Appendix

A. Proofs for Convergence under Gaussian Noise (Theorem 1)

A.1. Proof Overview

The main proof of Theorem 1 is contained in Appendix A.4.

Here, we outline the steps of our proof:

1. In Appendix A.2, we construct a coupling between (3) and (2) over a single step (i.e. for $t \in [k\delta, (k+1)\delta]$, for some k and δ).
2. Appendix A.3, we prove Lemma 1, which shows that under the coupling constructed in Step 1, a Lyapunov function $f(x_T - y_T)$ contracts exponentially with rate λ, plus a discretization error term. The function f is defined in Appendix E, and sandwiches $\|x_T - y_T\|_2$. In Corollary 2, we apply the results of Lemma 1 recursively over multiple steps to give a bound on $f(x_{k\delta} - y_{k\delta})$ for all k, and for sufficiently small δ.
3. Finally, in Appendix A.4, we prove Theorem 1 by applying the results of Corollary 2, together with the fact that $f(z)$ upper bounds $\|z\|_2$ up to a constant factor.

A.2. A coupling construction

In this subsection, we will study the evolution of (3) and (2) over a small time interval. Specifically, we will study

$$dx_t = -\nabla U(x_t)dt + M(x_t)dB_t \tag{20}$$
$$dy_t = -\nabla U(y_0)dt + M(y_0)dB_t \tag{21}$$

One can verify that (20) is equivalent to (3), and (21) is equivalent to a single step of (2) (i.e. over an interval $t \le \delta$).

We first give the explicit coupling between (20) and (21): (A similar coupling in the continuous-time setting is first seen in (Gorham et al., 2016) in their proof of contraction of (3).)

Given arbirary (x_0, y_0), define (x_t, y_t) using the following coupled SDE:

$$\begin{aligned} x_t &= x_0 + \int_0^t -\nabla U(x_s)ds + \int_0^t c_m dV_s + \int_0^t N(x_s)dW_s \\ y_t &= y_0 + \int_0^t -\nabla U(y_0)dt + \int_0^t c_m\big(I - 2\gamma_s\gamma_s^T\big)dV_s + \int_0^t N(y_0)dW_s \end{aligned} \tag{22}$$

Where dV_t and dW_t are two independent standard Brownian motion, and

$$\gamma_t := \frac{x_t - y_t}{\|x_t - y\|_2} \cdot \mathbb{1}\left\{\|x_t - y_t\|_2 \in [2\epsilon, \mathcal{R}_q)\right\} \tag{23}$$

By Lemma 6, we show that (20) has the same distribution as x_t in (22), and (21) has the same distribution as y_t in (22). Thus, for any t, the process (x_t, y_t) defined by (22) is a valid coupling for (20) and (21).

A.3. One step contraction

Lemma 1 *Let f be as defined in Lemma 18 with parameters ϵ satisfying $\epsilon \le \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$. Let x_t and y_t be as defined in (22). If we assume that $\mathbb{E}\left[\|y_0\|_2^2\right] \le 8\big(R^2 + \beta^2/m\big)$ and $T \le \min\left\{\frac{\epsilon^2}{\beta^2}, \frac{\epsilon}{6L\sqrt{R^2+\beta^2/m}}\right\}$, then*

$$\mathbb{E}\left[f(x_T - y_T)\right] \le e^{-\lambda T}\mathbb{E}\left[f(x_0 - y_0)\right] + 3T(L + L_N^2)\epsilon$$

Remark 8 *For ease of reference: m, L, L_R, R are from Assumption A, c_m, β are from Assumption B, $\alpha_q, \mathcal{R}_q, L_N, \lambda$ are defined in (7).*

Proof of Lemma 1

For notational convenience, for the rest of this proof, let us define $z_t := x_t - y_t$ and $\nabla_t := \nabla U(x_t) - \nabla U(y_t)$, $\Delta_t := \nabla U(y_0) - \nabla U(y_t)$ $N_t := N(x_t) - N(y_t)$.

It follows from (22) that

$$dz_t = -\nabla_t dt + \Delta_t dt + 2c_m \gamma_t \gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t \tag{24}$$

Using Ito's Lemma, the dynamics of $f(z_t)$ is given by

$$\begin{aligned}
& df(z_t) \\
=& \langle \nabla f(z_t), dz_t\rangle + 2c_m^2 \text{tr}\left(\nabla^2 f(z_t)\left(\gamma_t\gamma_t^T\right)\right)dt + \frac{1}{2}\text{tr}\left(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\right)dt \\
=& \underbrace{-\langle \nabla f(z_t), \nabla_t\rangle}_{①} dt + \underbrace{\langle \nabla f(z_t), \Delta_t\rangle}_{②} dt + \underbrace{\left\langle \nabla f(z_t), 2c_m\gamma_t\gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t\right\rangle}_{③} \\
& + \underbrace{2c_m^2 \text{tr}\left(\nabla^2 f(z_t)\left(\gamma_t\gamma_t^T\right)\right)}_{④} dt + \underbrace{\frac{1}{2}\text{tr}\left(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\right)}_{⑤} dt
\end{aligned} \tag{25}$$

③ goes to 0 when we take expectation, so we will focus on ①, ②, ④, ⑤. We will consider 3 cases

Case 1: $\|z_t\|_2 \le 2\epsilon$

From item 1(c) of Lemma 18, $\|\nabla f(z)\|_2 \le 1$. Using Assumption A.1, $\|\nabla_t\| \le L\|z_t\|_2$, so that

$$① \le \|\nabla_t\|_2 \le L\|z_t\|_2 \le 2L\epsilon$$

Also by Cauchy Schwarz,

$$② = \langle \nabla f(z_t), \Delta_t\rangle \le \|\Delta_t\|_2 \le L\|y_t - y_0\|_2$$

Since $\gamma_t = 0$ in this case by definition in (23), ④ $= 0$.

Using Lemma 18.2.c. $\left\|\nabla^2 f(z_t)\right\|_2 \le \frac{2}{\epsilon}$, so that

$$\begin{aligned}
⑤ \le& \frac{1}{\epsilon}\left(\text{tr}\left(N_t^2 + N(y_t) - N(y_0)\right)^2\right) \\
\le& \frac{2}{\epsilon}\left(\text{tr}\left(N_t^2\right) + \text{tr}\left((N(y_t) - N(y_0))^2\right)\right) \\
\le& \frac{2L_N^2}{\epsilon}\left(\|z_t\|_2^2 + \|y_t - y_0\|_2^2\right) \\
\le& 4L_N^2\epsilon + \frac{2L_N^2}{\epsilon}\|y_t - y_0\|_2^2
\end{aligned}$$

Where the second inequality is by Young's inequality, the third inequality is by item 2 of Lemma 16, the fourth inequality is by our assumption that $\|z_t\|_2 \le 2\epsilon$.

Summing these,

$$① + ② + ④ + ⑤ \le 4\left(L + L_N^2\right)\epsilon + L\|y_t - y_0\|_2 + \frac{2L_N^2}{\epsilon}\|y_t - y_0\|_2^2$$

Case 2: $\|z_t\|_2 \in (2\epsilon, \mathcal{R}_q)$

In this case, $\gamma_t = \frac{z_t}{\|z_t\|_2}$. Let q be as defined in (39) and g be as defined in Lemma 20. By items 1(b) and 2(b) of Lemma 18 and items 1(b) and 2(b) of Lemma 20,

$$\begin{aligned}
\nabla f(z_t) =& q'(g(z_t))\nabla g(z_t) \\
=& q'(g(z_t))\frac{z_t}{\|z_t\|_2} \\
\nabla^2 f(z_t) =& q''(g(z_t))\nabla g(z_t)\nabla g(z_t)^T + q'(g(z_t))\nabla^2 g(z_t) \\
=& q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)
\end{aligned}$$

Once again, by Assumption A.3,

$$① \leq q'(g(z_t))\|\nabla_t\|_2 \leq q'(g(z_t)) \cdot L_R \cdot \|z_t\|_2 \leq L \cdot q'(g(z_t))g(z_t) + 2L\epsilon$$

Where the last inequality uses Lemma 20.4. We can also verify that

$$② \leq L\|y_t - y_0\|_2$$

Using the expression for $\nabla^2 f(z_t)$,

$$④ = 2c_m^2 \mathrm{tr}\left(\nabla^2 f(z_t)\gamma_t\gamma_t^T\right) = 2c_m^2 \cdot q''(g(z_t))$$

Finally,

$$\begin{aligned}
⑤ =& \frac{1}{2}\mathrm{tr}\Big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\Big)\\
=& \frac{1}{2}\mathrm{tr}\left(\left(q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right)\\
\leq& \frac{1}{2}\mathrm{tr}\left(\left(q'(g(z_t))\frac{1}{\|z_t\|_2}\left(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\right)\right)(N_t + N(y_t) - N(y_0))^2\right)\\
\leq& \frac{q'(g(z_t))}{\|z_t\|_2} \cdot \Big(\mathrm{tr}\big(N_t^2\big) + \mathrm{tr}\Big((N(y_t) - N(y_0))^2\Big)\Big)\\
\leq& q'(g(z_t)) \cdot L_N^2\|z_t\|_2 + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon}\\
\leq& q'(g(z_t)) \cdot L_N^2 g(z_t) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon} + 2L_N^2\epsilon
\end{aligned}$$

The above uses multiples times the fact that $0 \leq q' \leq 1$ and $q'' \leq 0$ (proven in items 3 and 4 of Lemma 21). The second inequality is by Young's inequality, the third inequality is by item 2 of Lemma 16, the fourth inequality uses item 4 of Lemma 20.

Summing these,

$$\begin{aligned}
① + ② + ④ + ⑤ \leq& \big(L_R + L_N^2\big)q'(g(z_t))g(z_t) + 2c_m^2 q''(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon} + 2\big(L + L_N^2\big)\epsilon\\
\leq& -\frac{2c_m^2 \exp\left(-\frac{7\alpha_q\mathcal{R}_q{}^2}{3}\right)}{32\mathcal{R}_q{}^2}q(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon} + 2\big(L + L_N^2\big)\epsilon\\
\leq& -\lambda q(g(z_t)) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon} + 2(L + L_N^2)\epsilon\\
=& -\lambda f(z_t) + \frac{L_N^2\|y_t - y_0\|_2^2}{2\epsilon} + 2(L + L_N^2)\epsilon + L\|y_t - y_0\|_2
\end{aligned}$$

Where the last inequality follows from Lemma 21.1. and the definition of λ in (7).

Case 3: $\|z_t\|_2 \geq \mathcal{R}_q$
In this case, $\gamma_t = 0$. Similar to case 2,

$$\nabla f(z_t) = q'(g(z_t))\frac{z_t}{\|z_t\|_2}$$

Thus by Assumption A.3,

$$\begin{aligned}
① =& \left\langle q'(g(z_t))\frac{z_t}{\|z_t\|_2}, -\nabla_t\right\rangle\\
\leq& -mq'(g(z_t))\|z_t\|_2
\end{aligned}$$

Where the inequality is by Assumption A.3.

For identical reasons as in Case 1, ② $\leq L_R\|y_t - y_0\|_2$, and ④ $= 0$. Finally,

$$\begin{aligned}
⑤ =& \frac{1}{2}\text{tr}\Big(\nabla^2 f(z_t)(N_t + N(y_t) - N(y_0))^2\Big) \\
=& \frac{1}{2}\text{tr}\bigg(\bigg(q''(g(z_t))\frac{z_t z_t^T}{\|z_t\|_2^2} + q'(g(z_t))\frac{1}{\|z_t\|_2}\bigg(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\bigg)\bigg)(N_t + N(y_t) - N(y_0))^2\bigg) \\
\leq& \frac{1}{2}\text{tr}\bigg(\bigg(q'(g(z_t))\frac{1}{\|z_t\|_2}\bigg(I - \frac{z_t z_t^T}{\|z_t\|_2^2}\bigg)\bigg)(N_t + N(y_t) - N(y_0))^2\bigg) \\
\leq& \frac{q'(g(z_t))}{\|z_t\|_2} \cdot \Big(\text{tr}\big(N_t^2\big) + \text{tr}\Big((N(y_t) - N(y_0))^2\Big)\Big)
\end{aligned}$$

Where the first inequality is because $q'' \leq 0$ from item 4 of Lemma 21, the second inequality is by Young's inequality. (These steps are identical to Case 2). Continuing from above, and using item 2 and 3 of Lemma 16,

$$\begin{aligned}
⑤ \leq& q'(g(z_t)) \cdot \bigg(\frac{8\beta^2 L_N}{c_m} + \frac{L_N^2\|y_t - y_0\|_2^2}{\epsilon}\bigg) \\
\leq& q'(g(z_t)) \cdot \Big(\frac{m}{2}\|z_t\|_2\Big) + q'(g(z_t)) \cdot \bigg(\frac{L_N^2\|y_t - y_0\|_2^2}{\epsilon}\bigg)
\end{aligned}$$

Where the second inequality is by our definition of $\mathcal{R}_q$ in the Lemma statement, which ensures that $\frac{8\beta^2 L_N}{c_m} \leq \frac{m}{2}\mathcal{R}_q \leq \frac{m}{2}\|z_t\|_2$.

Thus

$$\begin{aligned}
&①+②+④+⑤ \\
\leq& -mq'(g(z_t))\|z_t\|_2 + L_R\|y_t - y_0\|_2 + \frac{m}{2}q'(g(z_t))\|z_t\|_2 + q'(g(z_t)) \cdot \bigg(\frac{L_N^2\|y_t - y_0\|_2^2}{\epsilon}\bigg) \\
\leq& -\frac{m}{2}q'(g(z_t))\|z_t\|_2 + \frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2 \\
\leq& -\lambda f(z_t) + \frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2
\end{aligned}$$

where the second inequality uses $q' \leq 1$ from item 3 of Lemma 21, the third inequality uses our definition of λ in (7).

