APPENDIX

Anonymous authors
Paper under double-blind review

A RELATED WORK

Influence of Normalization on Optimization Given the straightforward way to obtain the gradient of the normalized model, a number of works focus on how normalization affect the learning dynamics of the model during training, and explore the mechanism from different aspects. Bjorck et al. (2018); Santurkar et al. (2018) claim normalization can improve optimization by smoothing the loss landscape; Yang et al. (2018) interprets the benefit of normalization by mean field theory; Kohler et al. (2019); Cai et al. (2019); Wu et al. (2020) derive the convergence rate of gradient descent algorithm in normalized quadratic model, which can be seen as variants of Rayleigh Quotient, but they do not take into account weight decay and stochastic gradient; Dukler et al. (2020) gives the convergence result for the two-layer rule network with weight normalization (Salimans & Kingma, 2016). Based the spherical perspective of normalized weight, Hoffer et al. (2018); Wu et al. (2018b); Arora et al. (2018) regard gradient descent algorithm with normalization as an adaptive step-size algorithm. Some works acknowledge the joint effect of normalization and WD in SGD: Li & Arora (2019) find with normalization, SGD with WD and constant learning rate is equivalent to SGD with increasing learning rate algorithm but no WD. Van Laarhoven (2017); Chiley et al. (2019); Kunin et al. (2021) discuss the concept of "equilibrium" of normalization and WD; Li et al. (2020); Wan et al. (2021) theoretically justify the existence of equilibrium, Wan et al. (2021) further proposes the concept of "spherical motion dynamics", and empirically shows its influence to learning dynamics of the model.

Evolution Dynamics of SGD Beyond the convergence of SGD, many works focus on interpreting the possible benefit of gradient noise to improve the generalization of the model. Zhang et al. (2019) discuss the effect of batch size and momentum on SGD in a quadratic model; A more popular way is approximating SGD as SDE (Li et al., 2019; 2021), then studying the evolution of SDE instead. Hu et al. (2017); Jastrzebski et al. (2017); Wu et al. (2018a); Zhu et al. (2019); Xie et al. (2020) use diffusion model to characterize escaping the behavior of SGD; Liu et al. (2021); Kunin et al. (2021) discuss the limiting dynamics of quadratic model by deriving its stationary distribution.

B PRELIMINARIES

In this paper we relate the update of SGD to the discretized simulation of stochastic differential equations (SDEs). Finding the explicit solution of an SDE is the most direct way to analyze its dynamics, but is notoriously intractable in general cases. We adopt in this paper two alternative treatments which we illustrate as follows.

B.1 FROM MARTINGALE PROPERTY TO EXPECTATION BOUNDS

Lemma B.1.1 (Itô formula)**.** *Let*

$$d\boldsymbol{X}_t = \boldsymbol{A}(\boldsymbol{X}_t, t)dt + \boldsymbol{B}(\boldsymbol{X}_t, t)d\boldsymbol{B}_t \tag{1}$$

be an n-dimensional Itô process. Let $\boldsymbol{F}(t, \boldsymbol{x})$ be a $\mathcal{C}^2$ map from $[0, \infty) \times \mathbb{R}^n$ to $\mathbb{R}^m$. Then the process

$$\boldsymbol{Y}(t, \omega) = \boldsymbol{F}(t, \boldsymbol{X}_t) \tag{2}$$

is again an Itô process given by

$$d\boldsymbol{Y}_{i,t} = \partial_t \boldsymbol{F}_i dt + (\nabla_{\boldsymbol{x}} \boldsymbol{F}_i)^T \cdot d\boldsymbol{X}_t + \frac{1}{2} tr[\boldsymbol{B}(\nabla^2_{\boldsymbol{xx}} \boldsymbol{F}_i)\boldsymbol{B}^T]dt, \quad i = 1, 2, \cdots, m \tag{3}$$

Definition B.1.2. *Given a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a filtration* $\{\mathcal{F}_t | \, t \in [0, \infty)\}$ *on it, let* $\mathcal{V}(0, T)$ *be the class of functions* $f(t, \omega) : \, [0, \infty) \times \Omega \to \mathbb{R}$ *such that*
1). f *is* $\mathcal{B}([0, \infty)) \times \mathcal{F}$*-measurable.*
2). $f(t, \omega)$ *is* $\mathcal{F}_t$*-adapted.*
3). $\mathbb{E}\left[\int_0^T f(t, \omega)^2 dt\right] < \infty$.

Lemma B.1.3 (Martingale property of Itô integral)**.** *Given* $f \in \mathcal{V}(0, T)$*, the Itô integral*

$$I_t = \int_0^t f(\tau, \omega) dB_\tau, \quad 0 \le t \le T \tag{4}$$

is an $\mathcal{F}_t$*-martingale, which means*

$$\mathbb{E}\left[\int_s^t f(\tau, \omega) dB_\tau \middle| \mathcal{F}_s\right] = 0, \quad \forall \; 0 \le s \le t \le T \tag{5}$$

and consequently

$$\mathbb{E}\left[\int_s^t f(\tau, \omega) dB_\tau\right] = \mathbb{E}\left\{\mathbb{E}\left[\int_s^t f(\tau, \omega) dB_\tau \middle| \mathcal{F}_s\right]\right\} = 0 \tag{6}$$

Similar results hold in multivariate cases.

Suppose we are given an Itô process $d\boldsymbol{Y}_t = \boldsymbol{C}(\boldsymbol{Y}_t, t)dt + \boldsymbol{D}(\boldsymbol{Y}_t, t)d\boldsymbol{B}_t$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which means

$$\boldsymbol{Y}_t - \boldsymbol{Y}_0 = \int_0^t \boldsymbol{C}(\boldsymbol{Y}_s, s)ds + \int_0^t \boldsymbol{D}(\boldsymbol{Y}_s, s)d\boldsymbol{B}_s \tag{7}$$

We set $\mathcal{F}_t = \sigma(\boldsymbol{Y}_t)$ and take expectation on both sides of equation 7. By Lemma B.1.3 and with sufficient regularity constraints, we can cancel off the diffusion term and exchange the order of integral and expectation to obtain

$$\mathbb{E}\boldsymbol{Y}_t - \mathbb{E}\boldsymbol{Y}_0 = \mathbb{E}\left[\int_0^t \boldsymbol{C}(\boldsymbol{Y}_s, s)ds\right] = \int_0^t \mathbb{E}\left[\boldsymbol{C}(\boldsymbol{Y}_s, s)\right] ds \tag{8}$$

When the drift term $\boldsymbol{C}(\boldsymbol{Y}_t, t)$ is not linear in $\boldsymbol{Y}_t$, it is generally impossible to solve $\mathbb{E}\boldsymbol{Y}_t$ explicitly from the integral equation equation 8, while upper and lower bounds for $\mathbb{E}\boldsymbol{Y}_t$ may still be given.
Following the above preliminaries, in this paper we first apply Itô formula to transform the original dynamics of the NRQ weight $\boldsymbol{X}_t$ to the dynamics of other quantities we are interested in, such as $\tilde{\boldsymbol{X}}_t = \frac{\boldsymbol{X}_t}{||\boldsymbol{X}_t||}$, $M_t = ||\boldsymbol{X}_t||^2$ and $f_t = (\boldsymbol{e}_1^T \tilde{\boldsymbol{X}}_t)^2$, and then derive bounds for their expected value based on twe procedures in equation 7 $\sim$ equation 8.
Taking advantage of the martingale property of Itô integral, We first prove Lemma 1 of the main paper.

Lemma B.1.4. *In the evolution of*

$$d\tilde{\boldsymbol{X}}_t = -\,[\frac{\eta}{M_t}\boldsymbol{P}_t\boldsymbol{A}\tilde{\boldsymbol{X}}_t + \frac{\eta^2}{2M_t^2}\,\mathrm{Tr}(\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}}\boldsymbol{P}_t)\tilde{\boldsymbol{X}}_t]dt - \frac{\eta}{M_t}\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}d\boldsymbol{B}_t \tag{9}$$

$$dM_t = [-2\lambda\eta M_t + \frac{\eta^2}{M_t}\,\mathrm{Tr}(\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}})]dt \tag{10}$$

, we have

$$\Delta_t^2 = \frac{\mathrm{Tr}(\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}})\eta^2}{M_t^2}. \tag{11}$$

If $\mathrm{Tr}(\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}})$ *is constant, then*

$$\lim_{t\to\infty} \Delta_t = \sqrt{2\lambda\eta}. \tag{12}$$

Proof. Using Itô lemma, for given $T > 0$, and equation 9, we have

$$d||\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T||_2^2 = [-\frac{2\eta}{M_t}(\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T)^T \boldsymbol{P}_t \boldsymbol{A} \tilde{\boldsymbol{X}}_t - \frac{2\eta^2}{M_t^2}(\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T)^T \tilde{\boldsymbol{X}}_t \operatorname{Tr}(\boldsymbol{P}_t \tilde{\boldsymbol{\Sigma}}) + \frac{\eta^2}{M_t^2} \operatorname{Tr}(\boldsymbol{P}_t \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_t)]dt + \frac{\eta}{M_t}(\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T)^T \boldsymbol{P}_t \tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} d\boldsymbol{B}_t \tag{13}$$

Then by Lemma B.1.3, we have

$$\begin{aligned}\mathbb{E}_T||\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T||_2^2 =& \mathbb{E}_T \int_T^t [-\frac{2\eta}{M_\tau}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \boldsymbol{P}_\tau \boldsymbol{A} \tilde{\boldsymbol{X}}_\tau - \frac{2\eta^2}{M_\tau^2}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \tilde{\boldsymbol{X}}_\tau \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}}) + \frac{\eta^2}{M_\tau^2} \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)]d\tau \\ =& \int_T^t \mathbb{E}_t[-\frac{2\eta}{M_\tau}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \boldsymbol{P}_\tau \boldsymbol{A} \tilde{\boldsymbol{X}}_\tau]d\tau + \int_T^t \mathbb{E}_t[-\frac{2\eta^2}{M_\tau^2}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \tilde{\boldsymbol{X}}_\tau \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}})]d\tau \\ &+ \mathbb{E}_t \int_T^t \mathbb{E}_t[\frac{\eta^2}{M_\tau^2} \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)]d\tau\end{aligned} \tag{14}$$

Since we have the following results based on Lemma B.1.3

$$\lim_{\tau \to T} \mathbb{E}_T[-\frac{2\eta}{M_\tau}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \boldsymbol{P}_\tau \boldsymbol{A} \tilde{\boldsymbol{X}}_\tau] = 0 \tag{15}$$

$$\lim_{\tau \to T} \mathbb{E}_T[-\frac{2\eta^2}{M_\tau^2}(\tilde{\boldsymbol{X}}_\tau - \tilde{\boldsymbol{X}}_T)^T \tilde{\boldsymbol{X}}_\tau \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}})] = 0 \tag{16}$$

$$\lim_{\tau \to T} \mathbb{E}_t[\frac{\eta^2}{M_\tau^2} \operatorname{Tr}(\boldsymbol{P}_\tau \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)] = \frac{\eta^2}{M_T^2} \operatorname{Tr}(\tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau) \tag{17}$$

Hence we have

$$\lim_{t \to T} \frac{\mathbb{E}_T||\tilde{\boldsymbol{X}}_t - \tilde{\boldsymbol{X}}_T||_2^2}{t - T} = \frac{\eta^2}{M_T^2} \operatorname{Tr}(\tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau) \tag{18}$$

i.e.

$$\Delta_T^2 = \frac{\operatorname{Tr}(\tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)\eta^2}{M_T^2}. \tag{19}$$

When $\operatorname{Tr}(\tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)$ is constant, then by equation 12 in main text, we have

$$\lim_{t \to \infty} M_t^2 = \eta \frac{\operatorname{Tr}(\tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_\tau)}{2\lambda} \tag{20}$$

Combining equation 19 and equation 20, we have

$$\lim_{t \to \infty} \Delta_t = \sqrt{2\lambda\eta} \tag{21}$$

□

B.2 Fokker-Planck equations provide further statistics

In Appendix B.1 we illustrate how to bound the expectation of an Itô process we are interested in. Sometimes with mere expectation information it is not enough to tell anything further. Theoretically, all of the statistics of an Itô process can be told from the evolution of its distribution, which is depicted by the Fokker-Planck equation.

