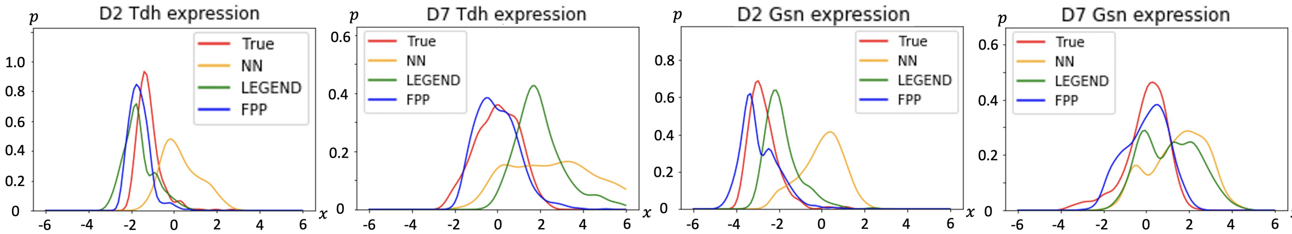
A. Supplementary Experiments

A.1. RNA-sequence



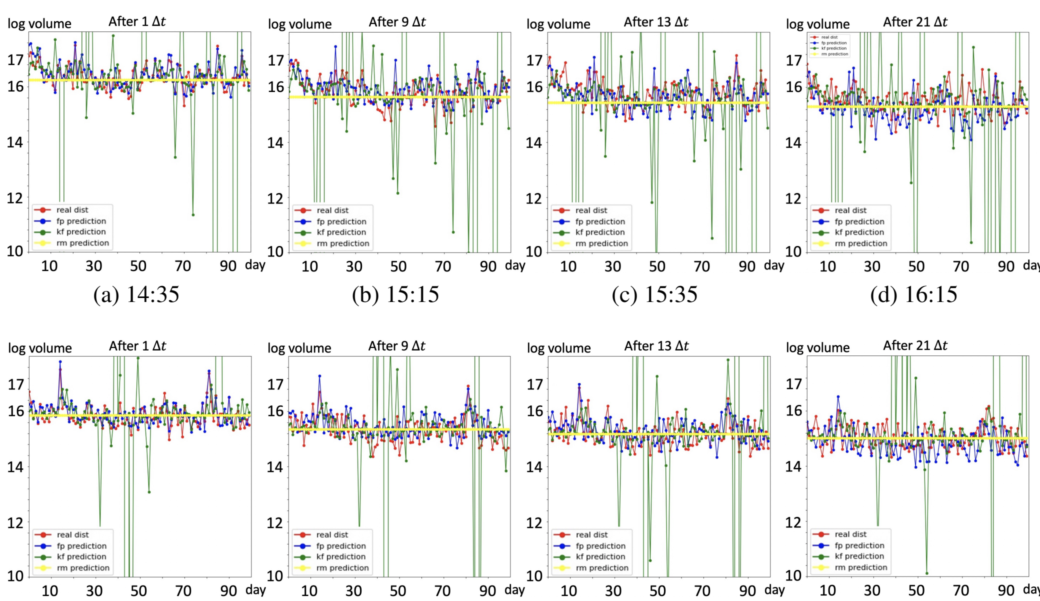
(a) D2 of Mt1 (b) D7 of Mt1 (c) D2 of Mt2 (d) D7 of Mt2

Figure 1: The performance comparisions among different models on D2 and D7 of Tdh and Gsn.

Table 1: The Wasserstein error of different models on Supplementary RNA-sequence data sets.