Combining the three cases, (25) can be upper bounded with probability 1:

$$df(z_t) \leq -\lambda f(z_t) + \frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2 + \big\langle \nabla f(z_t), 2c_m\gamma_t\gamma_t^T dV_t + (N_t + N(y_t) - N(y_0))dW_t \big\rangle$$

To simplify notation, let us define $G_t \in \mathbb{R}^{1\times 2d}$ as $G_t := \big[\nabla f(z_t)^T 2c_m\gamma_t\gamma_t^T, \nabla f(z_t)^T(N_t + N(y_t) - N(y_0))\big]$, and let A_t be a $2d$-dimensional Brownian motion from concatenating $A_t = \begin{bmatrix} V_t \\ W_t \end{bmatrix}$. Thus

$$df(z_t) \leq -\lambda f(z_t)dt + \bigg(\frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2\bigg) + G_t dA_t.$$

We will study the Lyapunov function

$$\mathcal{L}_t := f(z_t) - \int_0^t e^{-\lambda(t-s)}\bigg(\frac{L_N^2}{\epsilon}\|y_s - y_0\|_2^2 + L\|y_s - y_0\|_2\bigg)ds - \int_0^t e^{-\lambda(t-s)} G_s dA_s.$$

By taking derivatives, we see that

$$\begin{aligned}
d\mathcal{L}_t \leq& -\lambda f(z_t)dt + \left(\frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2\right)dt + G_t dA_t \\
&+ \lambda\left(\int_0^t e^{-\lambda(t-s)}\left(\frac{L_N^2}{\epsilon}\|y_s - y_0\|_2^2 + L\|y_s - y_0\|_2\right)ds\right)dt - \left(\frac{L_N^2}{\epsilon}\|y_t - y_0\|_2^2 + L\|y_t - y_0\|_2\right)dt \\
&+ \lambda\left(\int_0^t e^{-\lambda(t-s)} G_s dA_s\right)dt - G_t dA_t \\
=& -\lambda \mathcal{L}_t dt
\end{aligned}$$

We can then apply Gronwall's Lemma to $\mathcal{L}_t$, so that

$$\mathcal{L}_T \leq e^{-\lambda T}\mathcal{L}_0,$$

which is equivalent to

$$f(z_T) - \int_0^T e^{-\lambda(T-s)}\left(\frac{L_N^2}{\epsilon}\|y_s - y_0\|_2^2 + L\|y_s - y_0\|_2\right)ds - \int_0^T e^{-\lambda(t-s)} G_s dA_s \leq e^{-\lambda T} f(z_0).$$

Observe that G_s is measurable wrt the natural filtration generated by A_s, so that $\int_0^T e^{-\lambda(T-s)} G_s dA_s$ is a martingale. Thus taking expectations,

$$\mathbb{E}\left[f(z_T)\right] \leq e^{-\lambda T}\mathbb{E}\left[f(z_0)\right] + \int_0^T \frac{L_N^2}{\epsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right] + L\mathbb{E}\left[\|y_s - y_0\|_2\right] ds$$

By Lemma 11, $\mathbb{E}\left[\|y_t - y_0\|_2^2\right] \leq t^2 L^2 \mathbb{E}\left[\|y_0\|_2^2\right] + t\beta^2$, so that

$$\begin{aligned}
\int_0^T \frac{L_N^2}{\epsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right] ds &\leq \frac{T^3 L_N^2 L^2}{\epsilon}\mathbb{E}\left[\|y_0\|_2^2\right] + \frac{T^2 L_N^2}{\epsilon}\beta^2 \\
L\mathbb{E}\left[\|y_s - y_0\|_2\right] &\leq T^2 L^2 \sqrt{\mathbb{E}\left[\|y_0\|_2^2\right]} + T^{3/2} L\beta
\end{aligned}$$

Furthermore, using our assumption in the Lemma statement that $T \leq \min\left\{\frac{\epsilon^2}{\beta^2}, \frac{\epsilon}{6L\sqrt{R^2+\beta^2/m}}\right\}$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8\left(R^2 + \beta^2/m\right)$, we can verify that

$$\begin{aligned}
\int_0^T \frac{L_N^2}{\epsilon}\mathbb{E}\left[\|y_s - y_0\|_2^2\right] ds &\leq \frac{1}{4}TL_N^2\epsilon + TL_N^2\epsilon \\
L\mathbb{E}\left[\|y_s - y_0\|_2\right] &\leq \frac{1}{2}TL\epsilon + TL\epsilon
\end{aligned}$$

Combining the above gives

$$\mathbb{E}\left[f(z_T)\right] \leq e^{-\lambda T}\mathbb{E}\left[f(z_0)\right] + 3T\left(L + L_N^2\right)\epsilon$$

■

Corollary 2 *Let f be as defined in Lemma 18 with parameter ϵ satisfying $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2 + 1}$.*

Let $\delta \leq \min\left\{\frac{\epsilon^2}{\beta^2}, \frac{\epsilon}{8L\sqrt{R^2+\beta^2/m}}\right\}$, and let $\bar{x}_t$ and $\bar{y}_t$ have dynamics as defined in (3) and (2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{y}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Then there exists a coupling between $\bar{x}_t$ and $\bar{y}_t$ such that

$$\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] \leq e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\epsilon$$

Proof of Corollary 2

From Lemma 7 and 8, our initial conditions imply that for all t, $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right] \leq 6\left(R^2 + \frac{\beta^2}{m}\right)$ and $\mathbb{E}\left[\|\bar{y}_{k\delta}\|_2^2\right] \leq 8\left(R^2 + \frac{\beta^2}{m}\right)$.

Consider an arbitrary k, and for $t \in [k\delta, (k+1)\delta)$, define

$$x_t := \bar{x}_{k\delta+t} \quad \text{and} \quad y_t := \bar{y}_{k\delta+t}$$

Under this definition, x_t and y_t have dynamics described in (20) and (21). Thus the coupling in (22), which describes a coupling between x_t and y_t, equivalently describes a coupling between $\bar{x}_t$ and $\bar{y}_t$ over $t \in [k\delta, (k+1)\delta)$.

We now apply Lemma 1. Given our assumed bound on δ and our proven bounds on $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right]$ and $\mathbb{E}\left[\|\bar{y}_t\|_2^2\right]$,

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{x}_{(k+1)\delta} - \bar{y}_{(k+1)\delta})\right] \\
=&\mathbb{E}\left[f(x_\delta - y_\delta)\right] \\
\leq& e^{-\lambda\delta}\mathbb{E}\left[f(x_0 - y_0)\right] + 6\delta(L + L_N^2)\epsilon \\
=& e^{-\lambda\delta}\mathbb{E}\left[f(\bar{x}_{k\delta} - \bar{y}_{k\delta})\right] + 6\delta(L + L_N^2)\epsilon
\end{aligned}$$

Applying the above recursively gives, for any i

$$\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] \leq e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\epsilon$$

■

A.4. Proof of Theorem 1

For ease of reference, we re-state Theorem 1 below as Theorem 3 below. We make a minor notational change: using the letters $\bar{x}_t$ and $\bar{y}_t$ in Theorem 3, instead of the letters x_t and y_t in Theorem 1. This is to avoid some notation conflicts in the proof.

Theorem 3 (Equivalent to Theorem 1) *Let $\bar{x}_t$ and $\bar{y}_t$ have dynamics as defined in (3) and (2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{y}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Let $\hat{\epsilon}$ be a target accuracy satisfying $\hat{\epsilon} \leq \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right) \cdot \exp\left(7\alpha_q\mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. Let δ be a step size satisfying*

$$\delta \leq \min\left\{\begin{array}{l}\frac{\lambda^2\hat{\epsilon}^2}{512\beta^2\left(L^2+L_N^4\right)\exp\left(\frac{14\alpha_q{\mathcal{R}_q}^2}{3}\right)} \\ \frac{2\lambda\hat{\epsilon}}{\left(L^2+L_N^4\right)\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\sqrt{R^2+\beta^2/m}}\end{array}\right. .$$

If we assume that $\bar{x}_0 = \bar{y}_0$, then there exists a coupling between $\bar{x}_t$ and $\bar{y}_t$ such that for any k,

$$\mathbb{E}\left[\|\bar{x}_{k\delta} - \bar{y}_{k\delta}\|_2\right] \leq \hat{\epsilon}$$

Alternatively, if we assume $k \geq \frac{3\alpha_q{\mathcal{R}_q}^2}{\delta}\log\frac{R^2+\beta^2/m}{\hat{\epsilon}}$, then

$$W_1(p^*, p_{k\delta}^y) \leq 2\hat{\epsilon}$$

where $p_t^y := \mathsf{Law}(\bar{y}_t)$.

Proof of Theorem 3

Let $\epsilon := \frac{\lambda}{16(L+L_N^2)}\exp\left(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\hat{\epsilon}$. Let f be defined as in Lemma 18 with the parameter ϵ.

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{y}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{y}_{i\delta})\right] + 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\epsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\epsilon\right) + 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\epsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \frac{16\left(L + L_N^2\right)}{\lambda}\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\cdot\epsilon \\
=& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{y}_0)\right] + \hat{\epsilon}
\end{aligned}
\tag{26}
$$

where the first inequality is by item 4 of Lemma 18, the second inequality is by Corollary 2 (notice that δ satisfies the requirement on T in Theorem 1, for the given ϵ). The third inequality uses the fact that $1 \leq L/m \leq \frac{(L+L_N^2)}{\lambda}$.

The first claim follows from substituting $\bar{x}_0 = \bar{y}_0$ into (26), so that the first term is 0, and using the definition of ϵ, so that the second term is 0.

For the second claim, let $\bar{x}_0 \sim p^*$, the invariant distribution of (3). From Lemma 7, we know that $\bar{x}_0$ satisfies the required initial conditions in this Lemma. Continuing from (26),

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{y}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\mathbb{E}\left[\|\bar{x}_0\|_2^2 + \|\bar{y}_0\|_2^2\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\epsilon\right) + \epsilon \\
\leq& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \frac{16}{\lambda}\exp\left(2\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(L + L_N^2\right)\epsilon \\
=& 4\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \hat{\epsilon}
\end{aligned}
$$

By our assumption that $i \geq \frac{1}{\delta}\cdot 3\alpha_q {\mathcal{R}_q}^2 \log\frac{R^2+\beta^2/m}{\hat{\epsilon}}$, the first term is also bounded by $\hat{\epsilon}$, and this proves our second claim. ■

A.5. Simulating the SDE

One can verify that the SDE in (2) can be simulated (at discrete time intervals) as follows:

$$
y_{(k+1)\delta} = y_{k\delta} - \delta\nabla U(y_{k\delta}) + \sqrt{\delta}M(y_{k\delta})\theta_k
$$

Where $\theta_k \sim \mathcal{N}(0, I)$. This however requires access to $M(y_{k,\delta})$, which may be difficult to compute.

If for any y, one is able to draw samples from some distribution p_y such that

1. $\mathbb{E}_{\xi\sim p_y}\left[\xi\right] = 0$
2. $\mathbb{E}_{\xi\sim p_y}\left[\xi\xi^T\right] = M(y)$
3. $\|\xi\|_2 \leq \beta$ almost surely, for some β.

then one might sample a noise that is δ close to $M(y_{k\delta})\theta_k$ through Theorem 5.

Specifically, if one draws n samples $\xi_1 ... \xi_n \overset{iid}{\sim} p_y$, and let $S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$, Theorem 5 guarantees that $W_2(S_n, M(y)\theta) \leq \frac{6\sqrt{d}\beta\sqrt{\log n}}{\sqrt{n}}$. We remark that the proof of Theorem 1 can be modified to accommodate for this sampling error. The number of samples needed to achieve ϵ accuracy will be on the order of $n \approxeq O(\delta\epsilon)^{-2} = O(\epsilon^{-6})$.

B. Proofs for Convergence under Non-Gaussian Noise (Theorem 2)

B.1. Proof Overview

The main proof of Theorem 2 is contained in Appendix B.4.

Here, we outline the steps of our proof:

1. In Appendix B.2, we construct a coupling between (3) and (1) over an epoch which consists of an interval $[k\delta, (k+n)\delta)$ for some k. The coupling in (B.2) consists of four processes (x_t, y_t, v_t, w_t), where y_t and v_t are auxiliary processes used in defining the coupling. Notably, the process (x_t, y_t) has the same distribution over the epoch as (22).
2. In Appendix B.3, we prove Lemma 3 and Lemma 4, which, combined with Lemma 1 from Appendix A.3, show that under the coupling constructed in Step 1, a Lyapunov function $f(x_T - w_T)$ contracts exponentially with rate λ, plus a discretization error term. In Corollary 5, we apply the results of Lemma 1, Lemma 3 and Lemma 4 recursively over multiple steps to give a bound on $f(x_{k\delta} - w_{k\delta})$ for all k, and for sufficiently small δ.
3. Finally, in Appendix B.4, we prove Theorem 2 by applying the results of Corollary 5, together with the fact that $f(z)$ upper bounds $\|z\|_2$ up to a constant.

B.2. Constructing a Coupling

In this subsection, we construct a coupling between (1) and (3), given arbitrary initialization (x_0, w_0). We will consider a finite time $T = n\delta$, which we will refer to as an *epoch*.

1. Let V_t and W_t be two independent Brownian motion.
2. Using V_t and W_t, define

$$x_t = x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t c_m dV_s + \int_0^t N(w_0) dW_s \tag{27}$$

3. Using the same V_t and W_t in (27), we will define y_t as

$$y_t = w_0 + \int_0^t -\nabla U(w_0) ds + \int_0^t c_m \left(I - 2\gamma_s \gamma_s^t\right) dV_s + \int_0^T N(x_s) dW_s \tag{28}$$

   Where $\gamma_t := \frac{x_t - y_t}{\|x_t - y_t\|_2} \cdot \mathbb{1}\left\{\|x_t - y_t\|_2 \in [2\epsilon, \mathcal{R}_q)\right\}$. The coupling (x_t, y_t) defined in (27) and (28) is identical to the coupling in (22) (with $y_0 = w_0$).

4. We now define a process $v_{k\delta}$ for $k = 0...n$:

$$v_{k\delta} = w_0 + \sum_{i=0}^{k-1} -\delta \nabla U(w_0) + \sqrt{\delta} \sum_{i=0}^{k-1} \xi(w_0, \eta_i) \tag{29}$$

   where marginally, the variables $(\eta_0 ... \eta_{n-1})$ are drawn $i.i.d$ from the same distribution as in (1).

   Notice that $y_T - w_0 - T\nabla U(w_0) = \int_0^T c_m dB_t + \int_0^T N(w_0) dW_t$, so that $\mathsf{Law}(y_T - w_0 - T\nabla U(w_0)) = \mathcal{N}(0, TM(w_0)^2)$. Notice also that $v_T - w_0 - T\nabla U(w_0) = \sqrt{\delta}\sum_{i=0}^{n-1} \xi(w_0, \eta_i)$. By Corollary 24, $W_2(y_T - w_0 - T\nabla U(w_0), v_T - w_0 - T\nabla U(w_0)) = 6\sqrt{d\delta}\beta\sqrt{\log n}$. Let the joint distribution between (29) and (28) be the one induced by the optimal coupling between $y_T - w_0 - T\nabla U(w_0)$ and $v_T - w_0 - T\nabla U(w_0)$, so that

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]} \\
=&\sqrt{\mathbb{E}\left[\|y_T - T\nabla U(w_0) - v_T + T\nabla U(w_0)\|_2^2\right]} \\
=&W_2(y_T - w_0 - T\nabla U(w_0), v_T - w_0 - T\nabla U(w_0)) \\
\le&6\sqrt{d\delta}\beta\sqrt{\log n}
\end{aligned} \tag{30}$$

   where the last inequality is by Corollary 24.

5. Given the sequence $(\eta_0 ... \eta_{n-1})$ from (29), we can define

$$w_{k\delta} = w_0 + \sum_{i=0}^{k-1} -\delta \nabla U(w_{i\delta}) + \sqrt{\delta} \sum_{i=0}^{k-1} \xi(w_{i\delta}, \eta_i) \tag{31}$$

specifically, $(w_0...w_{n\delta})$ in (31) and $(v_0...v_{n\delta})$ in (29) are coupled through the shared $(\eta_0...\eta_{n-1})$ variables.

For convenience, we will let $v_t := v_{i\delta}$ and $w_t := w_{i\delta}$, where i is the unique integer satisfying $t \in [i\delta, (i+1)\delta)$.

