n$ and $1 \leq k \leq \infty$ and $\phi_i(s)_k$ is the k-th coordinate of $\phi_i(s)$. If $X$ is a Rademacher variable we may choose $C = \sqrt{2}$, if $X$ is a standard normal $C = \sqrt{\frac{\pi}{2}}$.*

**Corollary 2.** *[17, Corollary 4] Let $\mathcal{X}$ be any set, $(x_1, \cdots, x_n) \in \mathcal{X}^n$, let $F$ be a class of functions $f : \mathcal{X} \to l_2$ and let $h_i : l_2 \to \mathbb{R}$ have Lipschitz norm $L$. Then*

$$\mathbb{E}\sup_{f\in F}\sum_i \epsilon_i h_i(f(x_i)) \leq \sqrt{2}L\mathbb{E}\sup_{f\in F}\sum_{i,k}\epsilon_{ik} f_k(x_i) \tag{109}$$

*where $\epsilon_{ik}$ is an independent doubly indexed Rademacher sequence and $f_k(x_i)$ is the k-th component of $f(x_i)$.*

**Lemma 5.** *[17] Consider the function class $\mathcal{F} = \{x \to \frac{A}{m}\phi(x, W, U) : A \in \mathcal{B}(H, \mathbb{R}), \|A\|_F \leq B\}$. Then the empirical Rademacher complexity of $F$ is*

$$\hat{Rad}_n(\mathcal{F}) = \mathbb{E}\sup_{f\in F}\sum_{i,k}\epsilon_{ik} f_k(x_i) \leq \frac{B}{\sqrt{m}}\sqrt{\sum_i \|\phi(x, W, U)\|^2} \tag{110}$$

*Moreover, if $\mathbb{E}_x \|\phi(x, W, U)\|^2 \leq C^2$, the Rademacher Complexity of $\mathcal{F}$ is*

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{BC}{\sqrt{mn}} \tag{111}$$

**Proof:**

$$\hat{Rad}_n(\mathcal{F}) = \mathbb{E}\sup_{f\in F}\sum_{i,k}\epsilon_{ik} f_k(x_i) = \frac{1}{m}\mathbb{E}\sup_{\|A\|_F\leq B}\sum_k \langle a_k, \sum_i \epsilon_{ik} x_i\rangle \tag{112}$$

$$= \frac{1}{m}\mathbb{E}\sup_{\|A\|_F\leq B} tr(D^* A) \leq B\mathbb{E}\|D^*\|_* \tag{113}$$

where $D \in \mathcal{B}(H, \mathbb{R}^K)$ is the random transformation

$$v \to \left(\langle v, \sum_i \epsilon_{i1} x_i\rangle, \cdots, \langle v, \sum_i \epsilon_{iK} x_i\rangle\right) \tag{114}$$

Thus,

$$\mathbb{E}\|D^*\|_* = \mathbb{E}\sqrt{\sum_m \left\|\sum_i \epsilon_{ik}\phi(x_i, W, U)\right\|^2} \leq \sqrt{m\sum_i \|\phi(x_i, W, U)\|^2} \tag{115}$$

Thus,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{x_1,\cdots,x_n}\frac{1}{n}\hat{Rad}_n(\mathcal{F}) \leq \frac{B}{\sqrt{mn}}\mathbb{E}_{x_1,\cdots,x_n}\sqrt{\sum_i \|\phi(x_i,W,U)\|^2} \tag{116}$$

$$\leq \frac{B}{n\sqrt{m}}\sqrt{\sum_i \mathbb{E}_{x_1,\cdots,x_n}\|\phi(x_i,W,U)\|^2} \tag{117}$$

$$\leq \frac{BC}{\sqrt{mn}} \tag{118}$$

Suppose $0 < t_1 < \cdots < t_N = T$ are the chosen points of discretization for training, we have from the forward process

$$X(t) = e^{-t}X(0) + \sqrt{1-e^{-2t}}Z, Z \sim N(0,1) \tag{119}$$

$$\implies \mathbb{E}_Z[X^2(t)] = e^{-2t}x^2(0) + \frac{1-e^{-2t}}{2} \tag{120}$$

$$\implies \mathbb{E}_{X(0)}\mathbb{E}_Z[X^2(t_j)] \leq K^2 + \frac{1-e^{-2T}}{2}, \forall 0 < t_j < T \tag{121}$$