Data	Task	Dimension	NN	LEGEND	Ours
RNA-Tdh	D2	10	16.28	5.75	**2.15**
	D7	10	28.19	22.49	**1.03**
RNA-Gsn	D2	10	34.94	10.77	**3.31**
	D7	10	15.74	10.42	**2.07**

A.2. Daily Trading Volume



(a) 14:35 (b) 15:15 (c) 15:35 (d) 16:15

(e) 14:35 (f) 15:15 (g) 15:35 (h) 16:15

Figure 2: (a) to (d): TSLA stock. (e) to (h): GOOGL stock. We predictions of traded volume in next 100 days, RM(yellow) fails to capture the regularities of traded volume in time series, kalman filter based model(green) fails to capture noise information and make reasonable predictions, our model(blue) is able to seize the movements of traded volume and yield better predictions.

Table 2: The Mean absolute percentage error(MAPE) of different models on Daily Trading Volume data sets.

Stock	Time	RM	KF	Ours
JPM	14:35	0.52	0.28	**0.01**
	15:15	0.54	0.36	**0.04**
	15:35	0.51	0.42	**0.06**
	16:15	0.52	0.49	**0.12**
TSLA	14:35	0.53	0.31	**0.02**
	15:15	0.55	0.36	**0.03**
	15:35	0.53	0.39	**0.08**
	16:15	0.52	0.38	**0.14**
GOOGL	14:35	0.49	0.35	**0.01**
	15:15	0.51	0.38	**0.03**
	15:35	0.53	0.44	**0.05**
	16:15	0.51	0.42	**0.11**

B. Definition of G in Synthetic-2

Synthetic-2 (Nonlinear, converging to mixed-Gaussian):

$$\hat{\boldsymbol{x}}_0 \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0), \quad \hat{\boldsymbol{x}}_{t+\Delta t} = \hat{\boldsymbol{x}}_t - \boldsymbol{G}\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1)$$

$$\boldsymbol{G}_{11} = \frac{1}{\sigma_1}\frac{N_1}{N_1+N_2}(\hat{x}_t^1 - \mu_{11}) + \frac{1}{\sigma_2}\frac{N_2}{N_1+N_2}(\hat{x}_t^1 - \mu_{21})$$
$$\boldsymbol{G}_{22} = \frac{1}{\sigma_1}\frac{N_1}{N_1+N_2}(\hat{x}_t^2 - \mu_{12}) + \frac{1}{\sigma_2}\frac{N_2}{N_1+N_2}(\hat{x}_t^2 - \mu_{22})$$
$$N_1 = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(\hat{x}_t^1 - \mu_{11})^2}{2\sigma_1^2} - \frac{(\hat{x}_t^1 - \mu_{12})^2}{2\sigma_1^2}\right)$$
$$N_2 = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(\hat{x}_t^2 - \mu_{21})^2}{2\sigma_2^2} - \frac{(\hat{x}_t^2 - \mu_{22})^2}{2\sigma_2^2}\right)$$

C. Error Analysis

In this section, we provide an error analysis of our model. Suppose the hidden dynamics is driven by $g_r(\boldsymbol{x})$, the dynamics that we learn from data is $g_f(\boldsymbol{x})$, then original Itô process, Euler processes computed by true g_r and estimated g_f are:

$$d\boldsymbol{X} = g(\boldsymbol{X})dt + \sigma d\boldsymbol{W}$$
$$\boldsymbol{x}^r_{t+\Delta t} = \boldsymbol{x}^r_t + g_r(\boldsymbol{x}^r_t)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1)$$
$$\boldsymbol{x}^f_{t+\Delta t} = \boldsymbol{x}^f_t + g_f(\boldsymbol{x}^f_t)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1)$$

where $\boldsymbol{X}$ is the ground truth, $\boldsymbol{x}^r$ is computed by true g_r and $\boldsymbol{x}^f$ is computed by estimated g_f. Estimating the error between original Itô process and its Euler form can be very complex, hence we cite the conclusion from (Milstein & Tretyakov, 2013) and focus more on the error between original form and our model.