We can verify that, marginally, the process x_t in (27) has the same distribution as (3), using the proof as Lemma 6. It is also straightforward to verify that $w_{k\delta}$, as defined in (31), has the same marginal distribution as (1), due to the definition of η_i in (29).

B.3. One Epoch Contraction

In Lemma 3, we prove a discretization error bound between $f(x_T - y_T)$ and $f(x_T - v_T)$, for the coupling defined in (27), (28) and (29).

In Lemma 4, we prove a discretization error bound between $f(x_T - v_T)$ and $f(x_T - w_T)$, for the coupling defined in (27), (29) and (31).

Lemma 3 *Let f be as defined in Lemma 18 with parameter ϵ satisfying $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q \mathcal{R}_q{}^2+1}$. Let x_t, y_t and v_t be as defined in (27), (28), (29). Let n be any integer and δ be any step size, and let $T := n\delta$.*

If $\mathbb{E}\left[\|x_0\|_2^2\right] \leq 8\left(R^2 + \beta^2/m\right)$, $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8\left(R^2 + \beta^2/m\right)$ and $T \leq \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}\right\}$ and

$$\delta \leq \min\left\{\frac{T\epsilon^2 L}{36d\beta^2 \log\left(\frac{36d\beta^2}{\epsilon^2 L}\right)}, \frac{T\epsilon^4 L^2}{2^{14}d\beta^4 \log\left(\frac{2^{14}d\beta^4}{\epsilon^4 L^2}\right)}\right\}$$

Then

$$\mathbb{E}\left[f(x_T - v_T)\right] - \mathbb{E}\left[f(x_T - y_T)\right] \leq 4TL\epsilon$$

Proof
By Taylor's Theorem,

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - v_T)\right]\\
=&\mathbb{E}\left[f(x_T - y_T) + \langle \nabla f(x_T - y_T), y_T - v_T\rangle + \int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - y_T + s(y_T - v_T)), (y_T - v_T)(y_T - v_T)^T \right\rangle dsdt\right]\\
=&\mathbb{E}\left[f(x_T - y_T) + \underbrace{\langle \nabla f(x_0 - y_0), y_T - v_T\rangle}_{①} + \underbrace{\langle \nabla f(x_T - y_T) - \nabla f(x_0 - y_0), y_T - v_T\rangle}_{②}\right]\\
&+ \mathbb{E}\left[\underbrace{\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - y_T + s(y_T - v_T)), (y_T - v_T)(y_T - v_T)^T \right\rangle dsdt}_{③}\right]
\end{aligned}$$

We will bound each of the terms above separately.

$$\begin{aligned}
&\mathbb{E}\left[①\right]\\
=&\mathbb{E}\left[\langle \nabla f(x_0 - y_0), y_T - v_T\rangle\right]\\
=&\mathbb{E}\left[\left\langle \nabla f(x_0 - y_0), n\delta\nabla U(y_0) - n\delta\nabla U(v_0) + \int_0^T -\nabla U(w_0)dt + \int_0^T c_m dV_t + \int_0^T N(w_0)dW_t + \sum_{i=0}^{n-1} \sqrt{\delta}\xi(v_0, \eta_i)\right\rangle\right]\\
=&\mathbb{E}\left[\langle \nabla f(x_0 - y_0), n\delta\nabla U(y_0) - n\delta\nabla U(v_0)\rangle\right]\\
=&0
\end{aligned}$$

where the third equality is because $\int_0^T dB_t$, $\int_0^T dW_t$ and $\sum_{k=1}^T \xi(v_0, \eta_i)$ have zero mean conditioned on the information at time 0, and the fourth equality is because $y_0 = v_0$ by definition in (28) and (29).

$$
\begin{aligned}
&\mathbb{E}\left[\textcircled{2}\right]\\
=&\mathbb{E}\left[\langle \nabla f(x_T - y_T) - \nabla f(x_0 - y_0), y_T - v_T\rangle\right]\\
\leq&\sqrt{\mathbb{E}\left[\|\nabla f(x_T - y_T) - \nabla f(x_0 - y_0)\|_2^2\right]}\sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]}\\
\leq&\frac{2}{\epsilon}\sqrt{2\mathbb{E}\left[\|x_T - x_0\|_2^2 + \|y_T - y_0\|_2^2\right]}\sqrt{\mathbb{E}\left[\|y_T - v_T\|_2^2\right]}\\
\leq&\frac{2}{\epsilon}\sqrt{(32T\beta^2 + 4T\beta^2)}\cdot\left(6\sqrt{d\delta}\beta\log n\right)\\
\leq&\frac{128}{\epsilon}\sqrt{T}\beta^2\cdot\left(\sqrt{d\delta}\log n\right)
\end{aligned}
$$

Where the second inequality is by $\left\|\nabla^2 f\right\|_2 \leq \frac{2}{\epsilon}$ from item 2(c) of Lemma 18 and Young's inequality. The third inequality is by Lemma 10 and Lemma 11 and (30).

Finally, we can bound

$$
\begin{aligned}
&\mathbb{E}\left[\textcircled{3}\right]\\
\leq&\int_0^1\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(x_T - y_T + s(y_T - v_T))\right\|_2\|y_T - v_T\|_2^2\right] dsdt\\
\leq&\frac{2}{\epsilon}\mathbb{E}\left[\|y_T - v_T\|_2^2\right]\\
\leq&\frac{72 d\delta\beta^2\log^2 n}{\epsilon}
\end{aligned}
$$

Where the second inequality is by $\left\|\nabla^2 f\right\|_2 \leq \frac{2}{\epsilon}$ from item 2(c) of Lemma 18, the third inequality is by (30).

Summing these 3 terms,

$$
\begin{aligned}
&\mathbb{E}\left[f(x_T - v_T) - f(x_T - y_T)\right]\\
\leq&\frac{128}{\epsilon}\sqrt{T}\beta^2\cdot\left(\sqrt{d\delta}\sqrt{\log n}\right) + \frac{36 d\delta\beta^2\log n}{\epsilon}\\
=&\frac{128}{\epsilon}\sqrt{T}\beta^2\cdot\left(\sqrt{d\delta}\sqrt{\log\frac{T}{\delta}}\right) + \frac{36 d\delta\beta^2\log\frac{T}{\delta}}{\epsilon}
\end{aligned}
$$

Let us bound the first term. We apply Lemma 25 (with $x = \frac{T}{\delta}$ and $c = \frac{\epsilon^4}{2^{14}d\beta^4}$), which shows that

$$
\frac{T}{\delta} \geq \frac{2^{14}d\beta^4}{\epsilon^4}\log\left(\frac{2^{14}d\beta^4}{\epsilon^4 L^2}\right) \quad\Rightarrow\quad \frac{T}{\delta}\frac{1}{\log\frac{T}{\delta}} \geq \frac{2^{14}d\beta^4}{\epsilon^4 L^2} \quad\Leftrightarrow\quad \frac{128}{\epsilon}\sqrt{T}\beta^2\cdot\left(\sqrt{d\delta}\log\frac{T}{\delta}\right) \leq TL\epsilon
$$

For the second term, we can again apply Lemma 25 ($x = \frac{T}{\delta}$ and $c = \frac{\epsilon^2 L}{36 d\beta^2}$), which shows that

$$
\frac{T}{\delta} \geq \frac{36 d\beta^2}{\epsilon^2 L}\log\left(\frac{36 d\beta^2}{\epsilon^2 L}\right) \quad\Rightarrow\quad \frac{T}{\delta}\frac{1}{\log\frac{T}{\delta}} \geq \frac{36 d\beta^2}{\epsilon^2 L} \quad\Rightarrow\quad \frac{36 d\delta\beta^2\log\frac{T}{\delta}}{\epsilon} \leq TL\epsilon
$$

The above imply that

$$
\mathbb{E}\left[f(x_T - v_T) - f(x_T - y_T)\right] \leq 2TL\epsilon
$$

■

Lemma 4 *Let f be as defined in Lemma 18 with parameter ϵ satisfying $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q {\mathcal{R}_q}^2+1}$. Let x_t, v_t and w_t be as defined in (27), (29), (31). Let n be an integer and δ be a step size, and let $T := n\delta$.*

If we assume that $\mathbb{E}\left[\|x_0\|_2^2\right]$, $\mathbb{E}\left[\|v_0\|_2^2\right]$, and $\mathbb{E}\left[\|w_0\|_2^2\right]$ are each upper bounded by $8\left(R^2 + \beta^2/m\right)$ and that $T \leq \min\left\{\frac{1}{16L}, \frac{\epsilon}{32\sqrt{L}\beta}, \frac{\epsilon^2}{128\beta^2}, \frac{\epsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$, then

$$\mathbb{E}\left[f(x_T - w_T)\right] - \mathbb{E}\left[f(x_T - v_T)\right] \leq 4T(L + L_N^2)\epsilon$$

Remark 9 *For sufficiently small ϵ, our assumption on T boils down to $T = o(\epsilon^4)$*

Proof
First, we can verify using Taylor's theorem that for any x, y,

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \int_0^s \left\langle \nabla^2 f(x + s(y-x)), (y-x)(y-x)^T \right\rangle dsdt \tag{32}$$

$$\nabla f(y) = \nabla f(x) + \left\langle \nabla^2 f(x), y - x\right\rangle + \int_0^1 \int_0^s \left\langle \nabla^3 f(x + s(y-x)), (y-x)(y-x)^T \right\rangle dsdt \tag{33}$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[f(x_T - w_T)\right] \\
=&\mathbb{E}\left[f(x_T - v_T) + \langle \nabla f(x_T - v_T), v_T - w_T\rangle + \int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T \right\rangle dsdt\right] \\
=&\mathbb{E}\left[f(x_T - v_T) + \underbrace{\langle \nabla f(x_0 - v_0), v_T - w_T\rangle}_{①} + \underbrace{\langle \nabla f(x_T - v_T) - \nabla f(x_0 - v_0), v_T - w_T\rangle}_{②}\right] \\
&+ \mathbb{E}\left[\underbrace{\int_0^1 \int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T \right\rangle dsdt}_{③}\right]
\end{aligned}$$

Recall from (29) and (31) that

$$\begin{aligned}
v_{n\delta} =& w_0 + \sum_{i=0}^{n-1} \delta \nabla U(w_0) + \sqrt{\delta} \sum_{i=0}^{n-1} \xi(w_0, \eta_i) \\
w_{n\delta} =& w_0 + \sum_{i=0}^{n-1} \delta \nabla U(w_{i\delta}) + \sqrt{\delta} \sum_{i=0}^{n-1} \xi(w_{i\delta}, \eta_i)
\end{aligned}$$

Note that conditioned on the randomness up to time 0, $\mathbb{E}\left[\sum_{i=0}^{n-1}\xi(w_0,\eta_i)\right] = \mathbb{E}\left[\sum_{i=0}^{n-1}\xi(w_{i\delta},\eta_i)\right] = 0$, so that

$$
\begin{aligned}
&\mathbb{E}\left[①\right]\\
=&\mathbb{E}\left[\langle \nabla f(x_0 - v_0), v_T - w_T\rangle\right]\\
=&\delta\mathbb{E}\left[\left\langle \nabla f(x_0 - v_0), \sum_{i=0}^{n-1} \nabla U(w_0) - \nabla U(w_{i\delta})\right\rangle\right] + \sqrt{\delta}\mathbb{E}\left[\left\langle \nabla f(x_0 - v_0), \sum_{i=0}^{n-1}\xi(w_0,\eta_i) - \sum_{i=0}^{n-1}\xi(w_{i\delta},\eta_i)\right\rangle\right]\\
=&\delta\mathbb{E}\left[\left\langle \nabla f(x_0 - v_0), \sum_{i=0}^{n-1} \nabla U(w_0) - \nabla U(w_{i\delta})\right\rangle\right]\\
\leq&\delta\sum_{i=0}^{n-1} L\mathbb{E}\left[\|w_0 - w_{i\delta}\|_2\right]\\
\leq& TL\sqrt{32T\beta^2} \leq 8T^{3/2}L\beta
\end{aligned}
$$

where the third equality is becayse $\xi(\cdot,\eta_i)$ has 0 mean conditioned on the randomness at time 0, and the second inequality is by Lemma 13.

Next,

$$
\begin{aligned}
&\mathbb{E}\left[②\right]\\
=&\mathbb{E}\left[\langle \nabla f(x_T - v_T) - \nabla f(x_0 - v_0), v_T - w_T\rangle\right]\\
\leq&\mathbb{E}\left[\|\nabla f(x_T - v_T) - \nabla f(x_0 - v_0)\|_2\|v_T - w_T\|\right]\\
\leq&\frac{4}{\epsilon}\sqrt{\mathbb{E}\left[\|x_T - x_0\|_2^2 + \|v_T - v_0\|_2^2\right]}\cdot\sqrt{\mathbb{E}\left[\|v_T - w_T\|_2^2\right]}\\
\leq&\frac{4}{\epsilon}\sqrt{16T\beta^2 + 2T\beta^2}\cdot\sqrt{32\left(T^2L^2 + TL_\xi^2\right)T\beta^2}\\
\leq&\frac{128}{\epsilon}T\beta^2\left(\sqrt{T}L_\xi + TL\right)
\end{aligned}
$$

where the second inequality is because $\left\|\nabla^2 f\right\|_2 \leq \frac{2}{\epsilon}$ from item 2(c) of Lemma 18 and by Young's inequality. The third inequality is by Lemma 10, Lemma 12 and Lemma 14.

Finally,

$$
\begin{aligned}
&\mathbb{E}\left[③\right]\\
=&\mathbb{E}\left[\int_0^1\int_0^s \left\langle \nabla^2 f(x_T - v_T + s(v_T - w_T)), (v_T - w_T)(v_T - w_T)^T\right\rangle dsdt\right]\\
\leq&\int_0^1\int_0^s \mathbb{E}\left[\left\|\nabla^2 f(x_T - v_T + s(v_T - w_T))\right\|_2\|v_T - w_T\|_2^2\right]ds\\
\leq&\frac{1}{\epsilon}\mathbb{E}\left[\|v_T - w_T\|_2^2\right]\\
\leq&\frac{32}{\epsilon}(T^2L^2 + TL_\xi^2)T\beta^2
\end{aligned}
$$

wehere the second inequality is because $\left\|\nabla^2 f\right\|_2 \leq \frac{2}{\epsilon}$ from item 2(c) of Lemma 18 and by Young's inequality. The third inequality is by Lemma 14.

