Quantitative Central Limit Theorems for Discrete Stochastic Processes

Xiang Cheng †‡ Peter L. Bartlett †‡ Michael I. Jordan †‡

February 5, 2019

Abstract

In this paper, we establish a generalization of the classical Central Limit Theorem for a family of stochastic processes that includes stochastic gradient descent and related gradient-based algorithms. Under certain regularity assumptions, we show that the iterates of these stochastic processes converge to an invariant distribution at a rate of $O\left(\frac{1}{\sqrt{k}}\right)$ where $k$ is the number of steps; this rate is provably tight.

1 Introduction

Many randomized algorithms in machine learning can be analyzed as some kind of stochastic process. For example, MCMC algorithms intentionally inject carefully designed randomness in order to sample from a desired target distribution. There is a second category of randomized algorithms for which the goal is optimization rather than sampling, and the randomness is viewed as a price to pay for computational tractability. For example, stochastic gradient methods for large scale optimization use noisy estimates of a gradient because they are cheap.

While such algorithms are not designed with the goal of sampling from a target distribution, an algorithm of this kind has random outputs, and its behavior is determined by the distribution of its output. Results in this paper provide tools for analyzing the convergence of such algorithms as stochastic processes.

We establish a quantitative Central Limit Theorem for stochastic processes that have the following form:

$$x_{k+1} = x_k - \delta \nabla U(x_k) + \sqrt{\delta} \xi_k(x_k),$$  

where $x_k \in \mathbb{R}^d$ is an iterate, $\delta$ is a stepsize, $U : \mathbb{R}^d \to \mathbb{R}$ is a potential function, and $\xi(\cdot)$ is a zero-mean, position-dependent noise variable. Under certain assumptions, we show that $\{x_k\}$ converges in 2-Wasserstein distance to the following SDE:

$$dx(t) = -\nabla U(x(t))dt + \sigma(x(t))dB_t,$$

where $\sigma(x) = \left(\mathbb{E}\left[\xi(x)\xi(x)^T\right]\right)^{1/2}$. The notion of convergence is summarized in the following informal statement of our main theorem:

**Theorem 1 (Informal)** Let $p_k$ denote the distribution of $x_k$ in $\{x_k\}$, and let $p^*$ denote the invariant distribution of $\{x_k\}$. Then there exist constants $c_1, c_2$, such that for all $\epsilon > 0$, if $\delta \leq c_1 \epsilon^2 / d^2$ and $k \geq c_2 d^7 / \epsilon^2$,

$$W_2(p_k, p^*) \leq \epsilon.$$

In other words, under the right scaling of the step size, the long-term distribution of $x_k$ depends only on the expected drift $\nabla U(x)$ and the covariance matrix of the noise $\sigma(x)$. As long as we know these two quantities, we can draw conclusions about the approximate behavior of $\{x_k\}$ through $p^*$, and ignore the other characteristics of $\xi$.

Our result can be viewed as a general, quantitative form of the classical Central Limit Theorem, which can be thought of as showing that $x_k$ in $\{x_k\}$ converges in distribution to $N(0, I)$, for the specific case of $U(x) = \|x\|^2 / 2$ and $\sigma_x = I$. Our result is more general: $U(x)$ can be any strongly convex function satisfying certain regularity assumptions and $\sigma_x$ can vary with position. We show that $x_k$ converges to the invariant distribution of $\{x_k\}$, which is not necessarily a normal distribution. The fact that the classical CLT is a special case implies that the $\epsilon^{-2}$ rate in our main theorem cannot be improved in general. We discuss this in more detail in Section 3.1.1.

2 Related Work

Most relevant to this work is the quantitative CLT result due to Zhai (2018). In that paper, he established that for iid random variables $x_k$ with mean zero and covariance $I$, $W_2\left(\frac{X_k}{\sqrt{k}}, Z\right) = \tilde{O}\left(\frac{1}{\sqrt{k}}\right)$, where $Z$ is the standard
Gaussian random variable. Prior to this, a number of other authors have also proved a $1/\sqrt{T}$ rate, but without establishing dimension dependence (see, e.g., Bonis, 2015 [Rio et al., 2009].