Lemma 2. *With the same initial $\boldsymbol{X}_{t_0} = \boldsymbol{x}_{t_0} = \boldsymbol{x}_0$, if there is a global Lipschitz constant K which satisfies:*

$$|g(\boldsymbol{x}, t) - g(\boldsymbol{y}, t)| \le K|\boldsymbol{x} - \boldsymbol{y}|$$

then after n steps, the expectation error between Itô process $\boldsymbol{x}_{t_n}$ and Euler forward process $\boldsymbol{x}^r_{t_n}$ is:

$$\mathbb{E}|\boldsymbol{x}_{t_n} - \boldsymbol{x}^r_{t_n}| \le K\left(1 + \mathbb{E}|X_0|^2\right)^{1/2}\Delta t$$

Lemma 2 illustrates that the expectation error between original Itô process and its Euler form is not related to total steps n but time step Δt.

Proposition 3. *With the same initial $\boldsymbol{x}_0$, suppose the generalization error of neural network g is ε and existence of global Lipschitz constant K:*

$$|g(\boldsymbol{x}) - g(\boldsymbol{y})| \leq K|\boldsymbol{x} - \boldsymbol{y}|$$

then after n steps with step size $\Delta t = T/n$, the expectation error between Itô process $\boldsymbol{x}_{t_n}$ and approximated forward process $\boldsymbol{x}^f_{t_n}$ is bounded by:

$$\mathbb{E}|\boldsymbol{x}_{t_n} - \boldsymbol{x}^f_{t_n}| \leq \frac{\varepsilon}{K}(e^{KT} - 1) + K(1 + \mathbb{E}|\boldsymbol{x}_0|^2)^{1/2}\Delta t \tag{1}$$

Proposition 3 implies that besides time step size Δt, our expectation error interacts with three factors, generalization error, Lipschitz constant of g and total time length. In our experiments, we find the best way to decrease the expectation error is reducing the value of K and n.

D. Proofs

D.1. Proof of Proposition 1

Proof. Suppose $\hat{x}^{(k)}_{t_m}$ and $\hat{x}^{(k)}_{t_{m-1}}$ are our observed samples at t_m and t_{m-1} respectively, then expectations could be approximated by:

$$\mathbb{E}_{\boldsymbol{x}\sim\hat{p}(\boldsymbol{x},t_m)}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})\hat{p}(\boldsymbol{x},t_m)d\boldsymbol{x} \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}^{(k)}_{t_m}) \tag{2}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x},t_m)}[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})\tilde{p}(\boldsymbol{x},t_m)d\boldsymbol{x} = \int f(\boldsymbol{x})\left[\hat{p}(\boldsymbol{x},t_{m-1}) + \int_{t_{m-1}}^{t_m}\frac{\partial p(\boldsymbol{x},\tau)}{\partial t}d\tau\right]d\boldsymbol{x} \\
&= \int f(\boldsymbol{x})\hat{p}(\boldsymbol{x},t_{m-1})d\boldsymbol{x} + \int f(\boldsymbol{x})\int_{t_{m-1}}^{t_m}\frac{\partial p(\boldsymbol{x},\tau)}{\partial t}d\tau d\boldsymbol{x} \\
&\approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{\boldsymbol{x}}^{(k)}_{t_{m-1}}) + \underbrace{\int f(\boldsymbol{x})\int_{t_{m-1}}^{t_m}\left\{-\sum_{i=1}^{D}\frac{\partial}{\partial x_i}\left[g^i_\omega(\boldsymbol{x})p(\boldsymbol{x},\tau)\right] + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}p(\boldsymbol{x},\tau)\right\}d\tau d\boldsymbol{x}}_{I}
\end{aligned} \tag{3}$$