Summing the above,

$$
\begin{aligned}
&\mathbb{E}\left[f(x_T - w_T) - f(x_T - v_T)\right]\\
\leq&8T^{3/2}L\beta + \frac{128}{\epsilon}T\beta^2\left(\sqrt{T}L_\xi + TL\right) + \frac{32}{\epsilon}(T^2L^2 + TL_\xi^2)T\beta^2\\
\leq&T^{3/2}\epsilon
\end{aligned}
$$

where the last inequality is by our assumption on T, specifically,

$$
\begin{aligned}
&T \le \frac{\epsilon^2}{128\beta^2} \Rightarrow T^{3/2}L\beta \le TL\epsilon\\
&T \le \frac{\epsilon^2}{128\beta^2} \Rightarrow \frac{128}{\epsilon}T^2 L\beta^2 \le TL\epsilon\\
&T \le \frac{\epsilon}{32\sqrt{L}\beta} \Rightarrow \frac{32}{\epsilon}(T^3L^2\beta^2) \le TL\epsilon\\
&T \le \frac{\epsilon^4 L_N^2}{2^{14}\beta^2 c_m^2} \Rightarrow \frac{128}{\epsilon}T^{3/2}\beta^2 L_\xi \le TL_N^2\epsilon\\
&T \le \frac{\epsilon^2}{128\beta^2} \Rightarrow T \le \frac{\epsilon^2}{128 c_m^2} \Rightarrow \frac{32}{\epsilon}T^2 L_\xi^2\beta^2 \le TL_N^2\epsilon
\end{aligned}
$$

where the last line uses the fact that $\beta \ge c_m^2$.

■

Corollary 5 *Let f be as defined in Lemma 18 with parameter ϵ satisfying $\epsilon \le \frac{\mathcal{R}_q}{\alpha_q {\mathcal{R}_q}^2+1}$.*
Let $T = \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\epsilon}{32\sqrt{L}\beta}, \frac{\epsilon^2}{128\beta^2}, \frac{\epsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$ and let $\delta \le \min\left\{\frac{T\epsilon^2 L}{36d\beta^2\log\left(\frac{36d\beta^2}{\epsilon^2 L}\right)}, \frac{T\epsilon^4 L^2}{2^{14}d\beta^4\log\left(\frac{2^{14}d\beta^4}{\epsilon^4 L^2}\right)}\right\}$, assume additionally that $n = T/\delta$ is an integer.
Let $\bar{x}_t$ and $\bar{w}_t$ have dynamics as defined in (3) and (2) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \le R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{w}_0\|_2^2\right] \le R^2 + \beta^2/m$. Then there exists a coupling between $\bar{x}_t$ and $\bar{w}_t$ such that

$$
\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{w}_{i\delta})\right] \le e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{6}{\lambda}(L + L_N^2)\epsilon
$$

Proof
From Lemma 7 and 9, our initial conditions imply that for all t, $\mathbb{E}\left[\|\bar{x}_t\|_2^2\right] \le 6\left(R^2 + \frac{\beta^2}{m}\right)$ and $\mathbb{E}\left[\|\bar{w}_{k\delta}\|_2^2\right] \le 8\left(R^2 + \frac{\beta^2}{m}\right)$.

Consider an arbitrary k, and for $t \in [0, T)$, define

$$
x_t := \bar{x}_{kT+t} \quad \text{and} \quad w_t := \bar{w}_{kT+t} \tag{34}
$$

Notice that as described above, x_t and w_t have dynamics described in (3) and (1). Let x_t, w_t have joint distribution as described in (27) and (31), and let (y_t, v_t) be the processes defined in (28) and (29). Notice that the joint distribution between x_t and w_t equivalently describes a coupling between $\bar{x}_t$ and $\bar{w}_t$ over $t \in [kT, (k+1)T)$.

First, notice that the processes (27) and (28) have the same distribution as (22). We can thus apply Lemma 1:

$$
\mathbb{E}\left[f(x_T - y_T)\right] \le e^{-\lambda T}\mathbb{E}\left[f(x_0 - y_0)\right] + 6T(L + L_N^2)\epsilon
$$

By Lemma 3,

$$
\mathbb{E}\left[f(x_T - v_T)\right] - \mathbb{E}\left[f(x_T - y_T)\right] \le 4TL\epsilon
$$

By Lemma 4,

$$
\mathbb{E}\left[f(x_T - w_T)\right] - \mathbb{E}\left[f(x_T - v_T)\right] \le 4T(L + L_N^2)\epsilon
$$

Summing the above three equations,

$$
\mathbb{E}\left[f(x_T - w_T)\right] \le e^{-\lambda\delta}\mathbb{E}\left[f(x_0 - w_0)\right] + 14T(L + L_N^2)
$$

Where we use the fact that $y_0 = w_0$ by construction in (28).

Recalling (34), this is equivalent to

$$\mathbb{E}\left[f(\bar{x}_{(k+1)T} - \bar{w}_{(k+1)T})\right] \leq e^{-\lambda\delta}\mathbb{E}\left[f(\bar{x}_{kT} - \bar{w}_{kT})\right] + 14T(L + L_N^2)$$

Applying the above recursively gives, for any i

$$\mathbb{E}\left[f(\bar{x}_{iT} - \bar{w}_{iT})\right] \leq e^{-\lambda iT}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{14}{\lambda}\left(L + L_N^2\right)\epsilon$$

■

B.4. Proof of Theorem 2

For ease of reference, we re-state Theorem 2 below as Theorem 4 below. We make a minor notational change: using the letters $\bar{x}_t$ and $\bar{y}_t$ in Theorem 4, instead of the letters x_t and y_t in Theorem 2. This is to avoid some notation conflicts in the proof.

Theorem 4 (Equivalent to Theorem 2) *Let $\bar{x}_t$ and w_t have dynamics as defined in (3) and (1) respectively, and suppose that the initial conditions satisfy $\mathbb{E}\left[\|\bar{x}_0\|_2^2\right] \leq R^2 + \beta^2/m$ and $\mathbb{E}\left[\|\bar{w}_0\|_2^2\right] \leq R^2 + \beta^2/m$. Let $\hat{\epsilon}$ be a target accuracy satisfying $\hat{\epsilon} \leq \left(\frac{16\left(L+L_N^2\right)}{\lambda}\right) \cdot \exp\left(7\alpha_q\mathcal{R}_q/3\right) \cdot \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. Let $\epsilon := \frac{\lambda}{16(L+L_N^2)}\exp\left(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\hat{\epsilon}$. Let $T := \min\left\{\frac{1}{16L}, \frac{\beta^2}{8L^2(R^2+\beta^2/m)}, \frac{\epsilon}{32\sqrt{L}\beta}, \frac{\epsilon^2}{128\beta^2}, \frac{\epsilon^4 L_N^2}{2^{14}\beta^2 c_m^2}\right\}$ and let δ be a step size satisfying*

$$\delta \leq \min\left\{\frac{T\epsilon^2 L}{36d\beta^2\log\left(\frac{36d\beta^2}{\epsilon^2 L}\right)}, \frac{T\epsilon^4 L^2}{2^{14}d\beta^4\log\left(\frac{2^{14}d\beta^4}{\epsilon^4 L^2}\right)}\right\}.$$

If we assume that $\bar{x}_0 = \bar{w}_0$, then there exists a coupling between $\bar{x}_t$ and $\bar{w}_t$ such that for any k,

$$\mathbb{E}\left[\|\bar{x}_{k\delta} - \bar{w}_{k\delta}\|_2\right] \leq \hat{\epsilon}.$$

Alternatively, if we assume that $k \geq \frac{3\alpha_q{\mathcal{R}_q}^2}{\delta} \cdot \log\frac{R^2+\beta^2/m}{\hat{\epsilon}}$, then

$$W_1(p^*, p_{k\delta}^w) \leq 2\hat{\epsilon},$$

where $p_t^w := \mathsf{Law}(\bar{w}_t)$.

Proof of Theorem 4

Let f be defined as in Lemma 18 with parameter ϵ.

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{w}_{i\delta}\|_2\right] \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\mathbb{E}\left[f(\bar{x}_{i\delta} - \bar{w}_{i\delta})\right] + 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\epsilon \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{6}{\lambda}\left(L + L_N^2\right)\epsilon\right) + 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)\epsilon \\
\leq& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \frac{16\left(L + L_N^2\right)}{\lambda}\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right) \cdot \epsilon \qquad (35) \\
=& 2\exp\left(\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\right)e^{-\lambda i\delta}\mathbb{E}\left[f(\bar{x}_0 - \bar{w}_0)\right] + \hat{\epsilon}
\end{aligned}$$

where the first inequality is by item 4 of Lemma 18, the second inequality is by Corollary 5 (notice that δ satisfies the requirement on T in Theorem 1, for the given ϵ). The third inequality uses the fact that $1 \leq L/m \leq \frac{(L+L_N^2)}{\lambda}$.

The first claim follows from substituting $\bar{x}_0 = \bar{w}_0$ into (35), so that the first term is 0, and using the definition of ϵ, so that the second term is 0.

For the second claim, let $\bar{x}_0 \sim p^*$, the invariant distribution of (3). From Lemma 7, we know that $\bar{x}_0$ satisfies the required initial conditions in this Lemma. Continuing from (35),

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{i\delta} - \bar{w}_{i\delta}\|_2\right] \\
\le& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\mathbb{E}\left[\|\bar{x}_0\|_2^2 + \|\bar{w}_0\|_2^2\right] + \frac{6}{\lambda}(L + L_N^2)\epsilon\right) + \epsilon \\
\le& 2\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(2e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \frac{16}{\lambda}\exp\left(2\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)(L + L_N^2)\epsilon \\
=& 4\exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)\left(e^{-\lambda i\delta}\left(R^2 + \beta^2/m\right)\right) + \hat{\epsilon}
\end{aligned}$$

By our assumption that $i \ge \frac{1}{\delta} \cdot 3\alpha_q {\mathcal{R}_q}^2 \log \frac{R^2+\beta^2/m}{\hat{\epsilon}}$, the first term is also bounded by $\hat{\epsilon}$, and this proves our second claim. ■

C. Coupling Properties

Lemma 6 *Consider the coupled (x_t, y_t) in (22). Let p_t denote the distribution of x_t, and q_t denote the distribution of y_t. Let p'_t and q'_t denote the distributions of (20) and (21).*

If $p_0 = p'_0$ and $q_0 = q'_0$, then $p_t = p'_t$ and $q_t = q'_t$ for all t.

Proof
Consider the coupling in (22), reproduced below for ease of reference:

$$\begin{aligned}
x_t =& x_0 + \int_0^t -\nabla U(x_s)ds + \int_0^t c_m dV_s + \int_0^t N(x_s)dW_s \\
y_t =& y_0 + \int_0^t -\nabla U(y_0)dt + \int_0^t c_m\left(I - 2\gamma_s\gamma_s^T\right)dV_s + \int_0^t N(y_0)dW_s
\end{aligned}$$

Let us define the stochastic process $A_t := \int_0^t M(x_s)^{-1}c_m dV_s + \int_0^t M(x_s)^{-1}N(x_s)dW_s$. We can verify using Levy's characterization that A_t is a standard Brownian motion: first, since V_t and W_t are Brownian motions, and $N(x)$ is differentiable with bounded derivatives, we know that A_t has continuous sample paths. We now verify that $A_t^i A_t^j - \mathbb{1}\{i=j\}t$ is a martingale.

Notice that $dA_t = c_m dV_t + M(x_s)^{-1}N(x_s)dW_s$. Then

$$\begin{aligned}
dA_t^i A_t^j =& dA_t^T\left(e_i e_j^T\right)A_t \\
=& A_t\left(e_i e_j^T\right)\left(c_m dV_t + M(x_s)^{-1}N(x_s)dW_s\right)^T + \left(c_m dV_t + M(x_s)^{-1}N(x_s)dW_s\right)\left(e_j e_i^T\right)a_t^T \\
&+ \frac{1}{2}\text{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\left(c_m^2 M(x_s)^{-2} + M(x_s)^{-1}N(x_s)^2 M(x_s)^{-1}\right)\right)dt
\end{aligned}$$

where the second inequality is by Ito's Lemma applied to $f(A_t) = A_t^T e_j e_j^T A_t$. Taking expectations,

$$\begin{aligned}
d\mathbb{E}\left[A_t^i A_t^j\right] =& \mathbb{E}\left[\frac{1}{2}\text{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\left(c_m^2 M(x_s)^{-2} + M(x_s)^{-1}N(x_s)N(x_s)^T\left(M(x_s)^{-1}\right)^T\right)\right)\right]dt \\
=& \mathbb{E}\left[\frac{1}{2}\text{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\left(M(x_s)^{-1}\left(c_m^2 I + N(x_s)^2\right)M(x_s)^{-1}\right)\right)\right]dt \\
=& \mathbb{E}\left[\frac{1}{2}\text{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\left(M(x_s)^{-1}\left(M(x_s)^2\right)M(x_s)^{-1}\right)\right)\right]dt \\
=& \mathbb{E}\left[\frac{1}{2}\text{tr}\left(\left(e_i e_j^T + e_j e_i^T\right)\right)\right]dt \\
=& \mathbb{1}\{i=j\}dt
\end{aligned}$$

This verifies that $A_t^i A_t^j - \mathbb{1}\{i = j\} t$ is a martingale, and hence by Levy's characterization, A_t is a standard Brownian motion. In turn, we verify that by definition of A_t,

$$\begin{aligned} x_t =& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t c_m dV_s + \int_0^t N(x_s) dW_s \\ =& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t M(x_s)\big(M(x_s)^{-1}(c_m dV_s + N(x_s) dW_s)\big) \\ =& x_0 + \int_0^t -\nabla U(x_s) ds + \int_0^t M(x_s) dA_s \end{aligned}$$

Since we showed that A_t is a standard Brownian motion, we verify that x_t as defined in (22) has the same distribution as (3).

On the other hand, we can verify that $A'_t := \int_0^T (I - 2\gamma_s \gamma_s^T) V_s$ is a standard Brownian motion by the reflection principle. Thus

$$\int_0^t c_m \big(I - 2\gamma_s \gamma_s^T\big) dV_s + \int_0^t N(y_0) dW_s \sim \mathcal{N}(0, \big(c_m^2 I + N(y_0)^2\big)) = \mathcal{N}(0, M(y_0)^2)$$

where the equality is by definition of N in (6).

It follows immediately that y_t in (22) has the same distribution as y_t in (2).

■

C.1. Energy Bounds

Lemma 7 *Consider x_t as defined in (3). If x_0 satisfies $\mathbb{E}\left[\|x_0\|_2^2\right] \le R^2 + \frac{\beta^2}{m}$, then Then for all t,*

$$\mathbb{E}\left[\|x_t\|_2^2\right] \le 6\left(R^2 + \frac{\beta^2}{m}\right)$$

We can also show that

$$\mathbb{E}_{p^*}\left[\|x\|_2^2\right] \le 4\left(R^2 + \frac{\beta^2}{m}\right)$$

Proof
We consider the potential function $a(x) = (\|x\|_2 - R)_+^2$ We verify that

$$\begin{aligned} \nabla a(x) =& (\|x\|_2 - R)_+ \frac{x}{\|x\|_2} \\ \nabla^2 a(x) =& \mathbb{1}\{\|x\|_2 \ge R\} \frac{xx^T}{\|x\|_2^2} + \frac{(\|x\|_2 - R)_+}{\|x\|_2}\left(I - \frac{xx^T}{\|x\|_2^2}\right) \end{aligned}$$

Observe that

1. $\left\|\nabla^2 a(x)\right\|_2 \le 2\mathbb{1}\{\|x\|_2 \ge R\} \le 2$
2. $\langle \nabla a(x), -\nabla U(x)\rangle \le -m a(x)$. This can be verified by considering 2 cases. If $\|x\|_2 \le R$, then $\nabla a(x) = 0$ and $a(x) = 0$. If $\|x\|_2 \ge R$, then by Assumption A,

$$\langle \nabla a(x), -\nabla U(x)\rangle \le -m(\|x\|_2 - R)_+ \|w\|_2 \le -m(\|x\|_2 - R)_+^2 = -m \cdot a(x)$$

3. $a(x) \ge \frac{1}{2}\|x\|_2^2 - 2R^2$. One can first verify that $a(x) \ge (\|x\|_2 - R)^2 - R^2$. Next, by Young's inequality, $(\|x\|_2 - R)^2 = \|x\|_2^2 + R^2 - 2\|x\|_2 R \ge \|x\|_2^2 + R^2 - \frac{1}{2}\|x\|_2^2 - 2R^2 = \frac{1}{2}\|x\|_2^2 - R^2$.