Lemma B.2.1 (Fokker-Planck equation)**.** *Consider an Itô process $\boldsymbol{X}_t$ defined as the solution of*

$$d\boldsymbol{X}_t = \boldsymbol{A}(\boldsymbol{X}_t, t)dt + \boldsymbol{B}(\boldsymbol{X}_t, t)d\boldsymbol{B}_t \tag{22}$$

Denote the density function of $\boldsymbol{X}_t$ by $\rho(\boldsymbol{x}, t)$. Then $\rho(\boldsymbol{x}, t)$ is given by the Fokker-Planck equation

$$\partial_t \rho = -\sum_{i=1}^p \partial_{x_i}[\rho \boldsymbol{A}(\boldsymbol{x}, t)] + \frac{1}{2} \sum_{i,j=1}^p \partial_{x_t x_j}^2 [\boldsymbol{B}(\boldsymbol{x}, t)\rho] \tag{23}$$

$$\triangleq -\nabla \cdot \boldsymbol{J}(\boldsymbol{x}, t) \tag{24}$$

where $\boldsymbol{J}_i(x,t) = \rho \boldsymbol{A}_i(\boldsymbol{x},t) - \frac{1}{2}\sum_{j=1}^{p} \partial_{x_j}[\boldsymbol{B}_{ij}(\boldsymbol{x},t)\rho]$ *is the probability current generated by the Itô process* $\boldsymbol{X}_t$.

Similar to solving an SDE, finding the explicit solution to a Fokker-Planck equation is usually an intractable task. When dealing with Fokker-Planck equation in this paper, we consider a simplified case under the following stronger condition.

Assumption B.2.2. *Under the setting of Corollary D.1.1, we assume additionally that that the spectrum of matrix* $\boldsymbol{A}$ *takes only 2 distinct real values* $a_1 = a_l < a_h = a_2 = a_3 = \cdots = a_p$.

Intuitively, the above assumption states that when solving the Fokker-Planck equation, we neglect the anisotropy of landscape along different geodesics on the unit sphere $\mathbb{S}^{p-1}$ spreading out from the polar $\boldsymbol{e}_1$.

C Proof of Theorem 1

Lemma C.0.1. $\forall T > 0$, $\mathbb{E}f_t$ *is continuous on* $[0,T]$.

Proof. This is a direct corollary of lemma B.1.3. □

Lemma C.0.2. *Given a positive number* $T > 0$, $\forall K \in (0,+\infty)$, *define* $h_K(t)$ *as*

$$h_K(t) = \frac{e^{Kt}}{\int_0^T e^{K\tau}d\tau}. \tag{25}$$

Then $\forall f(t) \in \mathcal{C}[0,T]$, *we have*

$$\lim_{K\to+\infty} \int_0^T h_K(t)f(t)dt = f(T). \tag{26}$$

Proof. Since $f(t)$ is continuous in $[0,T]$, then $f(t)$ is bounded on $[0,T]$, assume $\exists M > 0, \forall t \in [0,1]$, $|f(t)| < M$. Besides, since $f(t)$ is continuous on T, $\forall \varepsilon > 0$, $\exists \delta > 0$, when $|t-T| < \delta$, $|f(t) - f(T)| < \varepsilon$. Then $\forall K > \frac{1}{\delta}\log\frac{2M}{\varepsilon}$, we have:

$$\begin{aligned}
\int_0^T h_K(t)f(t)dt - f(T) &= \int_0^T h_K(t)(f(t)-f(T))dt \\
&= \int_0^(T-\delta) h_K(t)(f(t)-f(T))dt + \int_{T-\delta}^T h_K(t)(f(t)-f(T))dt \\
&\leq 2M\int_0^(T-\delta) h_K(t)dt + \varepsilon\int_{T-\delta}^T h_K(t)dt \\
&\leq 2M\frac{e^{K(T-\delta)}-1}{e^{KT}-1} + \delta \\
&\leq \frac{2M}{e^{K\delta}} + \varepsilon \\
&< 2\varepsilon
\end{aligned} \tag{27}$$

□

With the strategies introduced in Appendix B.1, we are now able to provide non-asymptotic bounds of $\mathbb{E}f_t$.

Theorem C.0.3 (Variants of Theorem 1 in main text)**.** *Then* $\forall t > 0$ *we have*

$$\mathbb{E}f_t \leq e^{-G_1(t)}[f_0 + \int_0^t g_2(\tau)e^{G_1(\tau)}d\tau], \tag{28}$$

where

$$M(t)^2 \triangleq \frac{\eta(p-1)\sigma^2}{2\lambda} + e^{-4\lambda\eta t}(M_0^2 - \frac{\eta(p-1)\sigma^2}{2\lambda}) = M_t^2; \tag{29}$$

$$g_1(t) \triangleq \frac{(a_p - a_1)\eta}{M(t)} + \frac{p\eta^2\sigma^2}{M(t)^2}; \tag{30}$$

$$g_2(t) \triangleq \frac{(a_p - a_1)\eta}{M(t)} + \frac{\eta^2\sigma^2}{M(t)^2}; \tag{31}$$

$$G_1(t) \triangleq \int_0^t g_1(\tau)d\tau. \tag{32}$$

Further assume $\exists \xi \in (\frac{1}{2}, 1)$, $\varepsilon(t) \triangleq P(f_t < \xi)$, *then* $\forall T > 0$, *we have*

$$\mathbb{E}f_t \geq e^{-\tilde{G}_1(t)}[f_0 + \int_0^t \tilde{g}_2(\tau)e^{\tilde{G}_1(\tau)}d\tau], \tag{33}$$

where

$$\tilde{g}_1(t) \triangleq \frac{(a_2 - a_1)\eta\xi}{M(t)} + \frac{p\eta^2\sigma^2}{M(t)^2}; \tag{34}$$

$$\tilde{g}_2(t) \triangleq \frac{(a_2 - a_1)\eta\xi}{M(t)}(1 - \varepsilon(t)) + \frac{\eta^2\sigma^2}{M(t)^2}; \tag{35}$$

$$\tilde{G}_1(t) \triangleq \int_0^t \tilde{g}_1(\tau)d\tau. \tag{36}$$

Proof. When $\tilde{\boldsymbol{\Sigma}} = \sigma^2\boldsymbol{I}$, M_t^2 is deterministic given t:

$$\begin{aligned} M_t^2 =& e^{-4\lambda\eta t}M_0^2 + 2\eta^2\int_0^t e^{-4\lambda\eta(t-\tau)}\operatorname{Tr}(\boldsymbol{P}_\tau\tilde{\boldsymbol{\Sigma}})d\tau, \\ =& e^{-4\lambda\eta t}M_0^2 + 2\eta^2\int_0^t e^{-4\lambda\eta(t-\tau)}(p-1)\sigma^2 d\tau, \\ =& \frac{\eta(p-1)\sigma^2}{2\lambda} + e^{-4\lambda\eta t}(M_0^2 - \frac{\eta(p-1)\sigma^2}{2\lambda}), \end{aligned} \tag{37}$$

therefore let $M(t)$ denote M_t when it is deterministic. It's obvious $M(t)$ is positive, monotonous, and continuous on $(0, +\infty)$, and converges to $M^* \triangleq \sqrt{\frac{\eta(p-1)\sigma^2}{2\lambda}}$. Hence, $g_1(t), g_2(t)$ are also positive, monotonous, and continuous on $(0, +\infty)$, and converges to $g_1^* \triangleq \frac{a_p - a_1}{M^*} + \frac{p\eta^2\sigma^2}{(M^*)^2}$, $g_2^* \triangleq \frac{a_p - a_1}{M^*} + \frac{\eta^2\sigma^2}{(M^*)^2}$ respectively.

(16) in main text can be rewritten as

$$df_t = \{\frac{\eta\delta_t}{M(t)}(1 - f_t)f_t + \frac{\eta^2}{M(t)^2}[\sigma^2 - p\sigma^2 f_t]\}dt + \frac{2\eta}{M(t)}\sqrt{f_t(1 - f_t)}dB_t. \tag{38}$$

where

$$\delta_t = \frac{L_t - a_1}{1 - f_t} = \frac{\sum_{i=2}^p (a_i - a_1)(\boldsymbol{X}_t^{(i)})^2}{\sum_{i=2}^p (\boldsymbol{X}_t^{(i)})^2}. \tag{39}$$

Obviously $\delta_t \in [a_2 - a_1, a_p - a_1]$.

Given an positve real number $\mathcal{T} \in \mathbb{R}^+$, arbitarily choose a positive continous-differential function $\mu(t) \in \mathcal{C}^1[0, \mathcal{T}]$, set $\hat{f}_t \triangleq \mu(t)f_t$, then by Itô lemma, we can derive the evolution of $\hat{f}_t$ based on equation 38, which is

$$d\hat{f}_t = \{\frac{\mu'(t)}{\mu(t)}\hat{f}_t + \frac{\eta\delta_t}{M(t)}(\mu(t) - \hat{f}_t)f_t + \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\hat{f}_t]\}dt + \frac{2\eta}{M(t)}\sqrt{\hat{f}_t(\mu(t) - \hat{f}_t)}dB_t. \tag{40}$$

Now derive the integration of equation 40, $\forall T \in [0, \mathcal{T}]$ we have

$$\hat{f}_T = \hat{f}_0 + \int_0^T \{\frac{\mu'(t)}{\mu(t)}\hat{f}_t + \frac{\eta\delta_t}{M(t)}(\mu(t) - \hat{f}_t)f_t + \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\hat{f}_t]\}dt + \frac{2\eta}{M(t)}\sqrt{\hat{f}_t(\mu(t) - \hat{f}_t)}dB_t, \tag{41}$$

where $\hat{f}_0 = \mu(0)f_0$. Note both the dirft part and diffusion part in equation 40 are continous and uniformly bounded on $[0, T]$, hence we can get rid of the diffusion part by expectation:

$$\mathbb{E}\hat{f}_T = \hat{f}_0 + \mathbb{E}\int_0^T \{\frac{\mu'(t)}{\mu(t)}\hat{f}_t + \frac{\eta\delta_t}{M(t)}(\mu(t) - \hat{f}_t)f_t + \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\hat{f}_t]\}dt. \tag{42}$$

Prove upper bound equation 28 Notice in $\frac{\eta\delta_t}{M(t)}(\mu(t) - \hat{f}_t)f_t = \frac{\eta\delta_t}{M(t)}\mu(t)(1 - f_t)$, $f_t \in [0, 1]$ $\eta, k_t, M(t), \mu(t) > 0$, $\delta_t \leq a_p - a_1$, hence we have

$$0 \leq \frac{\eta\delta_t}{M(t)}(\mu(t) - \hat{f}_t)f_t \leq \frac{\eta(a_p - a_1)}{M(t)}(\mu(t) - \hat{f}_t) \tag{43}$$

Bring equation 43 into equation 42 to remove f_t, then move the symbol $\mathbb{E}$ inside the integration, we have

$$\begin{aligned}\mathbb{E}\hat{f}_T \leq& \hat{f}_0 + \mathbb{E}\int_0^T \{\frac{\mu'(t)}{\mu(t)}\hat{f}_t + \frac{\eta(a_p - a_1)}{M(t)}(\mu(t) - \hat{f}_t) + \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\hat{f}_t]\}dt \\ =& \hat{f}_0 + \int_0^T \{\frac{\mu'(t)}{\mu(t)}\mathbb{E}\hat{f}_t + \frac{\eta(a_p - a_1)}{M(t)}(\mu(t) - \mathbb{E}\hat{f}_t) + \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\mathbb{E}\hat{f}_t]\}dt.\end{aligned} \tag{44}$$

equation 44 can be rewritten as

$$\mathbb{E}\hat{f}_T + \int_0^T (\frac{\eta(a_p - a_1)}{M(t)} + \frac{p\eta^2\sigma^2}{M(t)^2} - \frac{\mu'(t)}{\mu(t)})\mathbb{E}\hat{f}_t dt \leq \hat{f}_0 + \int_0^T (\frac{\eta(a_p - a_1)}{M(t)} + \frac{\eta^2\sigma^2}{M(t)^2})\mu(t)dt, \tag{45}$$

or rewritten as equation 46 using definition of $g_1(t), g_2(t)$:

$$\mathbb{E}\hat{f}_T + \int_0^T (g_1(t) - \frac{\mu'(t)}{\mu(t)})\mathbb{E}\hat{f}_t dt \leq \hat{f}_0 + \int_0^T g_2(t)\mu(t)dt. \tag{46}$$