Then for the second term I above, it is difficult to calculate directly, but we can use integration by parts to rewrite I as:

$$\begin{aligned}
I &= \int_{t_{m-1}}^{t_m}\int\left[\sum_{i=1}^{D} -f(\boldsymbol{x})\frac{\partial}{\partial x_i}g^i_\omega(\boldsymbol{x})p(\boldsymbol{x},\tau) + \frac{1}{2}\sigma^2\sum_{i=1}^{D} f(\boldsymbol{x})\frac{\partial^2}{\partial x_i^2}p(\boldsymbol{x},\tau)\right]d\boldsymbol{x}d\tau \\
&= \int_{t_{m-1}}^{t_m}\int\left[\sum_{i=1}^{D} g^i_\omega(\boldsymbol{x})p(\boldsymbol{x},\tau)\frac{\partial}{\partial x_i}f(\boldsymbol{x}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D} p(\boldsymbol{x},\tau)\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})\right]d\boldsymbol{x}d\tau \\
&= \int_{t_{m-1}}^{t_m}\left(\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x},\tau)}\left[\sum_{i=1}^{D} g^i_\omega(\boldsymbol{x})\frac{\partial}{\partial x_i}f(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x},\tau)}\left[\frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})\right]\right)d\tau \\
&\approx \int_{t_{m-1}}^{t_m}\frac{1}{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{D} g^i_\omega(x^{(k)})\frac{\partial}{\partial x_i}f(x^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}f(x^{(k)})\right)d\tau
\end{aligned} \tag{4}$$

To approximate the integral from t_{m-1} to t_m, we adopt trapezoid rule, then we could rewrite the expectation in Equation (3) as:

$$\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x},t_m)}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_{m-1}}^{(k)}) + \frac{\Delta t}{2}\Bigg[\frac{1}{N}\sum_{k=1}^{N}\Bigg(\sum_{i=1}^{D} g_{\omega}^{i}(\hat{x}_{t_{m-1}}^{(k)})\frac{\partial}{\partial x_i}f(\hat{x}_{t_{m-1}}^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}f(\hat{x}_{t_{m-1}}^{(k)})\Bigg)$$
$$+ \frac{1}{N}\sum_{k=1}^{N}\Bigg(\sum_{i=1}^{D} g_{\omega}^{i}(\tilde{x}_{t_m}^{(k)})\frac{\partial}{\partial x}f(\tilde{x}_{t_m}^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}f(\tilde{x}_{t_m}^{(k)})\Bigg)\Bigg]$$
$$= \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_{m-1}}^{(k)}) + \frac{\Delta t}{2}\left[\mathcal{F}_f(\hat{X}_{m-1}) + \mathcal{F}_f(\tilde{X}_m)\right] \tag{5}$$

We subtract (2) by (5) to finish the proof. □

D.2. Proof of Proposition 2

Proof. Given initial $\hat{\boldsymbol{x}}_{t_0}$, we generate $\tilde{\boldsymbol{x}}_{t_1}, \tilde{\boldsymbol{x}}_{t_2}, \tilde{\boldsymbol{x}}_{t_3}$ ... $\tilde{\boldsymbol{x}}_{t_n}$ sequentially by Euler-Maruyama scheme. Then the expectations can be rewritten as:

$$\mathbb{E}_{\boldsymbol{x}\sim\hat{p}(\boldsymbol{x},t_n)}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})\hat{p}(\boldsymbol{x},t_n)d\boldsymbol{x} \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_n}^{(k)}) \tag{6}$$

$$\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x},t_n)}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \int_{t_0}^{t_1}\frac{1}{N}\sum_{k=1}^{N}\left[\sum_{i=1}^{D} g_{\omega}^{i}(x^{(k)})\frac{\partial}{\partial x_i}f(x^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i\partial x_j}f(x^{(k)})\right]d\tau$$
$$+ \int_{t_1}^{t_2}\frac{1}{N}\sum_{k=1}^{N}\left[\sum_{i=1}^{D} g_{\omega}^{i}(x^{(k)})\frac{\partial}{\partial x_i}f(x^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{D}\frac{\partial^2}{\partial x_i^2}f(x^{(k)})\right]d\tau + ...$$
$$+ \int_{t_{n-1}}^{t_n}\frac{1}{N}\sum_{k=1}^{N}\left[\sum_{i=1}^{n} g_{\omega}^{i}(x^{(k)})\frac{\partial}{\partial x_i}f(x^{(k)}) + \frac{1}{2}\sigma^2\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2}f(x^{(k)})\right]d\tau \tag{7}$$