Therefore,

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[a(x_t)\right] = \mathbb{E}\left[\langle \nabla a(x_t), -\nabla U(x_t)dt\rangle\right] + \frac{1}{2}\mathbb{E}\left[\mathrm{tr}\big(M(x_t)^2\nabla^2 a(x)\big)\right] \leq -m\mathbb{E}\left[a(x_t)\right] + \beta^2\\
\Rightarrow \quad &\frac{d}{dt}\left(\mathbb{E}\left[a(x_t)\right] - \frac{\beta^2}{m}\right) \leq -m\left(\mathbb{E}\left[a(x_t)\right] - \frac{\beta^2}{m}\right)\\
\Rightarrow \quad &\frac{d}{dt}\left(\mathbb{E}\left[a(x_t)\right] - R^2 - \frac{\beta^2}{m}\right) \leq -m\left(\mathbb{E}\left[a(x_t)\right] - R^2 - \frac{\beta^2}{m}\right)
\end{aligned}$$

Thus if $\mathbb{E}\left[\|x_0\|_2^2\right] \leq R^2 + \frac{\beta^2}{m}$, then $\mathbb{E}\left[a(x_0)\right] \leq R^2 - \frac{\beta^2}{m}$, then $\left(\mathbb{E}\left[a(x_0)\right] - R^2 - \frac{\beta^2}{m}\right) \leq 0$, and $\left(\mathbb{E}\left[a(x_t)\right] - R^2 + \frac{\beta^2}{m}\right) \leq e^{-mt} \cdot 0 \leq 0$ for all t. This implies that, for all t,

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq \mathbb{E}\left[2a(x_t) + 4R^2\right] \leq 6\left(R^2 + \frac{\beta^2}{m}\right)$$

For our second claim that $\mathbb{E}_{p^*}\left[\|x\|_2^2\right] \leq R^2 + \frac{\beta^2}{m}$, we can use the fact that if $x_0 \sim p^*$, then $\mathbb{E}\left[a(x_t)\right]$ does not change as p^* is invariant, so that

$$0 = \frac{d}{dt}\mathbb{E}\left[a(x_t)\right] \leq -m\mathbb{E}\left[a(x_t)\right] + \beta^2$$

Thus

$$\mathbb{E}\left[a(x_t)\right] \leq \frac{\beta^2}{m}$$

Again,

$$\mathbb{E}_{p^*}\left[\|x\|_2^2\right] = \mathbb{E}\left[\|x_t\|_2^2\right] \leq 2\mathbb{E}\left[a(x_t)\right] + 4R^2 \leq 4\left(R^2 + \frac{\beta^2}{m}\right)$$

■

Lemma 8 *Let the sequence $y_{k\delta}$ be as defined in (1). Assuming that $\delta \leq m/(16L^2)$ and $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 2\left(R^2 + \frac{\beta^2}{m}\right)$ Then for all k,*

$$\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] \leq 8\left(R^2 + \frac{\beta^2}{m}\right)$$

Proof
Let $a(w) := (\|w\|_2 - R)_+^2$. We can verify that

$$\begin{aligned}
\nabla a(w) =& (\|w\|_2 - R)_+ \frac{w}{\|w\|_2}\\
\nabla^2 a(w) =& \mathbb{1}\left\{\|w\|_2 \geq R\right\}\frac{ww^T}{\|w\|_2^2} + (\|w\|_2 - R)_+ \frac{1}{\|w\|_2}\left(I - \frac{ww^T}{\|w\|_2^2}\right)
\end{aligned}$$

Observe that

1. $\left\|\nabla^2 a(w)\right\|_2 \leq 2\mathbb{1}\left\{\|w\|_2 \geq R\right\} \leq 2$
2. $\langle \nabla a(w), -\nabla U(w)\rangle \leq -ma(w)$.
3. $a(w) \geq \frac{1}{2}\|w\|_2^2 - 2R^2$.

The proofs are identical to the proof at the start of Lemma 9, so we omit them here.

Using Taylor's Theorem, and taking expectation of $y_{(k+1)\delta}$ conditioned on $y_{k\delta}$,

$$\begin{aligned}
&\mathbb{E}\left[a(y_{(k+1)\delta})\right]\\
=&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(y_{k\delta}), y_{(k+1)\delta} - y_{k\delta}\right\rangle\right]\\
&\quad + \mathbb{E}\left[\int_0^1\int_0^t \left\langle \nabla^2 a(y_{k\delta} + s(y_{(k+1)\delta} - y_{k\delta}), (y_{(k+1)\delta} - y_{k\delta})(y_{(k+1)\delta} - y_{k\delta})^T\right\rangle dtds\right]\\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(y_{k\delta}), y_{(k+1)\delta} - y_{k\delta}\right\rangle\right] + \mathbb{E}\left[\left\|(y_{(k+1)\delta} - y_{k\delta})\right\|_2^2 ds\right]\\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] + \mathbb{E}\left[\langle \nabla a(y_{k\delta}), -\delta\nabla U(y_{k\delta})\rangle\right] + 2\delta^2\|\nabla U(y_{k\delta})\|_2^2 + 2\delta\mathbb{E}\left[\mathrm{tr}\big(M(y_{k\delta})^2\big)\right]\\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 2\delta^2\mathbb{E}\left[\|\nabla U(y_{k\delta})\|_2^2\right] + 2\delta\mathbb{E}\left[\mathrm{tr}\big(M(y_{k\delta})^2\big)\right]\\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 2\delta^2L^2\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] + 2\delta\beta^2\\
\leq&\mathbb{E}\left[a(y_{k\delta})\right] - m\delta\mathbb{E}\left[a(y_{k\delta})\right] + 4\delta^2L^2\mathbb{E}\left[a(y_{k\delta})\right] + 8\delta^2L^2R^2 + 2\delta\beta^2\\
\leq&(1 - m\delta/2)\mathbb{E}\left[a(y_{k\delta})\right] + m\delta R^2 + 2\delta\beta^2
\end{aligned}$$

Where the first inequality uses the upper bound on $\left\|\nabla^2 a(y)\right\|_2$ above, the second inequality uses the fact that $y_{(k+1)\delta} \sim \mathcal{N}\big(y_{k\delta} - \delta\nabla U(y_{k\delta}), \delta M(y_{k\delta})^2\big)$, the third inequality uses claim 2. at the start of this proof, the fourth inequality uses item 2 of Assumption B. The fifth inequality uses claim 3. above, the sixth inequality uses our assumption that $\delta \leq \frac{m}{16L^2}$.

Taking expectation wrt $y_{k\delta}$,

$$\begin{aligned}
&\mathbb{E}\left[a(y_{(k+1)\delta})\right] \leq \mathbb{E}\left[a(y_k)\right] - m\delta\big(\mathbb{E}\left[a(y_{k\delta})\right] - 2R^2 + 2\beta^2/m\big)\\
\Rightarrow\qquad &\mathbb{E}\left[a(y_{(k+1)\delta})\right] - (2R^2/2 + 2\beta^2/m) \leq (1 - m\delta)\big(\mathbb{E}\left[a(y_{k\delta})\right] - (2R^2 + 2\beta^2/m)\big)
\end{aligned}$$

Thus, if $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 2R^2 + 2\beta^2/m$, then $\mathbb{E}\left[a(y_0)\right] - \big(2R^2 + 2\beta^2/m\big) \leq 0$, then $\mathbb{E}\left[a(y_{k\delta})\right] - \big(2R^2 + 2\beta^2/m\big) \leq 0$ for all k, which implies that

$$\mathbb{E}\left[\|y_{k\delta}\|_2^2\right] \leq 2\mathbb{E}\left[a(y_{k\delta})\right] + 4R^2 \leq 8\big(R^2 + \beta^2/m\big)$$

for all k. ■

Lemma 9 *Let the sequence $w_{k\delta}$ be as defined in (1). Assuming that $\delta \leq m/(16L^2)$ and $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 2\left(R^2 + \frac{\beta^2}{m}\right)$ Then for all k,*

$$\mathbb{E}\left[\|w_{k\delta}\|_2^2\right] \leq 8\left(R^2 + \frac{\beta^2}{m}\right)$$

Proof
The proof is almost identical to that of Lemma 8. Let $a(w) := (\|w\|_2 - R)_+^2$. We can verify that

$$\begin{aligned}
\nabla a(w) =&(\|w\|_2 - R)_+ \frac{w}{\|w\|_2}\\
\nabla^2 a(y) =&\mathbb{1}\left\{\|w\|_2 \geq R\right\}\frac{ww^T}{\|w\|_2^2} + (\|w\|_2 - R)_+ \frac{1}{\|w\|_2}\left(I - \frac{ww^T}{\|w\|_2^2}\right)
\end{aligned}$$

Observe that

1. $\left\|\nabla^2 a(w)\right\|_2 \leq 2\mathbb{1}\left\{\|w\|_2 \geq R\right\} \leq 2$
2. $\langle \nabla a(w), -\nabla U(w)\rangle \leq -ma(w)$.
3. $a(w) \geq \frac{1}{2}\|w\|_2^2 - 2R^2$.

The proofs are identical to the proof at the start of Lemma 9, so we omit them here.

Using Taylor's Theorem, and taking expectation of $w_{(k+1)\delta}$ conditioned on $w_{k\delta}$,

$$
\begin{aligned}
&\mathbb{E}\left[a(w_{(k+1)\delta})\right]\\
=&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(w_{k\delta}), w_{(k+1)\delta} - w_{k\delta}\right\rangle\right]\\
&+ \mathbb{E}\left[\int_0^1 \int_0^t \left\langle \nabla^2 a(w_{k\delta} + s(w_{(k+1)\delta} - w_{k\delta}), (w_{(k+1)\delta} - w_{k\delta})(w_{(k+1)\delta} - w_{k\delta})^T \right\rangle dtds\right]\\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(w_{k\delta}), w_{(k+1)\delta} - w_{k\delta}\right\rangle\right] + \mathbb{E}\left[\left\|(w_{(k+1)\delta} - w_{k\delta})\right\|_2^2 ds\right]\\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] + \mathbb{E}\left[\left\langle \nabla a(w_{k\delta}), -\delta\nabla U(w_{k\delta})\right\rangle\right] + 2\delta^2 \left\|\nabla U(w_{k\delta})\right\|_2^2 + 2\delta \mathbb{E}\left[\left\|\xi(w_{k\delta}, \eta_k)\right\|_2^2\right]\\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta\mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2\mathbb{E}\left[\left\|\nabla U(w_{k\delta})\right\|_2^2\right] + 2\delta\mathbb{E}\left[\left\|\xi(w_{k\delta}, \eta_k)\right\|_2^2\right]\\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta\mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2 L^2\mathbb{E}\left[\left\|w_{k\delta}\right\|_2^2\right] + 2\delta\beta^2\\
\leq&\mathbb{E}\left[a(w_{k\delta})\right] - m\delta\mathbb{E}\left[a(w_{k\delta})\right] + 2\delta^2 L^2 a(w_{k\delta}) + 2\delta^2 L^2 R^2 + 2\delta\beta^2\\
\leq&(1 - m\delta/2)a(w_{k\delta}) + m\delta R^2 + 2\delta\beta^2
\end{aligned}
$$

Where the first inequality uses the upper bound on $\left\|\nabla^2 a(y)\right\|_2$ above, the second inequality uses the fact that $w_{(k+1)\delta} = (y_{k\delta} - \delta\nabla U(y_{k\delta}) = \xi(w_{k\delta}, \eta_k))$, and $\mathbb{E}\left[\xi(w_{k\delta}, \eta_k) | w_{k\delta}\right] = 0$, the third inequality uses claim 2. at the start of this proof, the fourth inequality uses item 2 of Assumption B. The fifth inequality uses claim 3. above, the sixth inequality uses our assumption that $\delta \leq \frac{m}{16L^2}$.

Taking expectation wrt $w_{k\delta}$,

$$
\begin{aligned}
&\mathbb{E}\left[a(w_{(k+1)\delta})\right] \leq \mathbb{E}\left[a(w_k)\right] - m\delta\big(\mathbb{E}\left[a(w_{k\delta})\right] - 2R^2 + 2\beta^2/m\big)\\
\Rightarrow\quad &\mathbb{E}\left[a(w_{(k+1)\delta})\right] - (2R^2/2 + 2\beta^2/m) \leq (1 - m\delta)\big(\mathbb{E}\left[a(w_{k\delta})\right] - (2R^2 + 2\beta^2/m)\big)
\end{aligned}
$$

Thus, if $\mathbb{E}\left[\left\|w_0\right\|_2^2\right] \leq 2R^2 + 2\beta^2/m$, then $\mathbb{E}\left[a(w_0)\right] - \big(2R^2 + 2\beta^2/m\big) \leq 0$, then $\mathbb{E}\left[a(w_{k\delta})\right] - \big(2R^2 + 2\beta^2/m\big) \leq 0$ for all k, which implies that

$$
\mathbb{E}\left[\left\|w_{k\delta}\right\|_2^2\right] \leq 2\mathbb{E}\left[a(w_{k\delta})\right] + 4R^2 \leq 8\big(R^2 + \beta^2/m\big)
$$

for all k. ■

C.2. Divergence Bounds

Lemma 10 *Let x_t be as defined in (20) (or equivalently (22) or (27)), initialized at x_0. Then for any $T \leq \frac{1}{16L}$,*

$$
\mathbb{E}\left[\left\|x_T - x_0\right\|_2^2\right] \leq 8\big(T\beta^2 + T^2L^2\mathbb{E}\left[\left\|x_0\right\|_2^2\right]\big)
$$

If we additionally assume that $\mathbb{E}\left[\left\|x_0\right\|_2^2\right] \leq 8\big(R^2 + \beta^2/m\big)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$
\mathbb{E}\left[\left\|x_T - x_0\right\|_2^2\right] \leq 16T\beta^2
$$

Proof

By Ito's Lemma,

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\|x_t\|_2^2\right] \\
=&2\mathbb{E}\left[\langle \nabla U(x_t), x_t - x_0\rangle\right] + \mathbb{E}\left[\text{tr}\big(M(x_t)^2\big)\right] \\
\leq&2L\mathbb{E}\left[\|x_t\|_2\|x_t - x_0\|_2\right] + \beta^2 \\
\leq&2L\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + 2L\mathbb{E}\left[\|x_0\|_2\|x_t - x_0\|_2\right] + \beta^2 \\
\leq&2L\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + L^2T\mathbb{E}\left[\|x_0\|_2^2\right] + \frac{1}{T}\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + \beta^2 \\
\leq&\frac{2}{T}\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + \left(L^2T\mathbb{E}\left[\|x_0\|_2^2\right] + \beta^2\right)
\end{aligned}$$

where the first inequality is by item 1 of Assumption A and item 2 of Assumption B, the second inequality is by triangle inequality, the third inequality is by Young's inequality, the last inequality is by our assumption on T.