Notice $\mathbb{E}\hat{f}_t = \mu(t)\mathbb{E}f_t$, hence $(g_1(t) - \frac{\mu'(t)}{\mu(t)})\mathbb{E}\hat{f}_t = (g_1(t)\mu(t) - \mu'(t))\mathbb{E}f_t$, let

$$h(t) \triangleq g_1(t)\mu(t) - \mu'(t). \tag{47}$$

Obviously $h(t)$ is continous on $[0, \mathcal{T}]$. Then equation 46 can be rewritten as

$$\mathbb{E}\hat{f}_T + \int_0^T \frac{h(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt \leq \hat{f}_0 + \int_0^T g_2(t)\mu(t)dt. \tag{48}$$

The integration inequation equation 48 has an explicit solution: by equation 48, $\forall T \in [0, \mathcal{T}]$ we have

$$\mathbb{E}\hat{f}_T + \int_0^T \frac{h(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt \leq \hat{f}_0 + \int_0^T g_2(t)\mu(t)dt \tag{49}$$

$$\iff e^{\int^T \frac{h(t)}{\mu(t)}dt}\frac{h(T)}{\mu(T)}[\mathbb{E}\hat{f}_T + \int_0^T \frac{h(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt] \leq e^{\int^T \frac{h(t)}{\mu(t)}dt}\frac{h(T)}{\mu(T)}[\hat{f}_0 + \int_0^T g_2(t)\mu(t)dt] \tag{50}$$

$$\iff \frac{d}{dt}[e^{\int^T \frac{h(t)}{\mu(t)}dT}\int_0^T \frac{h(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt] \leq e^{\int^T \frac{h(t)}{\mu(t)}dt}\frac{h(T)}{\mu(T)}[\hat{f}_0 + \int_0^T g_2(t)\mu(t)dt] \tag{51}$$

$$\implies e^{\int^T \frac{h(t)}{\mu(t)}dt}\int_0^T \frac{h(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt \leq \int_0^T \{e^{\int^t \frac{h(\tau)}{\mu(\tau)}d\tau}\frac{h(t)}{\mu(t)}[\hat{f}_0 + \int_0^t g_2(\tau)\mu(\tau)d\tau]\}dt \tag{52}$$

$$\iff \int_0^T h(t)\mathbb{E}f_t dt \leq e^{-\int^T \frac{h(t)}{\mu(t)}dt}\int_0^T \{e^{\int^t \frac{h(\tau)}{\mu(\tau)}d\tau}\frac{h(t)}{\mu(t)}[\hat{f}_0 + \int_0^t g_2(\tau)\mu(\tau)d\tau]\}dt. \tag{53}$$

Note $\int^T \frac{h(t)}{\mu(t)}dt = C + \int_0^T \frac{h(t)}{\mu(t)}dt$ is primitive of $\frac{h(T)}{\mu(T)}$, we will determine its base C later. Now let's deal with $h(t)$ and $\mu(t)$ in equation 53. Recall $h(t) = g_1(t)\mu(t) - \mu'(t)$, which is an one-dimensional ODE, and has an explicit solution given $h(t) \in \mathcal{C}[0, \mathcal{T}]$ and $\mu(0) > 0$:

$$\mu(T) = e^{G_1(T)}[\mu(0) - \int_0^T h(t)e^{-G_1(t)}dt]. \tag{54}$$

To ensure $\mu(t) > 0$, we further assume $\forall t \in [0, \mathcal{T}], h(t) > 0, \int_0^{\mathcal{T}} h(t) = 1$, and $\mu(0) > 1$. Let's demonstrate these assumptions are sufficient to ensure $\mu(t) > 0$: since $g_1(t)$ is always postive regardless of $t \in [0, \infty)$, then $G_1(t) = \int_0^t g_1(\tau)d\tau$ is strictly monotonously increasing on $[0, \mathcal{T})$, and $G_1(0) = 0$. Therefore $\forall t \in [0, \mathcal{T}]$

$$\int_0^T h(t)e^{-G_1(t)}dt \leq \int_0^{\mathcal{T}} h(t)e^{-G_1(t)}dt < \int_0^T h(t)e^{-G_1(0)}dt = 1, \tag{55}$$

$$\implies \quad \mu(t) > e^{G_1(T)}(\mu(0) - 1) > 0. \tag{56}$$

Using equation 54, we can dervie the form of $\int^T \frac{h(t)}{\mu(t)}dt$ using $h(t)$:

$$\int^T \frac{h(t)}{\mu(t)}dt = C + \int_0^T \frac{h(t)}{\mu(t)}dt = C + \int_0^T \frac{h(t)e^{-G_1(t)}}{\mu(0) - \int_0^t h(\tau)e^{-G_1(\tau)}d\tau}dt = -\ln[\mu(0) - \int_0^t h(t)e^{-G_1(\tau)}d\tau]\Big|_0^T + C. \tag{57}$$

Here we can set $C = \ln(\mu(0))$ to ensure

$$\int^T \frac{h(t)}{\mu(t)}dt = -\ln[\mu(0) - \int_0^T h(t)e^{-G_1(t)}dt]. \tag{58}$$

Take equation 54, equation 58 into equation 53, we have

$$\begin{aligned}
\int_0^T h(t)\mathbb{E}f_t dt \leq & [\mu(0) - \int_0^T h(t)e^{-G_1(t)}dt] \int_0^T \frac{h(t)e^{-G_1(t)}}{[\mu(0) - \int_0^t h(\tau)e^{-G_1(\tau)}d\tau]^2}[\hat{f}_0 + \int_0^t g_2(\tau)\mu(\tau)d\tau]dt \\
\leq & [\mu(0) - \int_0^T h(t)e^{-G_1(t)}dt] \int_0^T \frac{h(t)e^{-G_1(t)}}{[\mu(0) - \int_0^{\mathcal{T}} h(\tau)e^{-G_1(\tau)}d\tau]^2}[\hat{f}_0 + \int_0^t g_2(\tau)\mu(\tau)d\tau]dt \\
= & \frac{\mu(0) - \int_0^T h(t)e^{-G_1(t)}dt}{[\mu(0) - \int_0^{\mathcal{T}} h(t)e^{-G_1(t)}dt]^2} \int_0^T h(t)e^{-G_1(t)}[\hat{f}_0 + \int_0^t g_2(\tau)\mu(\tau)d\tau]dt
\end{aligned} \tag{59}$$

Set $T = \mathcal{T}$, we have

$$\begin{aligned}
\int_0^{\mathcal{T}} h(t)dt\mathbb{E}f_t \leq & \frac{\mu(0) - \int_0^{\mathcal{T}} h(t)e^{-G_1(t)}dt}{[\mu(0) - \int_0^{\mathcal{T}} h(T)e^{-G_1(T)}dt]^2} \int_0^{\mathcal{T}} h(T)e^{-G_1(T)}[\hat{f}_0 + \int_0^T g_2(t)\mu(t)dt]dT \\
= & \frac{1}{\mu(0) - 1} \int_0^{\mathcal{T}} h(T)e^{-G_1(T)}\{\hat{f}_0 + \int_0^T g_2(t)e^{G_1(t)}[\mu(0) - \int_0^t h(\tau)e^{-G_1(\tau)}d\tau]dt\}dT \\
= & \frac{\mu(0)}{\mu(0) - 1} \int_0^{\mathcal{T}} h(T)e^{-G_1(T)}[f_0 + \int_0^T g_2(t)e^{G_1(t)}dt]dT \\
& - \frac{1}{\mu(0) - 1} \int_0^{\mathcal{T}} h(T)e^{-G_1(T)} \int_0^T g_2(t)e^{G_1(t)} \int_0^t h(\tau)e^{-G_1(\tau)}d\tau dt dT
\end{aligned} \tag{60}$$

Note we haven't determined the specific form of $h(t)$ and value of $\mu(0)$ yet. Now for a given $K \in (0, +\infty)$, define $h_K(t) = \frac{e^{Kt}}{\int_0^T e^{K\tau}d\tau}$, take $h_K(t)$ into equation 60, we have

$$\begin{aligned}
\int_0^{\mathcal{T}} h_K(T)dT\mathbb{E}f_T \leq & \frac{\mu(0)}{\mu(0) - 1} \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)}[f_0 + \int_0^T g_2(t)e^{G_1(t)}dt]dT \\
& - \frac{1}{\mu(0) - 1} \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)} \int_0^T g_2(t)e^{G_1(t)} \int_0^t h_K(\tau)e^{-G_1(\tau)}d\tau dt dT
\end{aligned} \tag{61}$$

By lemma C.0.1, $\mathbb{E}f_T$ is continous on $[0, \mathcal{T}]$; $e^{-G_1(T)}[f_0 + \int_0^T g_2(t)e^{G_1(t)}dt]$ is also continous on $[0, \mathcal{T}]$ because each part in it is continous. Therefore by lemma C.0.2, we have

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)\mathbb{E}f_T dT = \mathbb{E}f_{\mathcal{T}}, \tag{62}$$

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)}[f_0 + \int_0^T g_2(t)e^{G_1(t)}dt]dT = e^{-G_1(\mathcal{T})}[f_0 + \int_0^{\mathcal{T}} g_2(t)e^{G_1(t)}dt]. \tag{63}$$

Next we will prove

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)} \int_0^T g_2(t)e^{G_1(t)} \int_0^t h_K(\tau)e^{-G_1(\tau)}d\tau dt dT = 0 \tag{64}$$

Recall $g_1(t), g_2(t)$ are postive, continous and monotonous on $[0, +\infty)$, and converge to fixed values g_2^*, then

$$\underline{g}_1 \triangleq \min(g_1(0), g_1^*) \leq g_1(t) \leq \max(g_1(0), g_1^*) \tag{65}$$

$$\min(g_2(0), g_2^*) \leq g_2(t) \leq \max(g_2(0), g_2^*) \tag{66}$$

Hence by equation 66 we have

$$\begin{aligned}
&\Big| \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)} \int_0^T g_2(t)e^{G_1(t)} \int_0^t h(\tau)e^{-G_1(\tau)}d\tau dt dT \Big|, \\
=&\int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)} \int_0^T |g_2(t)|e^{G_1(t)} \int_0^t h_K(\tau)e^{-G_1(\tau)}d\tau dt dT, \\
\leq& \max(g_2(0), g_2^*) \int_0^{\mathcal{T}} h_K(T)e^{-G_1(T)} \int_0^T e^{G_1(t)} \int_0^t h_K(\tau)e^{-G_1(\tau)}d\tau dt dT, \\
\leq& \max(g_2(0), g_2^*) \int_0^{\mathcal{T}} h_K(T) \int_0^T e^{G_1(t)-G_1(T)} \int_0^t h_K(\tau)e^{-G_1(\tau)}d\tau dt dT;
\end{aligned} \tag{67}$$

By equation 65, for $0 \leq t \leq T$ we have

$$G_1(t) - G_1(T) = \int_T^t g_1(\tau)d\tau \leq \int_T^t \underline{g}_1 d\tau = \underline{g}_1(t - T); \tag{68}$$

$$-G_1(t) = -\int_0^t g_1(\tau)d\tau \leq -\int_0^t \underline{g}_1 d\tau = -\underline{g}_1 t, \tag{69}$$

Take equation 68, equation 69 into equation 67, assume $K > \underline{g}_1$ we have

$$\begin{aligned}
&\int_0^{\mathcal{T}} h_K(T) \int_0^T e^{G_1(t)-G_1(T)} \int_0^t h_K(\tau) e^{-G_1(\tau)} d\tau dt dT \\
\leq &\int_0^{\mathcal{T}} h_K(T) \int_0^T e^{\underline{g}_1(t-T)} \int_0^t h_K(\tau) e^{-\underline{g}_1\tau} d\tau dt dT \\
= &\frac{1}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2} \int_0^{\mathcal{T}} e^{(K-\underline{g}_1)T} \int_0^T e^{\underline{g}_1 t} \int_0^t e^{(K-\underline{g}_1)\tau} d\tau dt dT \\
\leq &\frac{1}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2} \int_0^{\mathcal{T}} e^{(K-\underline{g}_1)T} \int_0^T e^{\underline{g}_1 t} \frac{1}{K-\underline{g}_1} e^{(K-\underline{g}_1)t} dt dT \\
= &\frac{1}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2 (K-\underline{g}_1)} \int_0^{\mathcal{T}} e^{(K-\underline{g}_1)T} \int_0^T e^{Kt} dt dT \\
\leq &\frac{1}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2 (K-\underline{g}_1)} \int_0^{\mathcal{T}} e^{(K-\underline{g}_1)T} \frac{1}{K} e^{KT} dT \\
= &\frac{1}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2 (K-\underline{g}_1) K} \int_0^{\mathcal{T}} e^{(2K-\underline{g}_1)T} dT \\
\leq &\frac{e^{(2K-\underline{g}_1)\mathcal{T}}}{(\int_0^{\mathcal{T}} e^{Kt} dt)^2 (K-\underline{g}_1) K (2K-\underline{g}_1)} \\
= &\frac{e^{-\underline{g}_1\mathcal{T}}}{(1-e^{-K\mathcal{T}})^2 (2-\underline{g}_1/K)} \cdot \frac{1}{(K-\underline{g}_1)} \longrightarrow 0, \quad K \longrightarrow +\infty.
\end{aligned} \tag{70}$$

Therefore, equation 64 holds. Take equation 62, equation 63, equation 64 into equation 61, as $K \to +\infty$, we have

$$\mathbb{E} f_{\mathcal{T}} \leq \frac{\mu(0)}{\mu(0)-1} e^{-G_1(\mathcal{T})} [f_0 + \int_0^{\mathcal{T}} g_2(t) e^{G_1(t)} dt]. \tag{71}$$

Notice, $\forall \mu(0) \in (1, +\infty)$, equation 71 holds, which means when $\mu(0) \to +\infty$, we have

$$\mathbb{E} f_{\mathcal{T}} \leq e^{-G_1(\mathcal{T})} [f_0 + \int_0^{\mathcal{T}} g_2(t) e^{G_1(t)} dt]. \tag{72}$$

Since $\mathcal{T}$ is arbitarily given, the upper bound equation 28 has been proven.