Using the above bounds along with bounded support of embedding matrices $W, U$ and embedding function $e(t)$ and Assumption 1, it is easy to show that

$$\mathbb{E}_{X_0}\mathbb{E}_{\xi_j}\left[\|\sigma(Wx(t)+Ue(t))\|_2^2\right] \leq F_T^2, \forall 0 < t \leq T \tag{122}$$

for some constant $F_T^2$ and $x(t) = e^{-t}x(0) + \sqrt{1-e^{-2t}}\xi_j, \xi_j \sim \mathcal{N}(0, I)$

**Lemma 6.** *The term*

$$\mathcal{L}^y_{mut}(\theta_y, \theta_{-y}) = \sum_{j=1}^{N}\omega(t_j)(t_j - t_{j-1})\mathbb{E}_{X_0}\mathbb{E}_{X_{t_j}|X_0}\mathbb{E}_{y'\sim Q}\left[\left\|s_{t_j,\theta_y}(x_i(t_j)) - s_{t,\theta_{y'}}(x_i(t_j))\right\|_2^2\right] \tag{123}$$

*is* $\mathcal{O}\left(F_T B\sum_{j=1}^{N}\omega(t_j)(t_j - t_{j-1})\right)$

**Proof:** Using the fact of bounded support of embedding matrices $W, U$ and embedding function $e(t)$, bounded strategy space and Assumption 1 and eq 122, we get the desired bounded.

**Lemma 7.** *Suppose* $L_{C_1} = \bar{\sigma}_{t_j}^2 BF_T + \sqrt{d}\sqrt{\log\frac{2}{\pi\delta^2}})$. *Then, with probability* $1-\delta$, *the function* $h : \mathcal{A} \subset \mathbb{R}^d \to \mathbb{R}$

$$h(x) = \left\|\bar{\sigma}_{t_j}x + \xi_{ij}\right\|^2 \tag{124}$$

*is Lipschitz in* $x$, *where* $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq F_T B\}$.

**Proof.** It is sufficient to show the norm of the gradient of $h(x)$ is bounded for $x \in \mathcal{A}$. With probability $1-\delta$,

$$\|\nabla_x h(x)\|_2 = \bar{\sigma}_{t_j}\left\|\bar{\sigma}_{t_j}x + \xi_{ij}\right\|_2 \leq \bar{\sigma}_{t_j}^2 F_T B + \sqrt{d}\sqrt{\log\frac{2}{\pi\delta^2}}) \tag{125}$$

$$\tag{126}$$

**Lemma 8.** *Suppose* $L_{C_2} = 2F_T B|\mathcal{Y}|$. *Define* $g : \mathcal{A}^{\mathcal{Y}} \subset \mathbb{R}^{d|\mathcal{Y}|} \to \mathbb{R}$ *where*

$$g(x_1, x_2, \cdots, x_{|\mathcal{Y}|}) = \mathbb{E}_{y'}[\|x_i - x_{y'}\|^2], y' \in \{1, 2, \cdots, |\mathcal{Y}|\} - i \tag{127}$$

*is Lipschitz in* $x$, *where* $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq F_T B\}$.

**Proof:**

$$\nabla_{x_i} g(x_1, x_2, \cdots, x_{|\mathcal{Y}|}) = 2\mathbb{E}_{y'}[(x_i - x_{y'})] \tag{128}$$
$$\nabla_{x_j} g(x_1, x_2, \cdots, x_{|\mathcal{Y}|}) = 2p(x_j)(x_j - x_i), j \neq i \tag{129}$$

$$\|\nabla_x g(x)\| \leq \left\|\nabla_{x_i} g(x_1, x_2, \cdots, x_{|\mathcal{Y}|})\right\| + \sum_{j \neq i} \left\|\nabla_{x_j} g(x_1, x_2, \cdots, x_{|\mathcal{Y}|})\right\| \tag{130}$$
$$\leq 2\mathbb{E}_{y'}[\|x_i - x_{y'}\|] + 2\sum_{k \neq i} \|x_k - x_i\| \leq 2F_T B|\mathcal{Y}| \tag{131}$$