Applying Gronwall's inequality for $t \in [0, T]$,

$$\begin{aligned}
&\left(\mathbb{E}\left[\|x_t - x_0\|_2^2\right] + L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2\right) \\
\leq&e^2\left(\mathbb{E}\left[\|x_0 - x_0\|\right] + L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2\right) \\
\leq&8L^2T^2\mathbb{E}\left[\|x_0\|_2^2\right] + T\beta^2
\end{aligned}$$

This concludes our proof. ∎

Lemma 11 *Let y_t be as defined in (21) (or equivalently (22) or (27)), initialized at y_0. Then for any T,*

$$\mathbb{E}\left[\|y_T - y_0\|_2^2\right] \leq T^2L^2\mathbb{E}\left[\|y_0\|_2^2\right] + T\beta^2$$

If we additionally assume that $\mathbb{E}\left[\|y_0\|_2^2\right] \leq 8\big(R^2 + \beta^2/m\big)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|y_T - y_0\|_2^2\right] \leq 2T\beta^2$$

Proof
Notice from the definition in (21) that $y_T - y_0 \sim \mathcal{N}\big(-T\nabla U(y_0), TM(y_0)^2\big)$, the conclusion immediately follows from where the inequality is by item 1 of Assumption A and item 2 of Assumption B, and the fact that

$$\text{tr}\big(M(x)^2\big) = \text{tr}\big(\mathbb{E}\left[\xi(x,\eta)\xi(x,\eta)^T\right]\big) = \mathbb{E}\left[\|\xi(x,\eta)\|_2^2\right]$$

∎

Lemma 12 *Let v_t be as defined in (29), initialized at v_0. Then for any $T = n\delta$,*

$$\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \leq T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2$$

If we additionally assume that $\mathbb{E}\left[\|v_0\|_2^2\right] \leq 8\big(R^2 + \beta^2/m\big)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \leq 2T\beta^2$$

Proof
From (29),

$$v_T - v_0 = -T\nabla U(v_0) + \sqrt{\delta}\sum_{i=0}^{n-1}\xi(v_0, \eta_i)$$

Conditioned on the randomness up to time i, $\mathbb{E}\left[\xi(v_0, \eta_{i+1})\right] = 0$. Thus

$$
\begin{aligned}
&\mathbb{E}\left[\|v_T - v_0\|_2^2\right] \\
=&T^2\mathbb{E}\left[\|\nabla U(v_0)\|_2^2\right] + \delta\sum_{i=0}^{n-1}\mathbb{E}\left[\|\xi(v_0,\eta_i)\|_2^2\right] \\
\leq&T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2
\end{aligned}
$$

where the inequality is by item 1 of Assumption A and item 2 of Assumption B. ■

Lemma 13 *Let w_t be as defined in (31), initialized at w_0. Then for any $T = n\delta$ such that $T \leq \frac{1}{2L}$,*

$$
\mathbb{E}\left[\|w_T - w_0\|_2^2\right] \leq 16\Big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\Big)
$$

If we additionally assume that $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 8\big(R^2 + \beta^2/m\big)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$
\mathbb{E}\left[\|w_T - w_0\|_2^2\right] \leq 32T\beta^2
$$

Proof

$$
\begin{aligned}
&\mathbb{E}\left[\left\|w_{(k+1)\delta} - w_0\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|w_{k\delta} - \delta\nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_{k\delta},\eta_k) - w_0\right\|_2^2\right] \\
=&\mathbb{E}\left[\|w_{k\delta} - \delta\nabla U(w_{k\delta}) - w_0\|_2^2\right] + \delta\mathbb{E}\left[\|\xi(w_{k\delta},\eta_k)\|_2^2\right]
\end{aligned}
\tag{36}
$$

We can bound $\delta\mathbb{E}\left[\|\xi(w_{k\delta},\eta_k)\|_2^2\right] \leq \delta\beta^2$ by item 2 of Assumption B.

$$
\begin{aligned}
&\mathbb{E}\left[\|w_{k\delta} - \delta\nabla U(w_{k\delta}) - w_0\|_2^2\right] \\
\leq&\mathbb{E}\left[\left(\|w_{k\delta} - w_0 - \delta(\nabla U(w_{k\delta}) - \nabla U(w_0))\|_2 + \delta\|\nabla U(w_0)\|_2\right)^2\right] \\
\leq&\left(1+\frac{1}{n}\right)\mathbb{E}\left[\|w_{k\delta} - w_0 - \delta(\nabla U(w_{k\delta}) - \nabla U(w_0))\|_2^2\right] \\
&\quad + (1+n)\delta^2\mathbb{E}\left[\|\nabla U(w_0)\|_2^2\right] \\
\leq&\left(1+\frac{1}{n}\right)(1+\delta L)^2\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] \\
\leq&e^{1/n+2\delta L}\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right]
\end{aligned}
$$

where the first inequality is by triangle inequality, the second inequality is by Young's inequality, the third inequality is by item 1 of Assumption A.

Inserting the above into (36) gives

$$
\mathbb{E}\left[\left\|w_{(k+1)\delta} - w_0\right\|_2^2\right] \leq e^{1/n+2\delta L}\mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] + 2n\delta^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + \delta\beta^2
$$

Applying the above recursively for $k = 1...n$, we see that

$$\begin{aligned}
&\mathbb{E}\left[\|w_{n\delta} - w_0\|_2^2\right] \\
\leq &\sum_{k=0}^{n-1} e^{(n-k)\cdot(1/n+2\delta L)} \cdot \left(2n\delta^2 L^2 \mathbb{E}\left[\|w_0\|_2^2\right] + \delta\beta^2\right) \\
\leq &16\left(n^2\delta^2 L^2 \mathbb{E}\left[\|w_0\|_2^2\right] + n\delta\beta^2\right) \\
= &16\left(T^2 L^2 \mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\right)
\end{aligned}$$

■

C.3. Discretization Bounds

Lemma 14 *Let $v_{k\delta}$ and $w_{k\delta}$ be as defined in (29) and (31). Then for any δ, n, such that $T := n\delta \leq \frac{1}{16L}$,*

$$\mathbb{E}\left[\|v_T - w_T\|_2^2\right] \leq 8\left(2T^2 L^2\left(T^2 L^2 \mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\right) + TL_\xi^2\left(16\left(T^2 L^2 \mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\right)\right)\right)$$

If we additionally assume that $\mathbb{E}\left[\|v_0\|_2^2\right] \leq 8\left(R^2 + \beta^2/m\right)$, $\mathbb{E}\left[\|w_0\|_2^2\right] \leq 8\left(R^2 + \beta^2/m\right)$ and $T \leq \frac{\beta^2}{8L^2(R^2+\beta^2/m)}$, then

$$\mathbb{E}\left[\|v_T - w_T\|_2^2\right] \leq 32\left(T^2 L^2 + TL_\xi^2\right)T\beta^2$$

Proof
Using the fact that conditioned on the randomness up to step k, $\mathbb{E}\left[\xi(v_0, \eta_{k+1}) - \xi(w_{k\delta}, \eta_{k+1})\right] = 0$, we can show that for any $k \leq n$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|v_{(k+1)\delta} - w_{(k+1)\delta}\right\|_2^2\right] \\
= &\mathbb{E}\left[\left\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta}) + \sqrt{\delta}\xi(w_0, \eta_k) - \sqrt{\delta}\xi(w_{k\delta}, \eta_k)\right\|_2^2\right] \\
= &\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] + \delta\mathbb{E}\left[\|\xi(w_0, \eta_k) - \xi(w_{k\delta}, \eta_k)\|_2^2\right]
\end{aligned} \tag{37}$$

where the first inequality is by (Assumption on smoothness of U and xi).

Using (smoothness of xi), and Lemma 12, we can bound

$$\begin{aligned}
&\delta\mathbb{E}\left[\|\xi(w_0, \eta_k) - \xi(w_{k\delta}, \eta_k)\|_2^2\right] \\
\leq &\delta L_\xi^2 \mathbb{E}\left[\|w_{k\delta} - w_0\|_2^2\right] \\
\leq &\delta L_\xi^2\left(16\left(T^2 L^2 \mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\right)\right)
\end{aligned}$$

We can also bound

$$\begin{aligned}
&\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_0) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] \\
\leq &\left(1 + \frac{1}{n}\right)\mathbb{E}\left[\|v_{k\delta} - \delta\nabla U(v_{k\delta}) - w_{k\delta} + \delta\nabla U(w_{k\delta})\|_2^2\right] + (1+n)\delta^2\mathbb{E}\left[\|\nabla U(v_{k\delta}) - \nabla U(v_0)\|_2^2\right] \\
\leq &\left(1 + \frac{1}{n}\right)(1 + \delta L)^2\mathbb{E}\left[\|v_{k\delta} - w_{k\delta}\|_2^2\right] + 2n\delta^2 L^2\mathbb{E}\left[\|v_{k\delta} - v_0\|_2^2\right] \\
\leq &e^{1/n+2\delta L} E\|v_{k\delta} - w_{k\delta}\|_2^2 + 2n\delta^2 L^2\mathbb{E}\left[\|v_{k\delta} - v_0\|_2^2\right] \\
\leq &e^{1/n+2\delta L} E\|v_{k\delta} - w_{k\delta}\|_2^2 + 2n\delta^2 L^2\left(T^2 L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\right)
\end{aligned}$$

where the first inequality is by Young's inequality and the second inequality is by item 1 of Assumption A, the fourth inequality uses Lemma 12.

Substituting the above two equation blocks into (37), and applying recursively for $k = 0...n-1$ gives

$$
\begin{aligned}
&\mathbb{E}\left[\|v_T - w_T\|_2^2\right]\\
=&\mathbb{E}\left[\|v_{n\delta} - w_{n\delta}\|_2^2\right]\\
\le& e^{1+2n\delta L}\Big(2n^2\delta^2 L^2\Big(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\Big) + n\delta L_\xi^2\Big(16\Big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\Big)\Big)\Big)\\
\le& 8\Big(2T^2L^2\Big(T^2L^2\mathbb{E}\left[\|v_0\|_2^2\right] + T\beta^2\Big) + TL_\xi^2\Big(16\Big(T^2L^2\mathbb{E}\left[\|w_0\|_2^2\right] + T\beta^2\Big)\Big)\Big)
\end{aligned}
$$

the last inequality is by noting that $T = n\delta \le \frac{1}{4L}$. ■

D. Regularity of M and N

Lemma 15

$$
\begin{aligned}
&1.\ tr\big(M(x)^2\big) \le \beta^2\\
&2.\ tr\big((M(x)^2 - M(y)^2)^2\big) \le 16\beta^2 L_\xi^2\|x-y\|_2^2\\
&3.\ tr\big((M(x)^2 - M(y)^2)^2\big) \le 32\beta^3 L_\xi\|x-y\|_2
\end{aligned}
$$

Proof

In this proof, we will use the fact that $\xi(\cdot,\eta)$ is L_ξ-Lipschitz from Assumption B.

The first property is easy to see:

$$
\begin{aligned}
&\mathrm{tr}\big(M(x)^2\big)\\
=&\mathrm{tr}\big(\mathbb{E}_\eta\left[\xi(x,\eta)\xi(x,\eta)^T\right]\big)\\
=&\mathbb{E}_\eta\left[\mathrm{tr}\big(\xi(x,\eta)\xi(x,\eta)^T\big)\right]\\
=&\mathbb{E}_\eta\left[\|\xi(x,\eta)\|_2^2\right]\\
\le&\beta^2
\end{aligned}
$$

We now prove the second and third claims. Consider a fixed x and fixed y, let $u_\eta := \xi(x,\eta)$, $v_\eta := \xi(y,\eta)$. Then

$$
\begin{aligned}
&\mathrm{tr}\Big(\big(M(x)^2 - M(y)^2\big)^2\Big)\\
=&\mathrm{tr}\Big(\big(\mathbb{E}_\eta\left[u_\eta u_\eta^T - v_\eta v_\eta^T\right]\big)^2\Big)\\
=&\mathrm{tr}\big(\mathbb{E}_{\eta,\eta'}\left[\big(u_\eta u_\eta^T - v_\eta v_\eta^T\big)\big(u_{\eta'}u_{\eta'}^T - v_{\eta'}v_{\eta'}^T\big)\right]\big)\\
=&\mathbb{E}_{\eta,\eta'}\left[\mathrm{tr}\big(\big(u_\eta u_\eta^T - v_\eta v_\eta^T\big)\big(u_{\eta'}u_{\eta'}^T - v_{\eta'}v_{\eta'}^T\big)\big)\right]
\end{aligned}
$$

For any fixed η and η', let's further simplify notation by letting u, u', v, v' denote $u_\eta, u_{\eta'}, v_\eta, v_{\eta'}$. Thus

$$
\begin{aligned}
&\mathrm{tr}\big(\big(uu^T - vv^T\big)\big(u'u'^T - v'v'^T\big)\big)\\
=&\mathrm{tr}\big(\big((u-v)v^T + v(u-v)^T + (u-v)(u-v)^T\big)\big((u'-v')v'^T + v'(u'-v')^T + (u'-v')(u'-v')^T\big)\big)\\
=&\mathrm{tr}\big((u-v)v^T(u'-v')v'^T\big) + \mathrm{tr}\big((u-v)v^Tv'(u'-v')^T\big) + \mathrm{tr}\big((u-v)v^T(u'-v')(u'-v')^T\big)\\
&+\mathrm{tr}\big(v(u-v)^T(u'-v')v'^T\big) + \mathrm{tr}\big(v(u-v)^Tv'(u'-v')^T\big) + \mathrm{tr}\big(v(u-v)^T(u'-v')(u'-v')^T\big)\\
&+\mathrm{tr}\big((u-v)(u-v)^T(u'-v')v'^T\big) + \mathrm{tr}\big((u-v)(u-v)^Tv'(u'-v')^T\big)\\
&+\mathrm{tr}\big((u-v)(u-v)^T(u'-v')(u'-v')^T\big)\\
\le&\min\left\{16\beta^2L_\xi^2\|x-y\|_2^2, 32\beta^3L_\xi\|x-y\|_2\right\}
\end{aligned}
$$

Where the last inequality uses Assumption B.2 and B.3; in particular, $\|v\|_2 \leq \beta$ and $\|u-v\|_2 \leq \min\{2\beta, L_\xi\|x-y\|_2\}$. This proves 2. and 3. of the Lemma statement. ■

Lemma 16 *Let $N(x)$ be as defined in (6) and L_N be as defined in (7). Then*

$$1.\ tr\big(N(x)^2\big) \leq \beta^2$$
$$2.\ tr\Big(\big(N(x)-N(y)\big)^2\Big) \leq L_N^2\|x-y\|_2^2$$
$$3.\ tr\Big(\big(N(x)-N(y)\big)^2\Big) \leq \frac{8\beta^2}{c_m}\cdot L_N\|x-y\|_2.$$

Proof of Lemma 16
The first inequality holds because $N(x)^2 := M(x)^2 - c_m^2 I$, and then applying Lemma 15.1, and the fact that $\mathrm{tr}\big(M(x)^2 - c_m^2 I\big) \leq \mathrm{tr}\big(M(x)^2\big)$ by Assumption B.4.