Prove lower bound equation 33 Let's go back to see equation 42, notice

$$\begin{aligned}
\mathbb{E}\hat{f}_T =&\hat{f}_0 + \mathbb{E} \int_0^T \{\frac{\mu'(t)}{\mu(t)} \hat{f}_t + \frac{\eta \delta_t}{M(t)} (\mu(t) - \hat{f}_t) f_t + \frac{\eta^2}{M(t)^2} [\sigma^2 \mu(t) - p\sigma^2 \hat{f}_t]\} dt \\
=&\hat{f}_0 + \int_0^T \frac{\mu'(t)}{\mu(t)} \mathbb{E}\hat{f}_t dt + \int_0^T \frac{\eta \mu(t)}{M(t)} \mathbb{E}\delta_t (1-f_t) f_t dt + \int_0^T \frac{\eta^2}{M(t)^2} [\sigma^2 \mu(t) - p\sigma^2 \mathbb{E}\hat{f}_t] dt.
\end{aligned} \tag{73}$$

To handle $\mathbb{E}\delta_t(1-f_t)f_t$, recall $\delta_t \in [a_2 - a_1, a_p - a_1]$ and $f_t \in [0,1]$, so we have

$$\mathbb{E}\delta_t(1-f_t)f_t \geq (a_2 - a_1)\mathbb{E}(1-f_t)f_t. \tag{74}$$

Notice

$$\begin{aligned}
\mathbb{E}(1-f_t)f_t =&\mathbb{E}(1-f_t)f_t \mathbb{1}_{\{f_t \geq \xi\}} + \mathbb{E}(1-f_t)f_t \mathbb{1}_{\{f_t < \xi\}} \\
\geq &\mathbb{E}(1-f_t)\xi \mathbb{1}_{\{f_t \geq \xi\}} + \mathbb{E}(1-f_t)\xi \mathbb{1}_{\{f_t < \xi\}} - \mathbb{E}(1-f_t)(\xi - f_t)\mathbb{1}_{\{f_t < \xi\}} \\
= &\xi \mathbb{E}(1-f_t) - \mathbb{E}(1-f_t)(\xi - f_t)\mathbb{1}_{\{f_t < \xi\}} \\
\geq &\xi \mathbb{E}(1-f_t) - \mathbb{E}\xi \mathbb{1}_{\{f_t < \xi\}} \\
= &\xi \mathbb{E}(1-f_t) - \xi P(f_t < \xi) \\
= &\xi(1 - \varepsilon(t) - \mathbb{E} f_t)
\end{aligned} \tag{75}$$

Take equation 75 into equation 73, we have

$$\begin{aligned}\mathbb{E}\hat{f}_T =&\hat{f}_0 + \int_0^T \frac{\mu'(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt + \int_0^T \frac{\eta\mu(t)}{M(t)}\mathbb{E}\delta_t(1-f_t)f_t dt + \int_0^T \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\mathbb{E}\hat{f}_t]dt \\ \geq&\hat{f}_0 + \int_0^T \frac{\mu'(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt + \int_0^T \frac{\eta\mu(t)}{M(t)}(a_2-a_1)\xi(1-\varepsilon(t)-\mathbb{E}f_t)dt + \int_0^T \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\mathbb{E}\hat{f}_t]dt \\ =&\hat{f}_0 + \int_0^T \frac{\mu'(t)}{\mu(t)}\mathbb{E}\hat{f}_t dt + \int_0^T \frac{\eta(a_2-a_1)\xi}{M(t)}[(1-\varepsilon(t))\mu(t) - \mathbb{E}\hat{f}_t]dt + \int_0^T \frac{\eta^2}{M(t)^2}[\sigma^2\mu(t) - p\sigma^2\mathbb{E}\hat{f}_t]dt,\end{aligned} \tag{76}$$

which can be written as

$$\mathbb{E}\hat{f}_T + \int_0^T (\frac{\eta(a_2-a_1)\xi}{M(t)} + \frac{p\eta^2\sigma^2}{M(t)^2} - \frac{\mu'(t)}{\mu(t)})\mathbb{E}\hat{f}_t dt \geq \hat{f}_0 + \int_0^T (\frac{\eta(a_2-a_1)\xi}{M(t)}(1-\varepsilon(t)) + \frac{\eta^2\sigma^2}{M(t)^2})\mu(t)dt, \tag{77}$$

or rewritten as equation 78 using definition of $\tilde{g}_1(t), \tilde{g}_2(t)$.

$$\mathbb{E}\hat{f}_T + \int_0^T (\tilde{g}_1(t) - \frac{\mu'(t)}{\mu(t)})\mathbb{E}\hat{f}_t dt \geq \hat{f}_0 + \int_0^T \tilde{g}_2(t)\mu(t)dt. \tag{78}$$

Similar to the procedure from equation 48 to equation 59, we can derive an explicit solution of equation 78 by auxiliary functions $h(t), \mu(t)$: $\forall T \in [0, \mathcal{T}]$

$$\int_0^T h(t)\mathbb{E}f_t dt \geq [\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt]\int_0^T \frac{h(t)e^{-\tilde{G}_1(t)}}{[\mu(0) - \int_0^t h(\tau)e^{-\tilde{G}_1(\tau)}d\tau]^2}[\hat{f}_0 + \int_0^t \tilde{g}_2(\tau)\mu(\tau)d\tau]dt; \tag{79}$$

where $h(t), \mu(t)$ satisfy

- $\forall t \in [0, \mathcal{T}]$, $h(t), \mu(t) > 0$;
- $\int_0^{\mathcal{T}} h(t)dt = 1$, $\mu(0) > 1$;
- $\mu(T) = e^{\tilde{G}_1(T)}[\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt]$.

since $\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt \leq \mu(0)$, we have

$$\begin{aligned}\int_0^T h(t)\mathbb{E}f_t dt \geq&[\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt]\int_0^T \frac{h(t)e^{-\tilde{G}_1(t)}}{[\mu(0) - \int_0^t h(\tau)e^{-\tilde{G}_1(\tau)}d\tau]^2}[\hat{f}_0 + \int_0^t \tilde{g}_2(\tau)\mu(\tau)d\tau]dt \\ \geq&[\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt]\int_0^T \frac{h(t)e^{-\tilde{G}_1(t)}}{\mu(0)^2}[\hat{f}_0 + \int_0^t \tilde{g}_2(\tau)\mu(\tau)d\tau]dt \\ =&\frac{\mu(0) - \int_0^T h(t)e^{-\tilde{G}_1(t)}dt}{\mu(0)^2}\int_0^T h(t)e^{-\tilde{G}_1(t)}[\hat{f}_0 + \int_0^t \tilde{g}_2(\tau)\mu(\tau)d\tau]dt.\end{aligned} \tag{80}$$

Set $T = \mathcal{T}$, we have

$$\begin{aligned}
\int_0^{\mathcal{T}} h(t)\mathbb{E}f_t dt \geq & \frac{\mu(0) - \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}dT}{\mu(0)^2} \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}[\hat{f}_0 + \int_0^T \tilde{g}_2(t)\mu(t)dt]dT \\
\geq & \frac{\mu(0) - 1}{\mu(0)^2} \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}[\hat{f}_0 + \int_0^T \tilde{g}_2(t)\mu(t)dt]dT \\
= & \frac{\mu(0) - 1}{\mu(0)^2} \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}\{\hat{f}_0 + \int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}[\mu(0) - \int_0^t h(\tau)e^{-\tilde{G}_1(\tau)}d\tau]dt\}dT \\
= & \frac{\mu(0) - 1}{\mu(0)} \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}dt]dT \\
& - \frac{\mu(0) - 1}{\mu(0)^2} \int_0^{\mathcal{T}} h(T)e^{-\tilde{G}_1(T)}\{\int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}[\int_0^t h(\tau)e^{-\tilde{G}_1(\tau)}d\tau]dt\}dT
\end{aligned} \tag{81}$$

Similar to equation 61, set $h(t)$ as $h_K(t) = \frac{e^{Kt}}{\int_0^T e^{K\tau}d\tau}$ in equation 81, we have

$$\begin{aligned}
\int_0^{\mathcal{T}} h_K(t)\mathbb{E}f_t dt \geq & \frac{\mu(0) - 1}{\mu(0)} \int_0^{\mathcal{T}} h_K(T)e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}dt]dT \\
& - \frac{\mu(0) - 1}{\mu(0)^2} \int_0^{\mathcal{T}} h_K(T)e^{-\tilde{G}_1(T)}\{\int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}[\int_0^t h_K(\tau)e^{-\tilde{G}_1(\tau)}d\tau]dt\}dT.
\end{aligned} \tag{82}$$

By lemma C.0.2, we have

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)\mathbb{E}f_T dT = \mathbb{E}f_{\mathcal{T}}, \tag{83}$$

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}dt]dT = e^{-\tilde{G}_1(\mathcal{T})}[f_0 + \int_0^{\mathcal{T}} \tilde{g}_2(t)e^{\tilde{G}_1(t)}dt]. \tag{84}$$

Similar to equation 64, we can prove

$$\lim_{K\to+\infty} \int_0^{\mathcal{T}} h_K(T)e^{-\tilde{G}_1(T)}\{\int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}[\int_0^t h_K(\tau)e^{-\tilde{G}_1(\tau)}d\tau]dt\}dT = 0. \tag{85}$$

The only difference in the proof is that unlike $g_2(t)$, $\tilde{g}_2(t)$ is not necessarily monotonous on $[0, +\infty)$, but we have $0 \leq \tilde{g}_2(t) < g_2(t), \forall t \in [0, +\infty)$, and $g_2(t)$ is uniformly bounded, so $\tilde{g}_2(t)$ is uniformly bounded too.

Therefore in equation 82, as $K \to, \mu(0) + \infty$, we have

$$\mathbb{E}f_{\mathcal{T}} \geq e^{-\tilde{G}_1(\mathcal{T})}[f_0 + \int_0^T \tilde{g}_2(t)e^{\tilde{G}_1(t)}dt]. \tag{86}$$

Since $\mathcal{T}$ is arbitarily given, the lower bound equation 33 has been proven. □

Put $r_t = 1 - \mathbb{E}f_t$, $\Delta_t = \sqrt{\frac{\mathrm{Tr}(\boldsymbol{P}_t\tilde{\boldsymbol{\Sigma}})\eta^2}{M_t^2}}$(equation 11) into Theorem C.0.3, we can get the form of Theorem 1 in main text.

D Proof of the results on equilibrium state dynamics

D.1 Expectation bounds

The following corollary follows directly from Theorem C.0.3.

Corollary D.1.1 (Variant of Corollary 1 in main text)**.** *Given* $\Delta = \sqrt{2\eta\lambda}$*, assume* $M(0) = M^* \triangleq \sqrt{\frac{\eta(p-1)\sigma^2}{2\lambda}}$*, and* $\exists \varepsilon > 0, \overline{\lim}_{t\to+\infty} \varepsilon(t) < \varepsilon$ *in theorem C.0.3, then we have*

$$\underline{f}^* - \varepsilon + e^{-\tilde{g}_1^* t}(f_0 - \underline{f}^* - C) \leq \mathbb{E}f_t \leq \bar{f}^* + e^{-g_1^* t}(f_0 - \bar{f}^*). \tag{87}$$

where $g_1^* = (\frac{a_p - a_1}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta)\Delta$, $\bar{f}^* = 1 - \frac{\Delta}{\frac{a_p-a_1}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta}$, $\tilde{g}_1^* = (\frac{a_p - a_1}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta)\Delta$, $\underline{f}^* = 1 - \frac{\Delta}{\frac{\xi(a_p-a_1)}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta}$.