which is:

$$\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x},t_n)}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \frac{\Delta t}{2}\left[\mathcal{F}_f(\hat{X}_0) + \mathcal{F}_f(\tilde{X}_1)\right] + \frac{\Delta t}{2}\left[\mathcal{F}_f(\tilde{X}_1) + \mathcal{F}_f(\tilde{X}_2)\right] + ...$$
$$+ \frac{\Delta t}{2}\left[\mathcal{F}_f(\tilde{X}_{n-1}) + \mathcal{F}_f(\tilde{X}_n)\right] \tag{8}$$

Finally it comes to:

$$\mathbb{E}_{\boldsymbol{x}\sim\tilde{p}(\boldsymbol{x},t_n)}[f(\boldsymbol{x})] \approx \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \frac{\Delta t}{2}\left(\mathcal{F}_f(\hat{X}_0) + \mathcal{F}_f(\tilde{X}_n) + 2\sum_{s=1}^{n-1}\mathcal{F}_f(\tilde{X}_s)\right) \tag{9}$$

We subtract (6) by (9) to finish the proof. □

D.3. Proof of Error Analysis

Proof. The proof process of Lemma 2 is quite long and out of the scope of this paper, for more details please see first two chapters in reference book (Milstein & Tretyakov, 2013). While for the proof of Proposition 3, with initial X and first one-step iteration:

$$\begin{cases} \boldsymbol{x}_{t_0}^{r} = \boldsymbol{x}_{t_0} \\ \boldsymbol{x}_{t_0}^{f} = \boldsymbol{x}_{t_0} \end{cases} \tag{10}$$

$$\begin{cases} \boldsymbol{x}_{t_1}^{r} = \boldsymbol{x}_{t_0}^{r} + g_r(\boldsymbol{x}_{t_0}^{r})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \\ \boldsymbol{x}_{t_1}^{f} = \boldsymbol{x}_{t_0}^{f} + g_f(\boldsymbol{x}_{t_0}^{f})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \end{cases} \tag{11}$$

Then we have:

$$\mathbb{E}|\boldsymbol{x}^r_{t_0} - \boldsymbol{x}^f_{t_0}| = \mathbb{E}|\boldsymbol{x}_{t_0} - \boldsymbol{x}_{t_0}| = 0 \tag{12}$$

$$\begin{aligned}
\mathbb{E}|\boldsymbol{x}^r_{t_1} - \boldsymbol{x}^f_{t_1}| &= \mathbb{E}|\boldsymbol{x}^r_{t_0} - \boldsymbol{x}^f_{t_0} + g_r(\boldsymbol{x}^r_{t_0})\Delta t - g_f(\boldsymbol{x}^f_{t_0})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) - \sigma\sqrt{\Delta t}\mathcal{N}(0,1)| \\
&\leq \mathbb{E}|\boldsymbol{x}^r_{t_0} - \boldsymbol{x}^f_{t_0}| + \mathbb{E}|g_r(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^f_{t_0})|\Delta t \\
&= \mathbb{E}|g_r(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^r_{t_0}) + g_f(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^f_{t_0})|\Delta t \\
&\leq \mathbb{E}|g_r(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^r_{t_0})|\Delta t + \mathbb{E}|g_f(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^f_{t_0})|\Delta t \\
&\leq \varepsilon\Delta t + \mathbb{E}|g_f(\boldsymbol{x}^r_{t_0}) - g_f(\boldsymbol{x}^f_{t_0})|\Delta t \\
&= \varepsilon\Delta t + \mathbb{E}|g'_f(\boldsymbol{x}^\xi_{t_0})(\boldsymbol{x}^r_{t_0} - \boldsymbol{x}^f_{t_0})|\Delta t \qquad (\boldsymbol{x}^\xi_{t_0} \in [\boldsymbol{x}^r_{t_0}, \boldsymbol{x}^f_{t_0}]) \\
&\leq \varepsilon\Delta t + K\mathbb{E}|\boldsymbol{x}^r_{t_0} - \boldsymbol{x}^f_{t_0}|\Delta t \\
&= \varepsilon\Delta t
\end{aligned} \tag{13}$$