The second inequality is a immediate consequence of Lemma 17, Lemma 15.2, and the fact that $\lambda_{min}\big(N(x)^2\big) = \lambda_{min}\big(M(x)^2 - c_m^2\big) \geq c_m^2$ by Assumption B.4.

The proof for the third inequality is similar to the second inequality, and follows from Lemma 15 and Lemma 17.

■

Lemma 17 (Simplified version of Lemma 1 from (Eldan et al., 2018)) *Let A, B be positive definite matrices. Then*

$$tr\bigg(\Big(\sqrt{A}-\sqrt{B}\Big)^2\bigg) \leq tr\big((A-B)^2A^{-1}\big)$$

E. Defining f and related inequalities

In this section, we define the Lyapunov function f which is central to the proof of our main results. Here, we give an overview of the various functions defined in this section:

1. $g(z):\mathbb{R}^d \to \mathbb{R}^+$: A smoothed version of $\|z\|_2$, with bounded derivatives up to third order.
2. $q(r):\mathbb{R}^+ \to \mathbb{R}^+$: A concave potential function, similar to the one defined in (Eberle, 2016), which has bounded derivatives up to third order everywhere except at $r=0$.
3. $f(z) = q(g(z)):\mathbb{R}^d \to \mathbb{R}^+$, a concave function which upper and lower bounds $\|z\|_2$ within a constant factor, has bounded derivatives up to third order everywhere.

Lemma 18 (Properties of f) *Let ϵ satisfy $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q{\mathcal{R}_q}^2+1}$. We define the function*

$$f(z) := q(g(z))$$

Where q is as defined in (39) Appendix E.1, and g is as defined in Lemma 20 (with parameter ϵ). Then

1. *(a)* $\nabla f(z) = q'(g(z))\cdot\nabla g(z)$
   (b) For $\|z\|_2 \geq 2\epsilon$, $\nabla f(z) = q'(g(z))\frac{z}{\|z\|_2}$
   (c) For all z, $\|\nabla f(z)\|_2 \leq 1$.
2. *(a)* $\nabla^2 f(z) = q''(g(z))\nabla g(z)\nabla g(z)^T + q'(g(z))\nabla^2 g(z)$
   (b) For $r \geq 2\epsilon$, $\nabla^2 f(z) = q''(g(z))\frac{zz^T}{\|z\|_2^2} + q'(g(z))\frac{1}{\|z\|_2}\Big(I - \frac{zz^T}{\|z\|_2^2}\Big)$
   (c) For all z, $\big\|\nabla^2 f(z)\big\|_2 \leq \frac{2}{\epsilon}$
   (d) For all z, v, $v^T\nabla^2 f(z)v \leq \frac{q'(g(z))}{\|z\|_2}$
3. *For any* z, $\big\|\nabla^3 f(z)\big\|_2 \leq \frac{9}{\epsilon^2}$
4. *For any* z, $f(z) \in \Big[\frac{1}{2}\exp\Big(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\Big)g(\|z\|_2), g(\|z\|_2)\Big] \in \Big[\frac{1}{2}\exp\Big(-\frac{7\alpha_q{\mathcal{R}_q}^2}{3}\Big)(\|z\|_2 - 2\epsilon), \|z\|_2\Big]$

Proof of Lemma 18

1. (a) chain rule
   (b) Use definition of $\nabla g(z)$ from Lemma 20.
   (c) By definition, $\nabla f(z) = q'(g(z))\nabla g(z)$. From Lemma 21, $|q'(g(z))| \leq 1$. By definition, $\nabla g(z) = h'(\|z\|_2)\frac{z}{\|z\|_2}$. Our conclusion follows from $h' \leq 1$ using item 2 of Lemma 19.
2. (a) chain rule
   (b) by item 2 b) of Lemma 20
   (c) by item 1 c) and item 2 d) of Lemma 20, and item 3 and item 4 of Lemma 21, and our assumption that $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q + \mathcal{R}_q{}^2 + 1}$.
   (d) by item 4 of Lemma 21), and items 2 c) and 2 d) of Lemma 20, and our expression for $\nabla^2 f(z)$ established in item 2 a).
3. It can be verified that

$$\begin{aligned}\nabla^3 f(z) =& q'''(g(z)) \cdot \nabla g(z)^{\bigotimes 3} + q''(g(z))\nabla g(z) \bigotimes \nabla^2 g(z) + q''(g(z))\nabla^2 g(z) \bigotimes \nabla g(z)\\ &+ q''(g(z))\nabla g(z) \bigotimes \nabla^2 g(z) + q'(g(z))\nabla^3 g(z)\end{aligned}$$

   Thus

$$\begin{aligned}\left\|\nabla^3 f(z)\right\|_2 \leq& |q'''(g(z))| \|\nabla g(z)\|_2^3 + 3q''(g(z))\|\nabla g(z)\|_2 \left\|\nabla^2 g(z)\right\|_2 + q'(g(z))\left\|\nabla^3 g(z)\right\|\\ \leq& 5\left(\alpha_q + \frac{1}{\mathcal{R}_q{}^2}\right)\left(\alpha_q \mathcal{R}_q{}^2 + 1\right) + 3\left(\frac{5\alpha_q\mathcal{R}_q}{4} + \frac{4}{\mathcal{R}_q}\right) \cdot \frac{1}{\epsilon} + \frac{1}{\epsilon^2}\\ \leq& \frac{9}{\epsilon^2}\end{aligned}$$

   Where the first inequality uses Lemma 21 and Lemma 20, and the second inequality assumes that $\epsilon \leq \frac{\mathcal{R}_q}{\alpha_q\mathcal{R}_q{}^2+1}$

4.

$$f(z) \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q\mathcal{R}_q{}^2}{3}\right)g(\|z\|_2), g(\|z\|_2)\right] \in \left[\frac{1}{2}\exp\left(-\frac{7\alpha_q\mathcal{R}_q{}^2}{3}\right)(\|z\|_2 - 2\epsilon), \|z\|_2\right]$$

   The first containment is by Lemma 21.2.: $\frac{1}{2}\exp\left(-\frac{7\alpha_q\mathcal{R}_q{}^2}{3}\right) \cdot g(z) \leq q(g(z)) \leq g(z)$. THe second containment is by Lemma 20.4: $g(\|z\|_2) \in [\|z\|_2 - 2\epsilon, \|z\|_2]$.

■

Lemma 19 (Properties of h) *Given a parameter ϵ, define*

$$h(r) := \begin{cases} \frac{r^3}{6\epsilon^2}, & \textit{for } r \in [0, \epsilon]\\ \frac{\epsilon}{6} + \frac{r-\epsilon}{2} + \frac{(r-\epsilon)^2}{2\epsilon} - \frac{(r-\epsilon)^3}{6\epsilon^2}, & \textit{for } r \in [\epsilon, 2\epsilon]\\ r, & \textit{for } r \geq 2\epsilon\end{cases}$$

1. The derivatives of h are as follows:

$$h'(r) = \begin{cases} \frac{r^2}{2\epsilon^2}, & \textit{for } r \in [0, \epsilon]\\ \frac{1}{2} + \frac{r-\epsilon}{\epsilon} - \frac{(r-\epsilon)^2}{2\epsilon^2}, & \textit{for } r \in [\epsilon, 2\epsilon]\\ 1, & \textit{for } r \geq 2\epsilon\end{cases}$$

$$h''(r) = \begin{cases} \frac{r}{\epsilon^2}, & \textit{for } r \in [0, \epsilon]\\ \frac{1}{\epsilon} - \frac{r-\epsilon}{\epsilon^2}, & \textit{for } r \in [\epsilon, 2\epsilon]\\ 0, & \textit{for } r \geq 2\epsilon\end{cases}$$

$$h'''(r) = \begin{cases} \frac{1}{\epsilon^2}, & \textit{for } r \in [0, \epsilon]\\ -\frac{1}{\epsilon^2}, & \textit{for } r \in [\epsilon, 2\epsilon]\\ 0, & \textit{for } r \geq 2\epsilon\end{cases}$$

2. (a) h' *is positive, motonically increasing.*
   (b) $h'(0) = 0$, $h'(r) = 1$ *for* $r \geq \epsilon$
   (c) $\frac{h'(r)}{r} \leq \min\left\{\frac{1}{\epsilon}, \frac{1}{r}\right\}$ *for all* r
3. (a) $h''(r)$ *is positive*
   (b) $h''(r) = 0$ *for* $r = 0$ *and* $r \geq 2\epsilon$
   (c) $h''(r) \leq \frac{1}{\epsilon}$
   (d) $\frac{h''(r)}{r} \leq \frac{1}{\epsilon^2}$
4. $|h'''(r)| \leq \frac{1}{\epsilon^2}$
5. $r - 2\epsilon \leq h(r) \leq r$

Proof of Lemma 19
The claims can all be verified with simple algebra. ■

Lemma 20 (Properties of g) *Given a parameter* ϵ*, let us define*

$$g(z) := h(\|z\|_2)$$

Where h *is as defined in Lemma 19 (using parameter* ϵ*). Then*

1. (a) $\nabla g(z) = h'(\|z\|_2)\frac{z}{\|z\|_2}$
   (b) *For* $\|z\|_2 \geq 2\epsilon$, $\nabla g(z) = \frac{z}{\|z\|_2}$.
   (c) *For any* $\|z\|_2$, $\|\nabla g(z)\|_2 \leq 1$
2. (a) $\nabla^2 g(z) = h''(\|z\|_2)\frac{zz^T}{\|z\|_2^2} + h'(\|z\|_2)\frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right)$
   (b) *For* $\|z\|_2 \geq 2\epsilon$, $\nabla^2 g(z) = \frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right)$.
   (c) *For* $\|z\|_2 \geq 2\epsilon$, $\left\|\nabla^2 g(z)\right\|_2 = \frac{1}{\|z\|_2}$
   (d) *For all* z, $\left\|\nabla^2 g(z)\right\|_2 \leq \frac{1}{\epsilon}$
3. $\left\|\nabla^3 g(z)\right\|_2 \leq \frac{5}{\epsilon^2}$
4. $\|z\|_2 - 2\epsilon \leq g(z) \leq \|z\|_2$.

Proof of Lemma 20
All the properties can be verified with algebra. We provide a proof for 3. since it is a bit involved.

Let us define the functions $\kappa^1(z) = \nabla(\|z\|_2), \kappa^2(z) = \nabla^2(\|z\|_2), \kappa^3(z) = \nabla^3(\|z\|_2)$. Specifically,

$$\begin{aligned}
\kappa^1(z) =& \frac{z}{\|z\|_2} \\
\kappa^2(z) =& \frac{1}{\|z\|_2}\left(I - \frac{zz^T}{\|z\|_2^2}\right) \\
\kappa^3(z) =& -\frac{1}{\|z\|_2^2}\frac{z}{\|z\|_2}\bigotimes\left(I - \frac{zz^T}{\|z\|_2^2}\right) + \frac{1}{\|z\|_2}\left(\frac{z}{\|z\|_2}\bigotimes\kappa^2(z) + \kappa^2(z)\bigotimes\frac{z}{\|z\|_2}\right)
\end{aligned}$$

It can be verified that

$$\begin{aligned}
\left\|\kappa^2(z)\right\|_2 =& \frac{1}{\|z\|_2} \\
\left\|\kappa^3(z)\right\|_2 =& \frac{1}{\|z\|_2^2}
\end{aligned}$$

It can be verified that $\nabla^2 g(z)$ has the following form:

$$\nabla^3 g(z) = h'''(\|z\|_2)\big(\kappa^1(z)\big)^{\bigotimes 3} + h''(\|z\|_2)\kappa^1(z) \bigotimes \kappa^2(z) + h''(\|z\|_2)\kappa^2(z) \bigotimes \kappa^1(z)$$
$$+ h'(\|z\|_2)\kappa^3(z) + h''(\|z\|_2)\kappa^1(z) \bigotimes \kappa^2(z)$$

Thus

$$\left\|\nabla^3 g(z)\right\|_2 \le |h'''(\|z\|_2)| + 3\frac{h''(\|z\|_2)}{\|z\|_2} + \frac{h'(\|z\|_2)}{\|z\|_2^2} \le \frac{5}{\epsilon^2}$$

Where we use properties of h from Lemma 19.

The last claim follows immediately from Lemma 19.4. ■

E.1. Defining q

In this section, we define the function q that is used in Lemma 18. Our construction is a slight modification to the original construction in (Eberle, 2011).

Let α_q and $\mathcal{R}_q$ be as defined in (7). We begin by defining auxiliary functions $\psi(r)$, $\Psi(r)$ and $\nu(r)$, all from $\mathbb{R}^+$ to $\mathbb{R}$:

$$\psi(r) := e^{-\alpha_q \tau(r)}, \qquad \Psi(r) := \int_0^r \psi(s)ds, \qquad \nu(r) := 1 - \frac{1}{2}\frac{\int_0^r \frac{\mu(s)\Psi(s)}{\psi(s)}ds}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)}ds}, \tag{38}$$

Where $\tau(r)$ and $\mu(r)$ are as defined in Lemma 22 and Lemma 23 with $\mathcal{R} = \mathcal{R}_q$.

Finally we define q as

$$q(r) := \int_0^r \psi(s)\nu(s)ds. \tag{39}$$

We now state some useful properties of the distance function q.

Lemma 21 *The function q defined in (39) has the following properties.*

1. *For all $r \le \mathcal{R}_q$, $q''(r) + \alpha_q q'(r) \cdot r \le -\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{32{\mathcal{R}_q}^2} q(r)$*
2. *For all r, $\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \cdot r \le q(r) \le r$*
3. *For all r, $\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \le q'(r) \le 1$*
4. *For all r, $q''(r) \le 0$ and $|q''(r)| \le \left(\frac{5\alpha_q \mathcal{R}_q}{4} + \frac{4}{\mathcal{R}_q}\right)$*
5. *For all r, $|q'''(r)| \le 5\alpha_q + 2\alpha_q\left(\alpha_q {\mathcal{R}_q}^2 + 1\right) + \frac{2(\alpha_q {\mathcal{R}_q}^2 + 1)}{{\mathcal{R}_q}^2}$*

Proof of Lemma 21

Proof of 1. It can be verified that

$$\psi'(r) = \psi(r)(-\alpha_q \tau'(r))$$
$$\psi''(r) = \psi(r)\Big(\big(\alpha_q \tau'(r)\big)^2 + \alpha_q \tau''(r)\Big)$$
$$\nu'(r) = -\frac{1}{2}\frac{\frac{\mu(r)\Psi(r)}{\psi(r)}}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)}ds}$$

For $r \in [0, \mathcal{R}_q]$, $\tau'(r) = r$, so that $\psi'(r) = \psi(r)(-\alpha_q r)$. Thus

$$\begin{aligned}
q'(r) =& \psi(r)\nu(r) \\
q''(r) =& \psi'(r)\nu(r) + \psi(r)\nu'(r) \\
=& \psi(r)\nu(r)(-\alpha_q r) + \psi(r)\nu'(r) \\
=& -\alpha_q r \nu'(r) + \psi(r)\nu'(r) \\
q''(r) + \alpha_q r q'(r) =& \psi(r)\nu'(r) \\
=& -\frac{1}{2}\frac{\mu(r)\Psi(r)}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \\
=& -\frac{1}{2}\frac{\Psi(r)}{\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds}
\end{aligned}$$

Where the last equality is by definition of $\mu(r)$ in Lemma 23 and the fact that $r \leq \mathcal{R}_q$.