Proof. When $M(0) = \sqrt{\frac{\eta(p-1)\sigma^2}{2\lambda}}$, based on the evolution of $M(t)$ (equation 29), $M(t) \equiv \sqrt{\frac{\eta(p-1)\sigma^2}{2\lambda}}$, then $G_1(t), g_2(t), \tilde{G}_1(t), \tilde{g}_2(t)$ will be

$$G_1(t) = \int_0^t [(a_p - a_1)\sqrt{\frac{2\lambda\eta}{(p-1)\sigma^2}} + \frac{2p\lambda\eta}{(p-1)}]d\tau = (\frac{a_p - a_1}{\sqrt{p-1}\sigma}\Delta + \frac{p}{p-1}\Delta^2) \cdot t = g_1^* t; \tag{88}$$

$$g_2(t) = (a_p - a_1)\sqrt{\frac{2\lambda\eta}{(p-1)\sigma^2}} + \frac{2\lambda\eta}{(p-1)} = \frac{a_p - a_1}{\sqrt{p-1}\sigma}\Delta + \frac{1}{p-1}\Delta^2; \tag{89}$$

$$\tilde{G}_1(t) = \int_0^t [\xi(a_2 - a_1)\sqrt{\frac{2\lambda\eta}{(p-1)\sigma^2}} + \frac{2p\lambda\eta}{(p-1)}]d\tau = (\frac{\xi(a_2 - a_1)}{\sqrt{p-1}\sigma}\Delta + \frac{p}{p-1}\Delta^2) \cdot t = \tilde{g}_1^* t; \tag{90}$$

$$\begin{aligned}\tilde{g}_2(t) =& \xi(a_2 - a_1)(1 - \varepsilon(t))\sqrt{\frac{2\lambda\eta}{(p-1)\sigma^2}} + \frac{2\lambda\eta}{(p-1)} \\ =& \frac{\xi(a_2 - a_1)}{\sqrt{p-1}\sigma}\Delta + \frac{1}{p-1}\Delta^2 - \frac{\xi(a_2 - a_1)\Delta}{\sqrt{p-1}\sigma}\varepsilon(t).\end{aligned} \tag{91}$$

Then the upper bound equation 28 can be written as

$$\begin{aligned}\mathbb{E}f_t \leq& e^{-g_1^* t}[f_0 + \int_0^t (\frac{a_p - a_1}{\sqrt{p-1}\sigma}\Delta + \frac{1}{p-1}\Delta^2)e^{g_1^* \tau}d\tau] \\ =& e^{-g_1^* t}f_0 + (1 - e^{-g_1^* t})\frac{\frac{a_p-a_1}{\sqrt{p-1}\sigma}\Delta + \frac{1}{p-1}\Delta^2}{\frac{a_p-a_1}{\sqrt{p-1}\sigma}\Delta + \frac{p}{p-1}\Delta^2} \\ =& e^{-g_1^* t}f_0 + (1 - e^{-g_1^* t})(1 - \frac{\Delta}{\frac{a_p-a_1}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta}) \\ =& e^{-g_1^* t}f_0 + (1 - e^{-g_1^* t})\bar{f}^* \\ =& \bar{f}^* + e^{-g_1^* t}(f_0 - \bar{f}^*).\end{aligned} \tag{92}$$

The lower bound equation 33 can be written as

$$\begin{aligned}\mathbb{E}f_t \geq& e^{-\tilde{g}_1^* t}[f_0 + \int_0^t (\frac{\xi(a_2 - a_1)}{\sqrt{p-1}\sigma}\Delta + \frac{1}{p-1}\Delta^2)e^{\tilde{g}_1^* \tau}d\tau] - \frac{\xi(a_2 - a_1)\Delta}{\sqrt{p-1}\sigma}e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}\varepsilon(\tau)d\tau \\ =& e^{-\tilde{g}_1^* t}f_0 + (1 - e^{-g_1^* t})(1 - \frac{\Delta}{\frac{\xi(a_2-a_1)}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta}) - \frac{\xi(a_2 - a_1)\Delta}{\sqrt{p-1}\sigma}e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}\varepsilon(\tau)d\tau \\ =& e^{-\tilde{g}_1^* t}f_0 + (1 - e^{-g_1^* t})\underline{f}^* - \frac{\xi(a_2 - a_1)\Delta}{\sqrt{p-1}\sigma}e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}\varepsilon(\tau)d\tau\end{aligned} \tag{93}$$

Now let's estimate the value of $e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}\varepsilon(\tau)d\tau$. Recall the assumptiion that $\overline{\lim}_{t\to\infty} \varepsilon(t) < \varepsilon$, which means $\exists T_0 > 0, \forall t > T_0, \varepsilon(t) < \varepsilon$, then if $t \leq T_0$,

$$e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}\varepsilon(\tau)d\tau \leq e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^* \tau}d\tau = \frac{1 - e^{-\tilde{g}^* t}}{\tilde{g}^*} \leq \frac{1}{\tilde{g}^*} \leq \frac{\varepsilon + e^{-\tilde{g}^* t + \tilde{g}^* T_0}}{\tilde{g}^*}; \tag{94}$$

If $t > T_0$, then we have

$$\begin{aligned} e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^*\tau}\varepsilon(\tau)d\tau =& e^{-\tilde{g}^* t}[\int_0^T e^{\tilde{g}_1^*\tau}\varepsilon(\tau)d\tau + \int_T^t e^{\tilde{g}_1^*\tau}\varepsilon(\tau)d\tau] \\ \leq& e^{-\tilde{g}^* t}[\int_0^T e^{\tilde{g}_1^*\tau}d\tau + \int_T^t e^{\tilde{g}_1^*\tau}\varepsilon d\tau] \\ =& \frac{e^{-\tilde{g}^* t}}{\tilde{g}^*}[(1-\varepsilon)(e^{\tilde{g}^* T}-1)+\varepsilon(e^{\tilde{g}^* t}-1)] \\ =& \frac{\varepsilon + e^{-\tilde{g}^* t}[(1-\varepsilon)(e^{\tilde{g}^* T}-1)-\varepsilon]}{\tilde{g}^*} \\ \leq& \frac{\varepsilon + e^{-\tilde{g}^* t+\tilde{g}^* T_0}}{\tilde{g}^*} \end{aligned} \tag{95}$$

Summarize the two cases above,we have $\forall t \in [0, +\infty)$,

$$e^{-\tilde{g}^* t}\int_0^t e^{\tilde{g}_1^*\tau}\varepsilon(\tau)d\tau \leq \frac{\varepsilon + e^{-\tilde{g}^* t+\tilde{g}^* T_0}}{\tilde{g}^*}. \tag{96}$$

Take equation 96 into equation 93, we have

$$\begin{aligned} \mathbb{E}f_t \geq& e^{-\tilde{g}_1^* t}f_0 + (1-e^{-g_1^* t})\underline{f}^* - \frac{\xi(a_2-a_1)\Delta}{\sqrt{p-1}\sigma}\cdot\frac{\varepsilon + e^{-\tilde{g}^* t+\tilde{g}^* T_0}}{\tilde{g}^*} \\ \geq& e^{-\tilde{g}_1^* t}f_0 + (1-e^{-g_1^* t})\underline{f}^* - (\varepsilon + e^{-\tilde{g}^* t+\tilde{g}^* T_0}) \\ =& \underline{f}^* - \varepsilon + e^{-\bar{g}^* t}(f_0 - \underline{f}^* - e^{\bar{g}^* T}). \end{aligned} \tag{97}$$

Set $C = e^{\tilde{g}^* T_0}$. Summarize equation 92, equation 97, equation 87 holds. □

Put $r_t = 1 - \mathbb{E}f_t$ in corollary 1 in main text, we can obtain the form of corollary 1 in main text.

D.2 Detailed depiction based on Fokker-Planck equation

In addition to the bounds on $\mathbb{E}f_t$, the exact stationary distribution of f_t can be solved with Fokker-Planck equation under Assumption B.2.2. Without loss of generality, we first set $\boldsymbol{A} = diag(0, 1, 1, \ldots, 1)$ and then convert the results to the general case $\boldsymbol{A} = diag(a_l, a_h, a_h, \ldots, a_h)$. Let $\beta_1 = \tilde{\boldsymbol{X}}_1 = \boldsymbol{e}_1^T\tilde{\boldsymbol{X}}$. In the case that $\boldsymbol{A} = diag(0, 1, 1, \ldots, 1)$, the dynamics of β_1 is given by

$$\begin{aligned} d\beta_1 =& -\frac{\eta}{M_t}(\beta_1^3 - \beta_1)dt - \frac{\eta^2(p-1)}{2M_t^2}\beta_1 dt \\ & - \frac{\eta\sigma}{M_t}\sqrt{1-\beta_1^2}dB_t \\ \triangleq& -a_t(\beta_1^3 - \beta_1)dt - b_t\beta_1 dt - q_t\sqrt{1-\beta_1^2}dB_t \end{aligned} \tag{98}$$

We introduce the transform $\beta_1 = \sin\theta$ to make the diffusion term space-homogeneous. Then the dynamics of θ_t reads

$$d\theta_t = \frac{a(t)}{2}\sin 2\theta_t dt - [b(t) - \frac{c(t)^2}{2}]\tan\theta_t dt - q(t)dB_t \tag{99}$$

$$= -\frac{1}{2}\nabla_\theta U dt - q(t)dB_t \tag{100}$$

in which

$$U(\theta, t) = \frac{a(t)}{2}\cos 2\theta + [2b(t) - q(t)^2]\ln\sec\theta \tag{101}$$

serves as a time-dependent potential function over the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. We apply the Fokker-Planck equation in Lemma B.2.1 to the dynamics of θ_t and obtain the following theorem.

Theorem D.2.1. *Define $\rho(\theta, t)$ as the density of θ_t at time t. Then ρ is given by the Fokker-Planck equation*

$$\partial_t \rho = \partial_\theta \left\{ (\frac{\rho}{2} \partial_\theta U) + \partial_\theta [\frac{q(t)^2}{2} \rho] \right\} \tag{102}$$

$$\triangleq -\partial_\theta J(\theta, t) \tag{103}$$

in which

$$J(\theta, t) = -\frac{\rho}{2} \partial_\theta U - \frac{q(t)^2}{2} \partial_\theta \rho \tag{104}$$

is the probability current generated by the dynamics of θ_t.