We know

$$\mathcal{L}^y(\theta_y) = \frac{1}{2}\sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} \mathbb{E}_{X_0}\mathbb{E}_{\xi_j}\left[\left\|\bar{\sigma}_{t_j} s_{t_j,\theta_y}(x(t_j)) + \xi_j\right\|_2^2\right] \tag{132}$$
$$+ \frac{1}{2}\sum_{j=1}^{N} \lambda(t_j)(t_j - t_{j-1}) C_{t_j}(y) \tag{133}$$

where $C_t(y) = \mathbb{E}_{X_t}\left\|\nabla \log p_t(.|y)\right\|^2 - \mathbb{E}_{X_0}\mathbb{E}_{X_t|X_0}\left\|\nabla \log p_t(x_t|x_0, y)\right\|^2$. Let $\bar{C}(y) = \frac{1}{2}\sum_{j=1}^{N} \lambda(t_j)(t_j - t_{j-1}) C_{t_j}(y)$

**Lemma 9.** *With probability $1 - Nn_y\delta$, an upper bound for the generalization gap i.e.*

$$\sup_{(\theta_y, \theta_{-y})} \mid \mathcal{L}^y_{reg}(\theta_y, \theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y}) \mid \tag{134}$$

*is*

$$\frac{2\sqrt{2}BF_T}{\sqrt{mn_y}} L_{C_1} \sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} + \frac{2\sqrt{2}BF_T|\mathcal{Y}|^2}{\sqrt{mn_y}} L_{C_2} \sum_{j=1}^{N} \omega(t_j)(t_j - t_{j-1}) + \bar{\mathcal{C}} \tag{135}$$

*where* $L_{C_1} = \bar{\sigma}_{t_j}^2 BF_T + \sqrt{d}\sqrt{\log \frac{2}{\pi\delta^2}}$,$L_{C_2} = 2F_T B|\mathcal{Y}|$,$\bar{\mathcal{C}} = \max_{y \in \mathcal{Y}} |\bar{C}(y)|$

**Proof.** Observe that, we can rewrite Eq. 134 using triangle inequality as

$$\sup_{(\theta_y, \theta_{-y})} \mid \mathcal{L}^y_{reg}(\theta_y, \theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{reg}(\theta_y, \theta_{-y}) \mid \leq \sup_{\theta_y} |\mathcal{L}^y(\theta_y) - \bar{\mathcal{L}}^{n_y}(\theta_y)| + \beta \sup_{(\theta_y, \theta_{-y})} |\mathcal{L}^y_{mut}(\theta_y, \theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{mut}(\theta_y, \theta_{-y})| \tag{136}$$

Further decomposing them, we get

$$\sup_{\theta_y} |\mathcal{L}^y(\theta_y) - \bar{\mathcal{L}}^{n_y}(\theta_y)| \leq \sum_{j=1}^{N} \frac{\lambda(t_j)(t_j - t_{j-1})}{\bar{\sigma}_{t_j}} \sup_{\theta_y} |\mathcal{L}^y(\theta_y)(j) - \bar{\mathcal{L}}^{n_y}(\theta_y)(j)| + \bar{\mathcal{C}} \tag{137}$$

where $|\mathcal{L}^y(\theta_y)(j) - \bar{\mathcal{L}}^{n_y}(\theta_y)(j)| = \mid \frac{1}{2n_y}\sum_{i=1}^{n_y}\left[\left\|\bar{\sigma}_{t_j} s_{t_j,\theta_y}(x_i(t_j)) + \xi_{ij}\right\|_2^2 - \mathbb{E}_{X_0}\mathbb{E}_{\xi_j}\left\|\bar{\sigma}_{t_j} s_{t_j,\theta_y}(e^{-t_j}X_0 + \sqrt{1 - e^{-2t_j}}\xi_j) + \xi_j\right\|^2\right] \mid$ and

$$\sup_{(\theta_y, \theta_{-y})} |\mathcal{L}^y_{mut}(\theta_y, \theta_{-y}) - \bar{\mathcal{L}}^{n_y}_{mut}(\theta_y, \theta_{-y})| \leq \sum_{j=1}^{N} \omega(t_j)(t_j - t_{j-1}) \sup_{(\theta_y, \theta_{-y})} |\mathcal{L}^y_{mut}(\theta_y, \theta_{-y})(j) - \bar{\mathcal{L}}^{n_y}_{mut}(\theta_y, \theta_{-y})(j) \mid \tag{138}$$