We can upper bound

$$\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds \leq \int_0^{4\mathcal{R}_q} \frac{\Psi(s)}{\psi(s)} ds \leq \frac{\int_0^{4\mathcal{R}_q} s ds}{\psi(4\mathcal{R}_q)} = \frac{16{\mathcal{R}_q}^2}{\psi(4\mathcal{R}_q)} \leq 16{\mathcal{R}_q}^2 \cdot \exp\left(\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)$$

Where the first inequality is by Lemma 23, the second inequality is by the fact that $\psi(s)$ is monotonically decreasing, the third inequality is by Lemma 22.

Thus

$$\begin{aligned}
q''(r) + \alpha_q r q'(r) \leq& -\frac{1}{2}\left(\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{16{\mathcal{R}_q}^2}\right)\Psi(r) \\
\leq& -\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{32{\mathcal{R}_q}^2} q(r)
\end{aligned}$$

Where the last inequality is by $\Psi(r) \geq q(r)$.

Proof of 2. Notice first that $\nu(r) \geq \frac{1}{2}$ for all r. Thus

$$\begin{aligned}
q(r) :=& \int_0^r \psi(s)\nu(s) ds \\
\geq& \frac{1}{2}\int_0^r \psi(s) ds \\
\geq& \frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \cdot r
\end{aligned}$$

Where the last inequality is by Lemma 22.

Proof of 3. By definition of f, $q'(r) = \psi(r)\nu(r)$, and

$$\frac{\exp\left(-\frac{7\alpha_q {\mathcal{R}_q}^2}{3}\right)}{2} \leq \psi(r)\nu(r) \leq 1$$

Where we use Lemma 22 and the fact that $\nu(r) \in [1/2, 1]$

Proof of 4. Recall that

$$q''(r) = \psi'(r)\nu(r) + \psi(r)\nu'(r)$$

That $q'' \leq 0$ can immediately be verified from the definitions of ψ and ν.

Thus

$$\begin{aligned}|q''(r)| \leq &|\psi'(r)\nu(r)| + |\psi(r)\nu'(r)| \\ \leq &\alpha_q\tau'(r) + |\psi(r)\nu'(r)|\end{aligned}$$

From Lemma 22, we can upperbound $\tau'(r) \leq \frac{5\mathcal{R}_q}{4}$. In addition, $\Psi(r) = \int_0^r \psi(s) \geq r\psi(r)$, so that

$$\frac{\Psi(r)}{\psi(r)} \geq r \tag{40}$$

(Recall again that $\psi(s)$ is monotonically decreasing). Thus $\Psi(r)/r \geq r$ for all r. In addition, using the fact that $\psi(r) \leq 1$,

$$\Psi(r) = \int_0^r \psi(s)ds \leq r \tag{41}$$

Combining the previous expressions,

$$\begin{aligned}|\psi(r)\nu'(r)| =&\left|\frac{1}{2}\frac{\mu(r)\Psi(r)}{\int_0^{4\mathcal{R}_q}\frac{\mu(s)\Psi(s)}{\psi(s)}ds}\right| \\ \leq&\left|\frac{1}{2}\frac{\mu(r)r}{\int_0^{\mathcal{R}_q}\frac{\Psi(s)}{\psi(s)}ds}\right| \\ \leq&\left|\frac{1}{2}\frac{4\mathcal{R}_q}{\int_0^{\mathcal{R}_q}sds}\right| \\ \leq&\frac{4}{\mathcal{R}_q}\end{aligned}$$

Where the first inequality are by definition of $\mu(r)$ and (41), and the second inequality is by (40) and the fact that $\mu(r) = 0$ for $r \geq 4\mathcal{R}_q$. Combining with our bound on $\psi'(r)\nu(r)$ gives the desired bound.

Proof of 5.

$$q'''(r) = \psi''(r)\nu(r) + 2\psi'(r)\nu'(r) + \psi(r)\nu''(r)$$

We first bound the middle term:

$$\begin{aligned}|\psi'(r)\nu'r)| =&|\psi(r)(\alpha_q\tau'(r))\nu'r)| \\ \leq&\alpha_q|\tau'(r)||\psi(r)\nu'r)| \\ \leq&\frac{5\alpha_q\mathcal{R}_q}{4}\cdot\frac{4}{\mathcal{R}_q} \\ \leq&5\alpha_q\end{aligned}$$

Where the second last line follows form Lemma 22 and our proof of 4..

Next,

$$\psi''(r) = \psi(r)\left(\alpha_q^2\tau'(r)^2 - \alpha_q\tau''(r)\right)$$

Thus applying Lemma 22.1 and Lemma 22.3,

$$|\psi''(r)\nu(r)| \leq 2\alpha_q^2{\mathcal{R}_q}^2 + \alpha_q$$

Finally,

$$\nu''(r) = \frac{1}{2\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \cdot \frac{d}{dr} \mu(r)\Psi(r)/\psi(r)$$

Expanding the numerator,

$$\begin{aligned}\frac{d}{dr}\frac{\mu(r)\Psi(r)}{\psi(r)} =& \mu'(r)\frac{\Psi(r)}{\psi(r)} + \mu(r) - \mu(r)\frac{\Psi(r)\psi'(r)}{\psi(r)^2} \\ =& \mu'(r)\frac{\Psi(r)}{\psi(r)} + \mu(r) + \mu(r)\frac{\Psi(r)\psi(r)\alpha_q\tau'(r)}{\psi(r)^2}\end{aligned}$$

Thus

$$\psi(r)\nu''(r) = \frac{1}{2\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \cdot (\mu'(r)\Psi(r) + \mu(r)\psi(r) + \mu(r)\Psi(r)\alpha_q\tau'(r))$$

Using the same argument as from the proof of 4., we can bound

$$\begin{aligned}\frac{1}{2\int_0^{4\mathcal{R}_q} \frac{\mu(s)\Psi(s)}{\psi(s)} ds} \leq& \frac{1}{2\int_0^{\mathcal{R}_q} s ds} \\ \leq& \frac{1}{{\mathcal{R}_q}^2}\end{aligned}$$

Finally, from Lemma 23, $|\mu'(r)| \leq \frac{\pi}{6\mathcal{R}_q}$, so

$$\begin{aligned}|\psi(r)\nu''(r)| \leq& \frac{\pi/6 + 1 + 5\alpha_q{\mathcal{R}_q}^2/4}{{\mathcal{R}_q}^2} \\ \leq& \frac{2(\alpha_q{\mathcal{R}_q}^2 + 1)}{{\mathcal{R}_q}^2}\end{aligned}$$

■

Lemma 22 *Let $\tau(r) : [0, \infty) \to \mathbb{R}$ be defined as*

$$\tau(r) = \begin{cases} \frac{r^2}{2}, & \textit{for } r \leq \mathcal{R} \\ \frac{\mathcal{R}^2}{2} + \mathcal{R}(r - \mathcal{R}) + \frac{(r-\mathcal{R})^2}{2} - \frac{(r-\mathcal{R})^3}{3\mathcal{R}}, & \textit{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ \frac{5\mathcal{R}^2}{3} + \mathcal{R}(r - 2\mathcal{R}) - \frac{(r-2\mathcal{R})^2}{2} + \frac{(r-2\mathcal{R})^3}{12\mathcal{R}}, & \textit{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ \frac{7\mathcal{R}^2}{3}, & \textit{for } r \geq 4\mathcal{R}] \end{cases}$$

Then

1. $\tau'(r) \in [0, \frac{5\mathcal{R}}{4}]$*, with maxima at* $r = \frac{3\mathcal{R}}{2}$*.* $\tau'(r) = 0$ *for* $r \in \{0\} \bigcup [4\mathcal{R}, \infty)$
2. *As a consequence of 1,* $\tau(r)$ *is monotonically increasing*
3. $\tau''(r) \in [-1, 1]$

Proof of Lemma 22
We provide the derivatives of τ below. The claims in the Lemma can then be immediately verified.

$$\tau'(r) = \begin{cases} r, & \text{for } r \leq \mathcal{R} \\ \mathcal{R} + (r - \mathcal{R}) - \frac{(r-\mathcal{R})^2}{\mathcal{R}}, & \text{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ \mathcal{R} - (r - 2\mathcal{R}) + \frac{(r-2\mathcal{R})^2}{4\mathcal{R}}, & \text{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ 0, & \text{for } r \geq 4\mathcal{R}] \end{cases}$$

$$\tau''(r) = \begin{cases} 1, & \text{for } r \le \mathcal{R} \\ 1 - \frac{2(r-\mathcal{R})}{\mathcal{R}}, & \text{for } r \in [\mathcal{R}, 2\mathcal{R}] \\ -1 + \frac{r-2\mathcal{R}}{2\mathcal{R}}, & \text{for } r \in [2\mathcal{R}, 4\mathcal{R}] \\ 0, & \text{for } r \ge 4\mathcal{R}] \end{cases}$$

■

Lemma 23 *Let*

$$\mu(r) := \begin{cases} 1, & \textit{for } r \le \mathcal{R} \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{\pi(r-\mathcal{R})}{3\mathcal{R}}\right), & \textit{for } r \in [\mathcal{R}, 4\mathcal{R}] \\ 0, & \textit{for } r \ge 4\mathcal{R} \end{cases}$$

Then

$$\mu'(r) := \begin{cases} 0, & \textit{for } r \le \mathcal{R} \\ -\frac{\pi}{6\mathcal{R}}\sin\left(\frac{\pi(r-\mathcal{R})}{\mathcal{R}}\right), & \textit{for } r \in [\mathcal{R}, 4\mathcal{R}] \\ 0, & \textit{for } r \ge 4\mathcal{R} \end{cases}$$

Furthermore, $\mu'(r) \in [-\frac{\pi}{6\mathcal{R}}, 0]$

This Lemma can be easily verified by algebra.

F. Miscellaneous

The following Theorem, taken from (Eldan et al., 2018), establishes a quantitative CLT.

Theorem 5 *Let* $X_1 ... X_n$ *be random vectors with mean 0, covariance* Σ*, and* $\|X_i\| \le \beta$ *almost surely for each* i*. Let* $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$*, and let* Z *be a Gaussian with covariance* Σ*, then*

$$W_2(S_n, Z) \le \frac{6\sqrt{d}\beta\sqrt{\log n}}{\sqrt{n}}$$

Corollary 24 *Let* $X_1 ... X_n$ *be random vectors with mean 0, covariance* Σ*, and* $\|X_i\| \le \beta$ *almost surely for each* i*. let* Y *be a Gaussian with covariance* $n\Sigma$*. Then*

$$W_2\left(\sum_i X_i, Y\right) \le 6\sqrt{d}\beta\sqrt{\log n}$$

This is simply taking the result of Theorem 5 and scaling the inequality by $\sqrt{n}$ on both sides.

The following Lemma is taken from (Cheng et al., 2019) and included here for completeness.

Lemma 25 *For any* $c > 0$*,* $x > 3\max\left\{\frac{1}{c}\log\frac{1}{c}, 0\right\}$*, the inequality*

$$\frac{1}{c}\log(x) \le x$$

holds.

Proof
We will consider two cases:

Case 1: If $c \ge \frac{1}{e}$, then the inequality

$$\log(x) \le cx$$

is true for all x.

Case 2: $c \leq \frac{1}{e}$.

In this case, we consider the Lambert W function, defined as the inverse of $f(x) = xe^x$. We will particularly pay attention to W_{-1} which is the lower branch of W. (See Wikipedia for a description of W and W_{-1}).

We can lower bound $W_{-1}(-c)$ using Theorem 1 from (Chatzigeorgiou, 2013):

$$\begin{aligned}
&\forall u > 0, \quad W_{-1}(-e^{-u-1}) > -u - \sqrt{2u} - 1 \\
\text{equivalently} \quad &\forall c \in (0, 1/e), \quad -W_{-1}(-c) < \log\left(\frac{1}{c}\right) + 1 + \sqrt{2\left(\log\left(\frac{1}{c}\right) - 1\right)} - 1 \\
&= \log\left(\frac{1}{c}\right) + \sqrt{2\left(\log\left(\frac{1}{c}\right) - 1\right)} \\
&\leq 3 \log \frac{1}{c}
\end{aligned}$$

Thus by our assumption,

$$\begin{aligned}
&x \geq 3 \cdot \frac{1}{c} \log\left(\frac{1}{c}\right) \\
\Rightarrow &x \geq \frac{1}{c}(-W_{-1}(-c))
\end{aligned}$$

then $W_{-1}(-c)$ is defined, so

$$\begin{aligned}
&x \geq \frac{1}{c} \max\left\{-W_{-1}(-c), 1\right\} \\
\Rightarrow &(-cx)e^{-cx} \geq -c \\
\Rightarrow &xe^{-cx} \leq 1 \\
\Rightarrow &\log(x) \leq cx
\end{aligned}$$

The first implication is justified as follows: $W_{-1}^{-1} : [-\frac{1}{\epsilon}, \infty) \to (-\infty, -1)$ is monotonically decreasing. Thus its inverse $W_{-1}^{-1}(y) = ye^y$, defined over the domain $(-\infty, -1)$ is also monotonically decreasing. By our assumption, $-cx \leq -3 \log \frac{1}{c} \leq -3$, thus $-cx \in (-\infty, -1]$, thus applying W_{-1}^{-1} to both sides gives us the first implication. ■

G. Experiment Details

In this section, we provide additional details of our experiments. In particular, we explain the CNN architecture that we use in our experiments. Denote a convolutional layer with p input filters and q output filters by conv(p, q), a fully connected layer with q outputs by fully_connect(q), and a max pooling operation with stride 2 as pool2. Let ReLU$(x) = \max\{x, 0\}$. Then the CNN architecture in our paper is the following:

conv$(3, 32)$ ⇒ ReLU ⇒ conv$(32, 64)$ ⇒ ReLU ⇒ pool2 ⇒ conv$(64, 128)$ ⇒ ReLU ⇒ conv$(128, 128)$
⇒ ReLU ⇒ pool2 ⇒ conv$(128, 256)$ ⇒ ReLU ⇒ conv$(256, 256)$ ⇒ ReLU ⇒ pool2 ⇒ fully_connect(1024)
⇒ ReLU ⇒ fully_connect(512) ⇒ ReLU ⇒ fully_connect(10).