D.3 Eigenfunction expansion and stationary distribution

In our analysis of equilibrium state, the weight norm is supposed to be initialized at its limiting value $\sqrt{M_*}$. Consequently, the coefficients $a(t)$, $b(t)$ and $q(t)$ in equation 99 also stays at their limiting value a_*, b_* and q_* throughout the training process, making the coefficients on the right-hand-side of equation 102 independent of t. We can then solve equation 102 by separation of variables. Set

$$\rho(\theta, t) = \Theta(\theta) T(t) \tag{105}$$

Then equation 102 can be rewritten as

$$\Theta T' = \frac{\Theta' T}{2} U' + \frac{\Theta T}{2} U'' + \frac{q_*^2}{2} \Theta'' T \tag{106}$$

For convenience, we introduce $V(\theta, t) = U(\theta, t)/q_t^2$, and that

$$V(\theta) = U_*(\theta)/q_*^2 \tag{107}$$

$$= \kappa \cos 2\theta + (p-2) \ln \sec \theta \tag{108}$$

where

$$\kappa = \frac{a_*}{2q_*^2} = \frac{\sqrt{p-1}}{2\sqrt{2\eta\lambda\sigma^2}} \tag{109}$$

Dividing both sides of equation 106 by ΘT, we have

$$\frac{2}{q_*^2} \frac{T'}{T} = \frac{\Theta''}{\Theta} + V' \frac{\Theta'}{\Theta} + V'' \tag{110}$$

Note that the left-hand-side of equation 110 involves only t, while the right-hand-side is a function of merely θ. Thus both sides must equal to a constant, which we denote by $-\lambda$:

$$\frac{2}{q_*^2} \frac{T'}{T} = \frac{\Theta''}{\Theta} + V' \frac{\Theta'}{\Theta} + V'' = -\lambda \tag{111}$$

or equivalently

$$-T' = \frac{q_*^2}{2} \lambda T \tag{112}$$

$$L_{FP} \Theta = \lambda \Theta \tag{113}$$

in which

$$L_{FP} = -\frac{d^2}{d\theta^2} - V' \frac{d}{d\theta} - V'' \tag{114}$$

$$= -\frac{d}{d\theta} (e^{-V} \frac{d}{d\theta} e^V) \tag{115}$$

is the Fokker Planck operator.
The basic solution to equation 112 is

$$T(t) = e^{-\frac{q_*^2}{2} \lambda t} \tag{116}$$

in which λ is determined by the boundary value problem:

$$L_{FP}\Theta = \lambda\Theta, \quad \Theta \in L_V^2(-\frac{\pi}{2}, \frac{\pi}{2}) \tag{117}$$

$$\lim_{\theta \to \pm\frac{\pi}{2}} e^{-V}\frac{d}{d\theta}(e^V\Theta) = 0 \tag{118}$$

in which $L_V^2(-\frac{\pi}{2}, \frac{\pi}{2})$ is the Hilbert space defined by

$$L_V^2(-\frac{\pi}{2}, \frac{\pi}{2}) = \left\{\rho(\theta) \left| \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\rho(\theta)|^2 e^{V(\theta)} d\theta < \infty \right.\right\} \tag{119}$$

The reason for choosing this Hilbert space is that once the initial density function $\rho(\theta, 0)$ is initialized in $L_V^2(-\frac{\pi}{2}, \frac{\pi}{2})$, the time-dependent solution $\rho(\theta, t)$ of Fokker-Planck equation equation 102 will remain in it as we will show later on. In fact, most common initial distributions including Gaussian initialization

$$\rho(\theta, 0) \propto \cos^{p-2}\theta \tag{120}$$

and initializing at certain fixed point

$$\rho(\theta, 0) = \delta_{\theta_0}(\theta) \tag{121}$$

fulfill this restriction. Boundary condition equation 118 follows from the reflecting boundary condition of probability flow

$$\lim_{\theta \to \pm\frac{\pi}{2}} J_*(\theta) = 0 \tag{122}$$

since the dynamics θ_t can not leave the interval $[\frac{\pi}{2}, \frac{\pi}{2}]$.
Based on the above analysis, we set out proving an important fact that the eigenvalues of of boundary value problem equation 118 and equation 122 are i). non-negative; ii). countable.

Lemma D.3.1. *The eigenvalues of boundary value problem equation 118 and equation 122 are non-negative. Moreover, $\lambda_0 = 0$ is an eigenvalue. The eigenfunction of $\lambda_0 = 0$ proportional to $\Theta_0 = e^{-V}$.*

Proof. Let $\Phi = \Phi(\theta)$ be an eigenfunction corresponding to an eigenvalue λ of the boundary value problem equation 118 and equation 122, that is,

$$L_{FP}\Phi = \lambda\Phi \tag{123}$$

$$\lim_{\theta \to \pm\frac{\pi}{2}} e^{-V}\frac{d}{d\theta}(e^V\Phi) = 0 \tag{124}$$

Multiplying both sides of equation 123 by Φe^V and then integrating by part, we have

$$\lambda \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\Phi|^2 e^V d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \overline{\Phi} e^V (L_{FP}\Phi) d\theta \tag{125}$$

$$= -\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \overline{\Phi} e^V \frac{d}{d\theta}[e^{-V}\frac{d}{d\theta}(e^V\Phi)] \tag{126}$$

$$= -\overline{\Phi} e^V [e^{-V}\frac{d}{d\theta}(e^V\Phi)]\Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\frac{d}{d\theta}(e^V\Phi)|^2 e^{-V} d\theta \tag{127}$$

$$= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\frac{d}{d\theta}(e^V\Phi)|^2 e^{-V} d\theta \geq 0 \tag{128}$$

which proves the self-ajointness of the operator LFP on $L_V^2(-\frac{\pi}{2}, \frac{\pi}{2})$, and the non-negativity of corresponding eigenvalues. Furthermore, the equality in equation 128 holds if and only if $\frac{d}{d\theta}(e^V\Phi) = 0$, that is, $\Phi \propto e^{-V}$. □

We then go on to prove the discreteness of the spectrum.

Lemma D.3.2. *Given a bounded interval $[-A, A] \subset \mathbb{R}$, and a potential $V_S(x)$ such that*

$$V_S \in \mathcal{C}^\infty(-A, A), \quad \inf_{x\in(-A,A)} V_S(x) > -\infty \tag{129}$$

Then the spectrum of the Schrödinger operator $H = -\frac{d^2}{dx^2} + V_S(x)$ in $L^2(-A, A)$ is purely discrete Simon (2008); Maz'ya (2007).

Corollary D.3.3. *For $p \geq 4$ boundary value problem equation 118 and equation 122 has at most countable eigenvalues which can be listed in ascending order*

$$0 = \lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \cdots \tag{130}$$

Proof. Define $\Theta = e^{-V/2}\Psi$, then equation 118 and equation 122 can be rewritten as

$$H\Psi = \lambda\Psi, \quad \Psi \in L^2(-\frac{\pi}{2}, -\frac{\pi}{2}) \tag{131}$$

$$\lim_{\theta\to\pm\frac{\pi}{2}} e^{-V}\frac{d}{d\theta}(e^{V/2}\Psi) = 0 \tag{132}$$

in which $H = -\frac{d^2}{d\theta^2} + V_S(\theta)$ is the Schrödinger operator, $V_S = \frac{V'^2}{4} = \frac{V''}{2}$ is the corresponding Schrödinger potential. Specifically,

$$\begin{aligned} V_S(\theta) = &\frac{(p-2)(p-4)}{4}\tan^2\theta + \kappa^2\sin^2 2\theta + \kappa p\cos 2\theta \\ &- (\kappa + \frac{1}{2})(p-2), \quad \theta \in (-\frac{\pi}{2}, \frac{\pi}{2}) \end{aligned} \tag{133}$$

For $p \geq 4$, condition equation 129 in Lemma D.3.2 holds. Then the theorem follows from Lemma D.3.2. □

Combining Theorem D.3.1 with Theorem D.3.3, we can expand the density $\rho(\theta, t)$ into Fourier series.

Theorem D.3.4. *For $p \geq 4$, if the initial distribution satisfies $\rho(\theta, 0) \in L^2_V(-\frac{\pi}{2}, \frac{\pi}{2})$, then the evolution of density function $\rho(\theta, t)$ over time is given by Fourier series in $L^2_V(-\frac{\pi}{2}, \frac{\pi}{2})$*

$$\rho(\theta, t) \propto \sum_{n\geq 0} T_n(t) \sum_{m\geq 0} c_{n,m}\Theta_{n,m}(\theta) \tag{134}$$

$$= c_{0,0}e^{-V(\theta)} + \sum_{n\geq 1} e^{-\frac{q_*^2}{2}\lambda_n t} \sum_{m\geq 0} c_{n,m}\Theta_{n,m}(\theta) \tag{135}$$

where $\Theta_{n,m}, m \in \mathbb{N}$ are the eigenfunctions corresponding to λ_n and $c_{n,m}$ are the coefficients determined by the initial condition

$$\rho(\theta, 0) = c_{0,0}e^{-V(\theta)} + \sum_{n\geq 1}\sum_{m\geq 0} c_{n,m}\Theta_{n,m}(\theta) \tag{136}$$

The infinite sums in equation 135 and equation 136 are interpreted as the limits in $L^2_V(-\frac{\pi}{2}, \frac{\pi}{2})$. $\Theta_{n_1,m}$ and $\Theta_{n_2,k}$ are orthogonal in $L^2_V(-\frac{\pi}{2}, \frac{\pi}{2})$ whenever $m \neq k$.

Corollary D.3.5 (Stationary distribution of θ_t)**.** *From equation 135 we define*

$$\rho_*(\theta) = c_{0,0}e^{-V(\theta)} \tag{137}$$

$$= c_{0,0}e^{-\kappa\cos 2\theta} \cdot \cos^{p-2}\theta, \quad \theta \in [\frac{\pi}{2}, \frac{\pi}{2}] \tag{138}$$

where κ follows the same definition in equation 109. Then ρ_* is the stationary distribution in the sense that

$$\lim_{t\to+\infty} \rho(\theta, t) = \rho_*(\theta), \quad \text{in } L^2_V(-\frac{\pi}{2}, \frac{\pi}{2}) \tag{139}$$

Specifically, we have

$$||\rho(\theta, t) - \rho_*(\theta)||^2_{L^2_V} \leq e^{-q_*^2\lambda_1 t}||\rho(\theta, 0) - \rho_*(\theta)||^2_{L^2_V} \tag{140}$$

Proof. Squaring both sides of initial condition equation 136 and taking integrals over $\theta \in (-\frac{\pi}{2}, -\frac{\pi}{2})$. For nonnegative measurable functions, we can use Tonelli's lemma to exchange the order of infinite sum and integral, and obtain

$$+\infty > ||\rho(\theta,t)||^2_{L^2_V} \tag{141}$$

$$= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\rho(\theta,t)|^2 e^{V(\theta)} d\theta \tag{142}$$

$$= \sum_{n,m} c^2_{n,m} a_{n,m}, \tag{143}$$

(orthogonality of eigenfunctions)

where

$$a_{n,m} = ||\Theta_{n,m}||^2_{L^2_V} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \Theta_{n,m}(\theta)^2 e^{V(\theta)} d\theta < +\infty \tag{144}$$

Then we have

$$||\rho(\theta,t) - \rho_*(\theta)||_{L^2_V} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} |\rho(\theta,t) - \rho_*(\theta)|^2 e^V d\theta \tag{145}$$

$$= \sum_{n\geq 1, m} e^{-q(\infty)^2 \lambda_n t} c^2_{n,m} a_{n,m} \tag{146}$$

$$\leq e^{-q_*^2 \lambda_1 t} \sum_{n\geq 1, m} c^2_{n,m} a_{n,m} \ \to 0 \tag{147}$$

$$= e^{-q_*^2 \lambda_1 t} ||\rho(\theta,t) - \rho(\theta,0)||_{L^2_V} \tag{148}$$

□

To further attain the stationary distribution of f_t, we change the variable with respect to the relation $f = \sin^2\theta$:

$$\rho(f,t) \propto \rho(\theta,t) \cdot |\frac{d\theta}{df}| \tag{149}$$

$$\propto \rho(\theta,t) \cdot |\frac{d(\pm \arcsin\sqrt{f})}{df}| \tag{150}$$

$$= \sum_{n\geq 0} T_n(t) \sum_{m\geq 0} b_{n,m} F_{n,m}(f), \quad f \in [0,1] \tag{151}$$

and specifically

$$\rho_*(f) \propto \rho_*(\theta) \cdot |\frac{d\theta}{df}| \tag{152}$$

$$\propto \rho_*(\theta) \cdot |\frac{d(\pm \arcsin\sqrt{f})}{df}| \tag{153}$$

$$\propto e^{-\kappa(1-2f)} (1-f)^{\frac{p-2}{2}} \cdot f^{-\frac{1}{2}} (1-f)^{-\frac{1}{2}} \tag{154}$$

$$\propto e^{2\kappa f} f^{-\frac{1}{2}} (1-f)^{\frac{p-3}{2}}, \quad f \in [0,1] \tag{155}$$

As is mentioned in Section B.2, in previous analysis we first assume that the NRQ matrix $\tilde{\boldsymbol{A}} = (0, 1, \cdots, 1)$. For a more general case with $\boldsymbol{A} = (a_l, a_h, \cdots, a_h)$, we define

$$\tilde{\mathcal{L}} = \frac{\boldsymbol{X}^T \mathrm{diag}(0, 1, \cdots, 1) \boldsymbol{X}}{2\boldsymbol{X}^T \boldsymbol{X}} \tag{156}$$

Hence

$$\mathcal{L} = a_l + (a_h - a_l)\tilde{\mathcal{L}} \tag{157}$$

and

$$d\boldsymbol{X}_t = -\eta(\nabla\mathcal{L}dt + \frac{P_{\boldsymbol{X}_t}\sigma}{||\boldsymbol{X}||}d\boldsymbol{B}_t) - \eta\lambda\boldsymbol{X}_t \tag{158}$$

$$= -(a_h - a_l)\eta\cdot(\nabla\tilde{\mathcal{L}}dt + \frac{P_{\boldsymbol{X}_t}}{||\boldsymbol{X}||}\frac{\sigma}{(a_h-a_l)}d\boldsymbol{B}_t) - (a_h-a_l)\eta\cdot\frac{\lambda}{a_h-a_l}\boldsymbol{X}_t dt \tag{159}$$

equation 159 indicates that the dynamics of NRQ with general matrix $\boldsymbol{A} = (a_l, a_h, \cdots, a_h)$ is equivalent to training with $\tilde{\boldsymbol{A}} = (0, 1, \cdots, 1)$ and rescaled learning rate, weight decay and noise scale:

$$\tilde{\eta} = (a_h - a_l)\eta \tag{160}$$

$$\tilde{\lambda} = \frac{\lambda}{a_h - a_l} \tag{161}$$

$$\tilde{\sigma^2} = \frac{\sigma^2}{(a_h-a_l)^2} \tag{162}$$

Substituting the above equivalence into equation 155, we obtain the following conclusion.