where

$$|\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})(j)-\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})(j)\ |=|\ \frac{1}{2n_y}\sum_{i=1}^{n_y}\Big[\mathbb{E}_{y'\sim Q}\Big[\ \Big\|s_{t_j,\theta_y}(x_i(t_j))-s_{t,\theta_{y'}}(x_i(t_j))\Big\|_2^2$$
$$-\mathbb{E}_{X_0}\mathbb{E}_{\xi_j}\mathbb{E}_{y'\sim Q}[\Big\|s_{t_j,\theta_y}(e^{-t_j}X_0+\sqrt{1-e^{-2t_j}}\xi_j)-s_{t_j,\theta_{y'}}(e^{-t_j}X_0+\sqrt{1-e^{-2t_j}})\Big\|^2]\Big]\ | \tag{139}$$

Finally using Corollary 2, Lemmas 5 7,8, [17, Section 4.1] we have

$$\sup_{\theta_y}|\mathcal{L}^y(\theta_y)-\bar{\mathcal{L}}^{n_y}(\theta_y)|\leq\frac{2\sqrt{2}BF_T}{\sqrt{mn_y}}L_{C_1}\sum_{j=1}^{N}\lambda(t_j)(t_j-t_{j-1})+\bar{\mathcal{C}} \tag{140}$$

and

$$\sup_{(\theta_y,\theta_{-y})}|\mathcal{L}^y_{mut}(\theta_y,\theta_{-y})-\bar{\mathcal{L}}^{n_y}_{mut}(\theta_y,\theta_{-y})|\leq\frac{2\sqrt{2}BF_T|\mathcal{Y}|^2}{\sqrt{mn_y}}L_{C_2}\sum_{j=1}^{N}\omega(t_j)(t_j-t_{j-1}) \tag{141}$$

# C Numerical Experiments

**Computing resources.** The numerical experiments were conducted on a MacBook Air (2023) and Gilbreth. Gilbreth has heterogeneous hardware comprising of Nvidia V100, A100, A10, and A30 GPUs in separate sub-clusters. All the nodes are connected by 100 Gbps Infiniband interconnects. We used sub-cluster B with 16 nodes, 24 cores per node, 192 GB memory per node, 3 A30 (24 GB) per node. For more information follow this link.

The width of network $m = 16$, learning rate $\eta_\tau = 10^{-4}, \forall\tau, T_{train} = 5000$ is fixed for Adam optimizer. We set $\lambda(t) = \bar{\sigma}_t, \omega(t) = e^t$, total number of training samples is 50.

**Case one** We perform more empirical experiments on $d = 1$, imbalance ratio $r = 2.5, \beta = 0.01$. We compute the KL-divergence between the ground truth distribution and the learned model using the procedure in [15]. $P(x|y = 1) \sim \mathcal{N}(-\mu, \sigma^2)$ and class 2 is $P(x|y = 2) \sim \mathcal{N}(\mu, \sigma^2)$. We observe Fig. 2 the worst case KL divergence for the mutual learning case is lower than the vanilla when we change the distance between mean and the variance of each class label. The performance of head class doesn't worsened for small $\mu$. However, the head class performance suffers for mutual learning case when the distance between the mean increases. This might be because when the support of class distribution are farther apart mutual learning is not advantageous as transfer of knowledge between the class is not useful.

**Case two** We now consider a case with two classes with imbalance ratio $r = 2.5, \beta = 0.01$. Class 1 itself is a uniform mixture of two Gaussian i.e $P(x|y = 1) \sim \frac{1}{2}\mathcal{N}(-4, 3) + \frac{1}{2}\mathcal{N}(4, 3)$ and class 2 is $P(x|y = 2) \sim \mathcal{N}(0, 2)$ as in Fig. 3. We observe the Mutual Learning objective with our formulation have lower KL-divergence for both the classes compared to the vanilla diffusion models trained on each class. In this case, mutual learning allows useful transfer of knowledge between the classes increasing the performance for both. We hypothesize that under some notion of similarity between various class distributions, mutual learning is advantageous in improving the performance of all classes.