Another relevant line of work is the recent work on quantitative rates for Langevin MCMC algorithms. Langevin MCMC algorithms can be thought of as discretizations of the Langevin diffusion SDE, which is essentially 2 for $\sigma(x) = I$. Authors such as Dalalyan (2017) and Durmus and Moulines (2016) were able to prove quantitative convergence results for Langevin MCMC by bounding its discretization error from the Langevin SDE. The processes we study in this paper differ from Langevin MCMC in two crucial ways: first, the noise $T_\eta(x)$ is not Gaussian, and second, the diffusion matrix in 3 varies with $x$.

Finally, this work is also motivated by results such as those due to Ruppert (1988), Polyak and Juditsky (1992), Fan et al. (2018), which show that iterates of the stochastic gradient algorithm with diminishing step size converge asymptotically to a normal distribution. (The limiting distribution of the appropriately rescaled iterates is Gaussian in this case, because a smooth $U$ is locally quadratic.) These classical results are asymptotic and do not give explicit rates.

3 Definitions and Assumptions

We will study the discrete process given by

$$x_{k+1} = x_k - \delta \nabla U(x_k) + \sqrt{2\delta T_{\eta_k}(x_k)},$$

where

1. $U(x): \mathbb{R}^d \rightarrow \mathbb{R}$ is the potential function,
2. $\eta_1, \eta_2, \ldots, \eta_k$ are iid random variables which take values in some set $\Omega$ and have distribution $q(\eta),$
3. $T: \Omega \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the noise map, and
4. $\delta > 0$ is a stepsize.

Let $\hat{\rho}(x)$ denote the invariant distribution of 3. Define

$$\sigma_x := \left( \mathbb{E}_{q(\eta)} \left[ T_\eta(x) T_\eta(x)^T \right] \right)^{1/2}.$$  (4)

We will also study the continuous SDE given by

$$dx(t) = -\nabla U(x(t)) dt + \sqrt{2\sigma_x(x)} dB_t,$$  (5)

where $B_t$ denotes the standard $d$-dimensional Brownian motion, and $\sigma_x: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is as defined in 4. Let $\rho^*$ denote the invariant distribution of 4.

For convenience of notation, we define the following:
1. Let $p_k$ be the distribution of $x_k$ in 3.
2. Let $F: \Omega \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ be the transition map:

$$F_\eta(x) := x - \delta \nabla U(x) + \sqrt{2\delta T_\eta(x)},$$  (6)

so that $x_{k+1} = F_\eta(x_k)$. Note that $F_\eta(x)$ also depends on $\delta$, but we do not write this explicitly; the choice of $\delta$ should be clear from context.
3. Define $\Phi_\delta$ as

$$\Phi_\delta(p) := (F_\eta)_# p,$$  (7)

where $\#$ denotes the pushforward operator; i.e., $\Phi(p)$ is the distribution of $F_\eta(x)$ when $x \sim p$, so that $p_{k+1} = \Phi_\delta(p_k)$.

We make the following assumptions about $U$.

Assumption 1 There exist constants $m$ and $L$ satisfying, for all $x$,

1. $\nabla U(0) = 0$,
2. $\nabla^2 U(x) \preceq LI$,
3. $\nabla^2 U(x) \succeq mI$,
4. $\|\nabla^3 U(x)\|_2 \leq L$.

where $\| \cdot \|_2$ denotes the operator norm; see 5 below.

We make the following assumptions about $q(\eta)$ and $T_\eta(x)$:

Assumption 2 There exists a constant $c_\eta$, such that for all $x$,

1. $\mathbb{E}_{q(\eta)}[T_\eta(x)] = 0$,
2. $\mathbb{E}_{q(\eta)} \left[ T_\eta(x) T_\eta(x)^T \right] \preceq c_\eta^2 I$. 


3.1 Basic Notation

For any two distributions \( p \) and \( q \), let \( W_2(p, q) \) be the 2-Wasserstein distance between \( p \) and \( q \). We overload the notation and sometimes use \( W_2(x, y) \) for random variables \( x \) and \( y \) to denote the \( W_2 \) distance between their distributions.