Corollary D.3.6 (Stationary distribution of f_t)**.** *The stationary distribution of f_t is given by*

$$\rho_*(f) = \frac{1}{\mathcal{N}(\kappa_{\boldsymbol{A}}, p)}\ e^{2\kappa_{\boldsymbol{A}}\cdot f} f^{-\frac{1}{2}}(1-f)^{\frac{p-3}{2}}, \quad f\in[0,1] \tag{163}$$

in which

$$\kappa_{\boldsymbol{A}} = \frac{\sqrt{p-1}}{2\tilde{\sigma}\sqrt{2\tilde{\eta}\tilde{\lambda}}} = \frac{(a_h-a_l)\sqrt{p-1}}{2\sigma\sqrt{2\eta\lambda}} \tag{164}$$

$\mathcal{N}(\kappa_{\boldsymbol{A}}, p)$ is the normalizing constant that equals to

$$\mathcal{N}(\kappa_{\boldsymbol{A}}, p) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{p-1}{2})}{\Gamma(\frac{p}{2})}\mathcal{M}(\frac{1}{2}, \frac{p}{2}, 2\kappa_{\boldsymbol{A}}) \tag{165}$$

where $\Gamma(z)$ is the gamma function, and $\mathcal{M}(a,b,z)$ is Kummer's confluent hypergeometric function.
Furthermore, the equilibrium expectation is given by

$$\mathbb{E}f_* = \frac{1}{p}\frac{\mathcal{M}(\frac{3}{2}, \frac{p}{2}+1, 2\kappa_{\boldsymbol{A}})}{\mathcal{M}(\frac{1}{2}, \frac{p}{2}, 2\kappa_{\boldsymbol{A}})} \tag{166}$$

$$= 1 - \frac{\sqrt{p-1}\sigma}{a_h-a_l}\sqrt{2\eta\lambda} + o(\sqrt{2\eta\lambda}) \tag{167}$$

Proof. For Kummer's confluent hypergeometric function, we have

$$\mathcal{M}(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)}\int_0^1 e^{zu}u^{a-1}(1-u)^{b-a-1}du \tag{168}$$

whenever $Re(b) > Re(a) > 0$. Hence

$$\mathcal{N}(\kappa_{\boldsymbol{A}}, p) = \int_0^1 e^{2\kappa_{\boldsymbol{A}}f} f^{-\frac{1}{2}}(1-f)^{\frac{p-3}{2}} \tag{169}$$

$$= \frac{\Gamma(\frac{1}{2})\Gamma(\frac{p-1}{2})}{\Gamma(\frac{p}{2})}\mathcal{M}(\frac{1}{2}, \frac{p}{2}, 2\kappa_{\boldsymbol{A}}) \tag{170}$$

$$\mathbb{E}f_* = \int_0^1 e^{2\kappa_{\boldsymbol{A}}f} f^{\frac{1}{2}}(1-f)^{\frac{p-3}{2}}/\mathcal{N}(\kappa_{\boldsymbol{A}}, p) \tag{171}$$

$$= \frac{\Gamma(\frac{3}{2})\Gamma(\frac{p-1}{2})}{\Gamma(\frac{p}{2}+1)}\mathcal{M}(\frac{3}{2}, \frac{p}{2}+1, 2\kappa_{\boldsymbol{A}})/\mathcal{N}(\kappa_{\boldsymbol{A}}, p) \tag{172}$$

$$= \frac{1}{p}\frac{\mathcal{M}(\frac{3}{2}, \frac{p}{2}+1, 2\kappa_{\boldsymbol{A}})}{\mathcal{M}(\frac{1}{2}, \frac{p}{2}, 2\kappa_{\boldsymbol{A}})} \tag{173}$$

Further we have

$$\mathcal{M}(a,b,z) = e^{z^2} z^{a-b} \frac{\Gamma(b)}{\Gamma(a)} [1 + \frac{(a-1)(a-b)}{z} + o(\frac{1}{z})] \tag{174}$$

and therefore

$$\mathbb{E} f_* = \frac{1 - \frac{p-1}{4} \frac{1}{2\kappa_{\boldsymbol{A}}}}{1 + \frac{p-1}{4} \frac{1}{2\kappa_{\boldsymbol{A}}}} + o(\frac{1}{2\kappa_{\boldsymbol{A}}}) = 1 - \frac{\sqrt{p-1}\sigma}{a_h - a_l} \sqrt{2\eta\lambda} + o(\sqrt{2\eta\lambda}) \tag{175}$$

□

D.4 Estimation of tail probability decay

From the expansion in Corollary D.3.4, we see that the convergence behaviors of the dynamics are governed by the lowest non-vanishing eigenvalue of the Fokker Planck equation $\mu_1 = \frac{q_*^2}{2}\lambda_1$.

Corollary D.4.1 (Linear decay of tail probability)**.** *Given any $\xi \in (0,1)$, $\varepsilon(t) = \mathbb{P}(f_t < \xi)$ is defined as the tail probability of f_t dynamics. It represents the probability that the trajectory of $\tilde{\boldsymbol{X}}_t$ stays outside certain neighbourhood of the optimal solution $\boldsymbol{e}_1$, and can be estimated with*

$$|\varepsilon(t) - \varepsilon_*| \leq Ce^{-\mu_1 t} \tag{176}$$

where C is a positive constant irrelevant of ξ.

Proof. Let $\delta = \arcsin\sqrt{\xi}$. By Cauchy-Schwartz inequality and the L_V^2 convergence of $\rho(\theta,t)$ to its stationary density in Corollary D.3.5, we have

$$|\varepsilon(t) - \varepsilon_*| = \left| \int_0^1 I_{(-\delta,\delta)} [\rho(\theta,t) - \rho_*(\theta)] df \right| \tag{177}$$

$$\leq \int_0^1 I_{(-\delta,\delta)} |\rho(\theta,t) - \rho_*(\theta)| df \tag{178}$$

$$\leq \left\{ \int_0^1 e^{-V} d\theta \right\}^{\frac{1}{2}} \cdot \left\{ \int_0^1 |\rho(\theta,t) - \rho_*(\theta)|^2 e^V d\theta) \right\}^{\frac{1}{2}} \tag{179}$$

$$\leq C_1 \cdot ||\rho(\theta,t) - \rho_*(\theta)||_{L_V^2} \cdot e^{-\mu_1 t}, \quad \text{by equation 140} \tag{180}$$

$$\to 0 \tag{181}$$

□

It is generally impossible to reach an exact expression of $\mu_1 = \frac{q_*^2}{2}\lambda_1$, but certain approximation can be made to estimate this value, as is introduced in Risken (1996).

Lemma D.4.2. *If $\kappa > \frac{p-2}{4}$, the normalized potential V is a symmetric double well potential, with local minima $\pm \arccos\sqrt{\frac{p-2}{4\kappa}}$ and local maxima 0.*

Proof. We have

$$V(-\theta) = \kappa \cos(-2\theta) + (p-2)\ln\sec(-\theta) = V(\theta) \tag{182}$$

which proves the symmetry. Also

$$\frac{d}{d\theta} V(\theta) = \tan\theta(p - 2 - 4\kappa\cos^2\theta) \tag{183}$$

proves the conclusion on critical points. □

Lemma D.4.3. *Given a symmetric double well potential $U(x)$ on finite interval $[-A, A]$ with local minima $\pm a$, $0 < a < A$. Further suppose reflecting within the potential is a related diffusion process*

$$dX_t = -\frac{1}{2}\nabla U(x)dt + \sqrt{D}dB_t \tag{184}$$

Then the lowest non-vanishing eigenvalue of the corresponding Fokker Planck equation can be approximated with

$$\hat{\mu}_1 = \frac{1}{\pi}\sqrt{|U''(0)U''(a)|}e^{-[U(0)-U(a)]/D} \tag{185}$$

$$= \frac{D}{\pi}\sqrt{|V''(0)V''(a)|}e^{-[V(0)-V(a)]} \tag{186}$$

in which $V = U/D$ is the normalized potential.

Remark 1. One should be cautious that the μ_1 in Lemma D.4.3 refers to the lowest non-vanishing eigenvalue of the original Fokker Planck equation, while λ_1 in D.3.4 is the lowest non-vanishing eigenvalue of the variable-separated boundary value problem. This means $\mu_1 = \frac{q(\infty)^2}{2}\lambda_1$.

Theorem D.4.4 (Estimation of μ_1)**.** *If $\kappa > \frac{p-2}{4}$, the lowest non-vanishing eigenvalue $\mu_1 = \frac{q_*^2}{2}\lambda_1$ of original Fokker Planck equation equation 102 can be approximated by*

$$\hat{\mu}_1 = \frac{4\sqrt{2}q(\infty)^2}{\pi}\kappa(1-\frac{p-2}{4\kappa})e^{-2\kappa(1-\frac{p-2}{4\kappa}+\frac{p-2}{4\kappa}\ln\frac{p-2}{4\kappa})} \tag{187}$$

Proof. Since

$$V(\theta) = \kappa\cos 2\theta + (p-2)\ln\sec\theta \tag{188}$$

$$V''(\theta) = -4\kappa\cos 2\theta + (p-2)\sec^2\theta \tag{189}$$

$$a = \arccos\sqrt{\frac{p-2}{4\kappa}} \tag{190}$$

we have the following calculations

$$V(0) = \kappa \tag{191}$$

$$V(a) = -\kappa - \frac{p-2}{2}\ln\frac{p-2}{4\kappa} \tag{192}$$

$$V''(0) = -4\kappa + p - 2 \tag{193}$$

$$V''(a) = 8\kappa - 2(p-2) \tag{194}$$

Then plunging equation 191 to equation 194 into equation 186, we complete the proof. □

Recall that under the equilibrium of SMD, the angular update (AU) is of fixed magnitude $\Delta = \sqrt{2\eta\lambda}$. Since $\kappa \propto C(p,\sigma^2)\Delta^{-1}$, the condition $\kappa > \frac{p-2}{4}$ means that the angular update Δ is rather small. This is generally true for commonly-adopted settings such as $\eta = 0.1,\ \lambda = 0.001$. For this reason, we make a further assumption

Assumption D.4.5. *In commonly adopted settings, $\Delta \ll 1$ so that $\frac{p-2}{4\kappa} \ll 1$.*

We can simplify the expression equation 187 with the above assumption.

Corollary D.4.6 (Small AU approximation of μ_1)**.** *Under Assumption D.4.5, the lowest non-vanishing eigenvalue of Fokker Planck equation can be approximated by*

$$\hat{\mu}_1 \approx C_1\Delta\cdot e^{-\frac{C_2}{\Delta}} \tag{195}$$

in which C_1 and C_2 are two constants depending only on dimension p and noise scale σ^2.

Proof. Since $q_* \propto \Delta$, $\kappa \propto \Delta^{-1}$ and $\frac{p-2}{4\kappa} = o(1)$, from equation 187 we have

$$\hat{\mu}_1 = C_1\Delta(1-B_1\Delta)\cdot e^{-\frac{C_2}{\Delta}[1-B_2\Delta+B_3\Delta\ln(B_3\Delta)]} \tag{196}$$

$$= C_1\Delta\cdot e^{-\frac{C_2}{\Delta}} + o(\Delta\cdot e^{-\frac{C_2}{\Delta}}) \tag{197}$$

□

The above corollary is crucial in the sense that it provides estimation of the decay rate of the density $\rho(f,t)$ together with its tail probability $\varepsilon(t)$.

E Proof of the results beyond equilibrium state dynamics

E.1 Escaping behavior

From the general bounds in Theorem C.0.3 it is straightforward to derive the following conclusion for the escaping behavior in non-equilibrium state.