Follow the pattern we have:

$$\begin{cases} \boldsymbol{x}^r_{t_2} = \boldsymbol{x}^r_{t_1} + g_r(\boldsymbol{x}^r_{t_1})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \\ \boldsymbol{x}^f_{t_2} = \boldsymbol{x}^f_{t_1} + g_f(\boldsymbol{x}^f_{t_1})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \end{cases} \tag{14}$$

...

$$\begin{cases} \boldsymbol{x}^r_{t_n} = \boldsymbol{x}^r_{t_{n-1}} + g_r(\boldsymbol{x}^r_{t_{n-1}})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \\ \boldsymbol{x}^f_{t_n} = \boldsymbol{x}^f_{t_{n-1}} + g_f(\boldsymbol{x}^f_{t_{n-1}})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) \end{cases} \tag{15}$$

Which leads to:

$$\begin{aligned}
\mathbb{E}|\boldsymbol{x}^r_{t_2} - \boldsymbol{x}^f_{t_2}| &= \mathbb{E}|\boldsymbol{x}^r_{t_1} - \boldsymbol{x}^f_{t_1} + g_r(\boldsymbol{x}^r_{t_1})\Delta t - g_f(\boldsymbol{x}^f_{t_1})\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0,1) - \sigma\sqrt{\Delta t}\mathcal{N}(0,1)| \\
&\leq \mathbb{E}|\boldsymbol{x}^r_{t_1} - \boldsymbol{x}^f_{t_1}| + \mathbb{E}|g_r(\boldsymbol{x}^r_{t_1}) - g_f(\boldsymbol{x}^f_{t_1})|\Delta t \\
&\leq \mathbb{E}|\boldsymbol{x}^r_{t_1} - \boldsymbol{x}^f_{t_1}| + \varepsilon\Delta t + K\mathbb{E}|\boldsymbol{x}^r_{t_1} - \boldsymbol{x}^f_{t_1}|\Delta t \\
&\leq (1 + K\Delta t)\varepsilon\Delta t + \varepsilon\Delta t
\end{aligned} \tag{16}$$

...

$$\mathbb{E}|\boldsymbol{x}^r_{t_n} - \boldsymbol{x}^f_{t_n}| \leq \varepsilon\Delta t \sum_{i=0}^{n-1}(1 + K\Delta t)^i \tag{17}$$

Now let $S = \sum_{i=0}^{n-1}(1 + K\Delta t)^i$, then consider followings:

$$\begin{aligned}
S(K\Delta t) &= S(1 + K\Delta t) - S \\
&= \sum_{i=1}^{n}(1 + K\Delta t)^i - \sum_{i=0}^{n-1}(1 + K\Delta t)^i \\
&= (1 + K\Delta t)^n - 1 \\
&= (1 + K\frac{T}{n})^n - 1 \\
&\leq e^{KT} - 1
\end{aligned} \tag{18}$$

Finally we have:

$$\mathbb{E}|\boldsymbol{x}^r_{t_n} - \boldsymbol{x}^f_{t_n}| \leq \frac{\varepsilon}{K}(e^{KT} - 1) \tag{19}$$

$$\mathbb{E}|\boldsymbol{x}_{t_n} - \boldsymbol{x}^f_{t_n}| \leq \frac{\varepsilon}{K}(e^{KT} - 1) + K(1 + E|\boldsymbol{x}_0|^2)^{1/2}\Delta t \tag{20}$$

□

References

Milstein, G. N. and Tretyakov, M. V. (eds.). *Stochastic numerics for mathematical physics*. Springer Science & Business Media, 2013.