For a \( k \)th-order tensor \( M \in \mathbb{R}^{k} \) and a vector \( v \in \mathbb{R}^d \), we define the product \( A = Mv \) such that \( [A]_{i_1...i_{k-1}j} = \sum_{i_{k-1}}[M]_{i_1...i_{k-1}i_{k-1,j}}v_j \). Sometimes, to avoid ambiguity, we will write \( A = \langle Mv \rangle \) instead.

We let \( \| \cdot \|_2 \) denote the operator norm:

\[
\|M\|_2 = \sup_{v \in \mathbb{R}^d, \|v\|_2 = 1} \|Mv\|_2.
\]

It can be verified that for all \( k \), \( \| \cdot \|_2 \) is a norm over \( \mathbb{R}^{kd} \).

Finally, we use the notation \( \{ \} \) to denote two kinds of inner products:

1. For vectors \( u, v \in \mathbb{R}^d \), \( \langle u, v \rangle = \sum_{i=1}^d u_i v_i \) (the dot product).
2. For matrices \( A, B \in \mathbb{R}^{k \times d} \), \( \langle A, B \rangle := \sum_{i=1}^k \sum_{j=1}^d A_{i,j} B_{j,i} \) (the trace inner product).

Although the notation is overloaded, the usage should be clear from context.

4 Main Results and Discussion

We will consider two settings: one in which the noise \( T_\eta \) in \( \mathcal{X} \) does not depend on \( x \), and one in which it does. We will treat these results separately in Theorem 2 and Theorem 3.

4.1 Homogeneous Noise

Our first theorem deals with the case when \( T_\eta \) is a constant with respect to \( x \). In addition to Assumption 1 and Assumption 2, we make the following assumptions:

**Assumption 3** For all \( x \),

1. \( T_\eta(x) = T_\eta \),
2. \( \|T_\eta\|_2 \leq \sqrt{T} \),
3. \( \delta_x = I \).

Under these assumptions, the invariant distribution \( p^*(x) \) of \( \mathcal{X} \) has the form

\[
p^*(x) \propto e^{-U(x)}.
\]

**Theorem 2** Let \( p_\eta \) be an arbitrary initial distribution, and let \( p_k \) be defined as in \( \mathcal{X} \) with step size \( \delta \). Recall the definition of \( \tilde{p} \) as the invariant distribution of \( \mathcal{X} \) and \( p^* \) as the invariant distribution of \( \mathcal{X} \).

For \( \delta \leq \frac{1}{\sqrt{d}} \cdot \text{poly} \left( \frac{1}{\delta}, L \right)^{-1} \),

\[
W_2(\tilde{p}, p^*) \leq \epsilon.
\]

If, in addition, \( k \geq \frac{\delta}{\sqrt{d}} \log \frac{W_2(p_\eta, p^*)}{\epsilon} \cdot \text{poly} \left( \frac{1}{\delta}, L \right) \),

\[
W_2(p_k, p^*) \leq \epsilon.
\]

An equivalent statement is that for a sufficiently large \( k \), and for sufficiently small \( \delta \), we can bound

\[
W_2(p_k, p^*) \leq \tilde{O} \left( \frac{\delta^{3/2}}{\sqrt{d}} \right).
\]

4.1.1 Relation to the Classical Central Limit Theorem

Our result can be viewed as a generalization of the classical central limit theorem, which deals with sequences of the form

\[
S_{k+1} = \frac{\sum_{i=0}^{k+1} \eta_i}{\sqrt{k+1}} = \frac{\sqrt{k}}{\sqrt{k+1}} \cdot S_k + \frac{\eta_{k+1}}{\sqrt{k+1}} \approx S_k - \frac{1}{2(k+1)} S_k + \frac{\sqrt{2}}{2(k+1)^{3/2}} \eta_{k+1} \]

for some \( \eta_k \) with mean 0 and covariance \( I \). Thus, the sequence \( S_k \) essentially has the same dynamics as \( x_k \) from \( \mathcal{X} \), with \( U(x) = -\frac{1}{2} \|x\|_2^2 \), \( T_{\eta_k} = \eta_k \) and variable stepsize \( \delta_k = \frac{1}{\sqrt{k}} \). To the best of our knowledge, the best rate for the classical CLT is proven in Theorem 1.1 of Zhai (2018), with a rate of \( W_2(S_k, Z) \leq \tilde{O} \left( \sqrt{d/k} \right) \). It is also essentially tight, as Proposition 1.2 of Zhai (2018) shows that the \( W_2(S_k, Z) \) is lower bounded by \( \Omega \left( \sqrt{d/k} \right) \).