Corollary E.1.1 (A sufficient condition for "escaping" behavior)**.** *When η, λ are given, if the following conditions hold:*

1) $\exists \varepsilon > 0$, $\forall t > 0, \varepsilon(t) < \varepsilon < 1 - \bar{f}^*$;

2) $f_0 = \bar{f}^*$;

3) $M(0) > \frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi}$;

where $\bar{f}^*$ is defined as in Corollary D.1.1. Then $\exists T > 0$, we have

$$\mathbb{E}f_T > f_0 \geq \overline{\lim_{t\to\infty}} \mathbb{E}f_t. \tag{198}$$

Proof. First of all, let's briefly demonstrate why $f_0 = \bar{f}^* \geq \overline{\lim}_{t\to\infty} \mathbb{E}f_t$: Acoording to the evolution of $M(t)$ shown in equation 29, $M(t)$ will convrege to M^* as $t \to +\infty$. Then by equation 87, when equilibrium has been achieved, as $t \to +\infty$, we have $\mathbb{E}f_t \leq \bar{f}^*$, hence $\overline{\lim}_{t\to\infty} \mathbb{E}f_t \leq \bar{f}^*$.

Now let's prove $\exists T > 0, \mathbb{E}f_T > \bar{f}^*$. First let's show $M^* < \frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi}$:

$$\begin{aligned}
\frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi} =& \frac{(p-1)\sigma^2\eta}{(\frac{\Delta}{\frac{a_p-a_1}{\sqrt{p-1}\sigma}+\frac{p}{p-1}\Delta} - \varepsilon)(a_2-a_1)\xi} \\
>& \frac{(p-1)\sigma^2\eta}{(\frac{\Delta}{\frac{a_p-a_1}{\sqrt{p-1}\sigma}})(a_2-a_1)\xi} \\
=& \frac{a_p-a_1}{(a_2-a_1)\xi} \cdot \frac{\sqrt{p-1}\sigma\eta}{\Delta} \\
=& \frac{a_p-a_1}{(a_2-a_1)\xi} M^* > M^*
\end{aligned} \tag{199}$$

which means $M(0) > \frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi} > M^*$. Recall $M(t)$ will continously and monotonously converge to M^* as equation 29 implies, hence $M(t)$ will strictly decrease to M^*, and $\exists T > 0, M(T) = \frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi}$. Then by equation 33, we have

$$\begin{aligned}
\mathbb{E}f_T \geq& e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \tilde{g}_2(\tau)e^{\tilde{G}_1(\tau)}d\tau], \\
=& e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \frac{\tilde{g}_2(\tau)}{\tilde{g}_1(\tau)}\tilde{g}_1(\tau)e^{\tilde{G}_1(\tau)}d\tau].
\end{aligned} \tag{200}$$

Let's estimate the bound of $\frac{\tilde{g}_2(\tau)}{\tilde{g}_1(\tau)}$ on $[0,T]$: let $\Delta(t)$ denote $\frac{\sqrt{p-1}\sigma\eta}{M(t)}$, so $\Delta(t)$ will increase from $\Delta(0)$ to $\Delta(T)$ on $[0,T]$, then we have

$$\begin{aligned}
\frac{\tilde{g}_2(\tau)}{\tilde{g}_1(\tau)} =& \frac{\frac{(a_2-a_1)\eta\xi}{M(t)}(1-\varepsilon(t)) + \frac{\eta^2\sigma^2}{M(t)^2}}{\frac{(a_2-a_1)\eta\xi}{M(t)} + \frac{p\eta^2\sigma^2}{M(t)^2}} \\
=& \frac{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}\Delta(t)(1-\varepsilon(t)) + \frac{1}{p-1}\Delta(t)^2}{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}\Delta(t) + \frac{p}{p-1}\Delta(t)^2} \\
=& 1 - \frac{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}\varepsilon(t) + \Delta(t)}{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta(t)} \\
>& 1 - \frac{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}\varepsilon + \Delta(t)}{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma} + \frac{p}{p-1}\Delta(t)} \\
>& 1 - \frac{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}\varepsilon + \Delta(t)}{\frac{(a_2-a_1)\xi}{\sqrt{p-1}\sigma}} \\
=& 1 - \varepsilon - \frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}\Delta(t) \\
\geq& 1 - \varepsilon - \frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}\Delta(T)
\end{aligned} \tag{201}$$

Take equation 201 into equation 200, we have

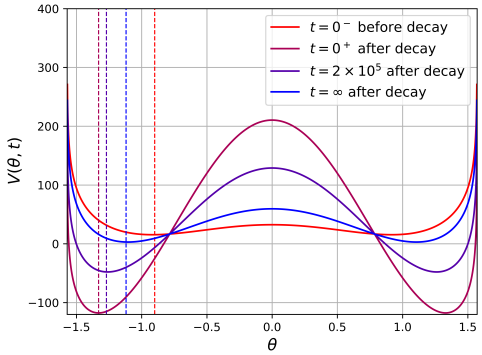
$$\begin{aligned}
\mathbb{E}f_T =& e^{-\tilde{G}_1(T)}[f_0 + \int_0^T \frac{\tilde{g}_2(\tau)}{\tilde{g}_1(\tau)}\tilde{g}_1(\tau)e^{\tilde{G}_1(\tau)}d\tau] \\
>& e^{-\tilde{G}_1(T)}\{f_0 + \int_0^T [1-\varepsilon - \frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}\Delta(T)]\tilde{g}_1(\tau)e^{\tilde{G}_1(\tau)}d\tau\} \\
=& e^{-\tilde{G}_1(T)}\{f_0 + [1-\varepsilon - \frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}\Delta(T)]\int_0^T \tilde{g}_1(\tau)e^{\tilde{G}_1(\tau)}d\tau\} \\
=& e^{-\tilde{G}_1(T)}\{f_0 + [1-\varepsilon - \frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}\Delta(T)](e^{\tilde{G}_1(T)} - 1)\}.
\end{aligned} \tag{202}$$

Note $M(T) = \frac{(p-1)\sigma^2\eta}{(1-\varepsilon-\bar{f}^*)(a_2-a_1)\xi}$, so $\Delta(T) = \frac{\sqrt{p-1}\sigma\eta}{M(T)} = (1-\varepsilon-\bar{f}^*)/\frac{\sqrt{p-1}\sigma}{(a_2-a_1)\xi}$, besides $f_0 = \bar{f}^*$, hence we have

$$\mathbb{E}f_T > e^{-\tilde{G}_1(T)}\{\bar{f}^* + [1-\varepsilon-(1-\varepsilon-\bar{f}^*)](e^{\tilde{G}_1(T)} - 1)\} = \bar{f}^*. \tag{203}$$

□

A supplementary qualitative explanation of this escaping behavior (or "pseudo-overfitting") can be derived from the view of back-and-forth distribution shift. In the Fokker-Planck equation equation 102 with respect to $\theta_t = \arcsin(\boldsymbol{e}_1^T\tilde{\boldsymbol{X}}_t)$, we have made the diffusion term space-homogeneous. Therefore, the dynamics can be regarded as the motion of a particle moving in a potential well $U(\theta,t)$ which changes with time. We plot below in Figure 1 the normalized potential well $V(\theta,t) = U(\theta,t)/q(t)^2$, the minima of which is indicated with a dashed line. It is clear from Figure 1 that at the moment when the learning rate get decayed, the double valleys of the potential soon jump away from each other towards $\pm\frac{\pi}{2}$ respectively. Remember that the x label of the figure is $\theta_t = \arcsin(\boldsymbol{e}_1^T\tilde{\boldsymbol{X}}_t)$. Therefore, this sudden shift is in favor of lower risk $r_t = (\boldsymbol{e}_1^T\tilde{\boldsymbol{X}}_t)^2$ and hence lower loss L_t. As the training goes on, however, the double valleys gradually moves closer to each other towards 0, which induces higher risk and loss and leads to the so-called "pseudo-overfitting" phenomenon.



(a) Double-well potential $V(\theta, t)$

Figure 1: Back-and-forth shift of potential well. Each dashed line indicates the left minima of the corresponding double-well potential.

E.2 Equivalent dynamics with adjusted hyperparameters

Corollary E.2.1. *$\forall k > 0$, if $\boldsymbol{X}_0$ is multiplied by k, enlarge η, λ by k^2, $\frac{1}{k^2}$ times respectively, r_t remains unchanged.*

Proof. Denote the new dynamics by $\boldsymbol{X}_t^{(2)}$. The adjustment implies

$$\frac{1}{k^2} M_0^{(2)} = M_0 \tag{204}$$

$$\frac{1}{k^2} \eta^{(2)} = \eta \tag{205}$$

$$k^2 \lambda^{(2)} = \lambda \tag{206}$$

Then from equation 9 we have

$$\begin{aligned} d\tilde{\boldsymbol{X}}_t^{(2)} =& -[\frac{\eta^{(2)}}{M_t^{(2)}} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \boldsymbol{A} \tilde{\boldsymbol{X}}_t^{(2)} + \frac{(\eta^{(2)})^2}{2(M_t^{(2)})^2} \operatorname{Tr}(\boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}}) \tilde{\boldsymbol{X}}^{(2)}] dt \\ & - \frac{\eta^{(2)}}{M_t^{(2)}} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}} d\boldsymbol{B}_t \end{aligned} \tag{207}$$

$$dM_t^{(2)} = [-2\lambda^{(2)} \eta^{(2)} M_t^{(2)} + \frac{(\eta^{(2)})^2}{M_t^{(2)}} \operatorname{Tr}(\boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}})] dt \tag{208}$$

that is

$$\begin{aligned} d\tilde{\boldsymbol{X}}_t^{(2)} =& -[\frac{\eta}{M_t^{(2)}/k^2} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \boldsymbol{A} \tilde{\boldsymbol{X}}_t^{(2)} + \frac{\eta^2}{2(M_t^{(2)}/k^2)^2} \operatorname{Tr}(\boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}}) \tilde{\boldsymbol{X}}^{(2)}] dt \\ & - \frac{\eta}{M_t^{(2)}/k^2} \boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}} d\boldsymbol{B}_t \end{aligned} \tag{209}$$

$$dM_t^{(2)}/k^2 = [-2\lambda\eta M_t^{(2)}/k^2 + \frac{\eta^2}{M_t^{(2)}/k^2} \operatorname{Tr}(\boldsymbol{P}_{\tilde{\boldsymbol{X}}_t^{(2)}} \tilde{\boldsymbol{\Sigma}})] dt \tag{210}$$

Therefore, $(\tilde{\boldsymbol{X}}_t, M_t)$ and $(\tilde{\boldsymbol{X}}_t^{(2)}, M_t^{(2)}/k^2)$ satisfies the same SDE, with identical initial condition

$$\tilde{\boldsymbol{X}}_0 = \tilde{\boldsymbol{X}}_0^{(2)} \tag{211}$$

$$M_0 = M_0^{(2)}/k^2 \tag{212}$$

Hence their dynamics will be identical. Since the risk r_t is entirely determined by the dynamics of $\tilde{\boldsymbol{X}}_t$, we claim that the aforementioned adjustment of hyperparameters will preserve the evolution of risk. □

References

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2018.

Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. *Advances in Neural Information Processing Systems*, 31:7694–7705, 2018.

Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. In *International Conference on Machine Learning*, pp. 882–890. PMLR, 2019.

Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32:8433–8443, 2019.

Yonatan Dukler, Quanquan Gu, and Guido Montúfar. Optimization theory for relu neural networks trained with normalization layers. In *International conference on machine learning*, pp. 2751–2760. PMLR, 2020.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 2164–2174, 2018.

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 806–815. PMLR, 2019.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2019.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=goEdyJ_nVQI`.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pp. 7045–7056. PMLR, 2021.

Vladimir Maz'ya. Analytic criteria in the qualitative spectral analysis of the... *Spectral Theory and Mathematical Physics: a Festschrift in honor of Barry Simon's 60th birthday*, 76:257, 2007.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pp. 63–95. Springer, 1996.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29:901–909, 2016.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Proceedings of the 32nd international conference on neural information processing systems*, pp. 2488–2498, 2018.

Barry Simon. Schrodinger operators with purely discrete spectrum. *arXiv preprint arXiv:0810.3275*, 2008.

Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. *Advances in Neural Information Processing Systems*, 34, 2021.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018a.

Xiaoxia Wu, Rachel Ward, and Léon Bottou. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018b.

Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization and convergence for weight normalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2018.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32:8196–8207, 2019.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. 2019.