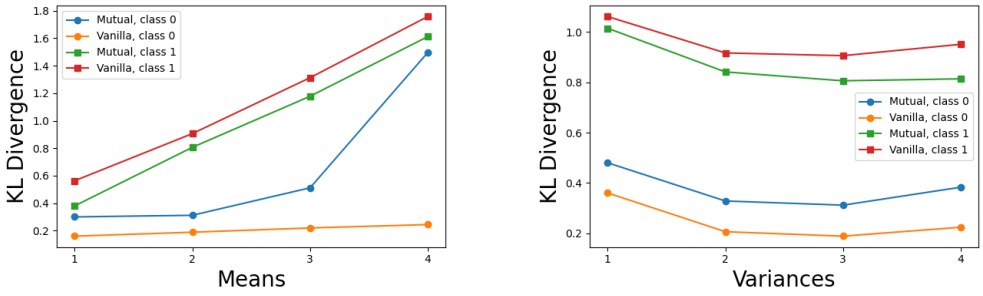


Figure 2: Case one: (Left) The first plot shows the KL-divergence for each class with and without mutual learning objective as $\mu$ is varied. (Right) shows the KL-divergence for each class with and without mutual learning objective as $\sigma$ is varied ($\mu = 2$ fixed).

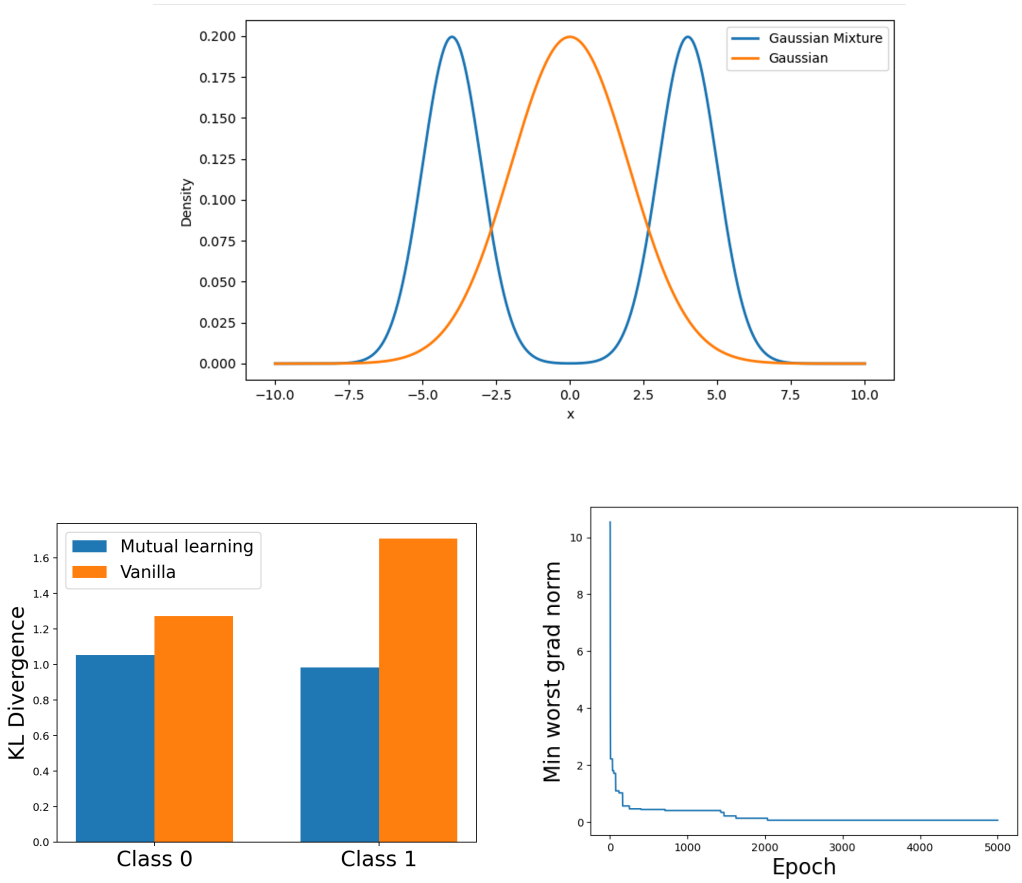


Figure 3: Case two: (Top) The first plot shows class 1 as a gaussian mixture with class 2 as Gaussian. (Bottom Left) Shows the KL-divergence for each class with and without mutual learning objective. (Bottom Right) Shows $\min_\tau \max_{y \in \mathcal{Y}} \left\| \nabla \bar{\mathcal{L}}^{n_y}_{reg}(\theta^\tau_y, \theta^\tau_{-y}) \right\|$ decreasing with training epoch.