Our bound in Theorem 2 (equivalently, (12)) also shrinks as \( 1/\sqrt{k} \). We note that the sequence \( x_k \) studied in Theorem 2 differs from \( S_k \), as the stepsize for \( x_k \) is constant (i.e., \( \delta \) does not depend on \( k \)). We stated Theorem 2 for constant step sizes mainly to simplify the proof. Our proof technique can also be applied to the variable step size setting; in Corollary 3 in the appendix, we use the results of Theorem 3 to prove that \( W_2(S_k, Z) \leq \tilde{O} \left( \sqrt{d/k} \right) \), which is the same as the constant-stepsize case.

This shows that the \( k \) dependence in Theorem 2 is tight. Our \( d \) dependence is \( d^{3/2} \), compared to the optimal rate of \( \sqrt{d} \). However, our bound is applicable to a much more general setting, not just for \( U(x) = 1/2 \|x\|_2^2 \).
4.2 Inhomogeneous Noise

We now examine the convergence of \((\ref{eq:2})\) under a general setting, in which the noise \(T_n(x)\) depends on the position.

In addition to the assumptions in Section \((\ref{eq:3})\) we make some additional assumptions about how \(T_n(x)\) depends on \(x\). We begin by defining some notation. For all \(x \in \mathbb{R}^d\) and \(\eta \in \Omega\), we will let \(G_n(x) \in \mathbb{R}^{2d}\) denote the derivative of \(T_n(x)\) wrt \(x\), \(M_n(x) \in \mathbb{R}^{4d}\) denote the derivative of \(G_n(x)\) wrt \(x\), and \(N_n(x) \in \mathbb{R}^{4d}\) denote the derivative of \(M_n(x)\) wrt \(x\), i.e.:

\[
\begin{align*}
1. & \ \forall x, i, j \text{ and for } \eta \text{ a.s., } [G_n(x)]_{i,j} := \frac{\partial}{\partial x_j} [T_n(x)]_i \\
2. & \ \forall x, i, j, k \text{ and for } \eta \text{ a.s., } [M_n(x)]_{i,j,k} := \frac{\partial^2}{\partial x_j \partial x_k} [T_n(x)]_i \\
3. & \ \forall x, i, j, k, l \text{ and for } \eta \text{ a.s., } [N_n(x)]_{i,j,k,l} := \frac{\partial^3}{\partial x_j \partial x_k \partial x_l} [T_n(x)]_i
\end{align*}
\]

We will assume that \(T_n(x), G_n(x), M_n(x)\) satisfy the following regularity:

**Assumption 4** There exists an \(L\) that satisfies Assumption \((\ref{eq:4})\) and, for all \(x\) and for \(\eta\) a.s.:

\[
\begin{align*}
1. & \ G_n(x) \text{ is symmetric,} \\
2. & \ \|T_n(x)\|_2 \leq \sqrt{L}(\|x\|_2 + 1), \\
3. & \ \|G_n(x)\|_2 \leq \sqrt{L}, \\
4. & \ \|M_n(x)\|_2 \leq \sqrt{L}, \\
5. & \ \|N_n(x)\|_2 \leq \sqrt{L}.
\end{align*}
\]

**Assumption 5** For any distributions \(p\) and \(q\), \(W_2(\Phi_4(p), \Phi_4(q)) \leq e^{-\lambda}W_2(p,q)\).

Finally, we assume that \(\log p^*(x)\) is regular in the following sense:

**Assumption 6** There exists a constant \(\theta\), such that the log of the invariant distribution of \((\ref{eq:3})\), \(f(x) := \log(p^*(x))\), satisfies, for all \(x\),

\[
\begin{align*}
1. & \ \|\nabla^3 f(x)\|_2 \leq \theta, \\
2. & \ \|\nabla^2 f(x)\|_2 \leq \theta (\|x\|_2 + 1), \\
3. & \ \|\nabla f(x)\|_2 \leq \theta (\|x\|_2^2 + 1).
\end{align*}
\]

**Remark 1** If \(\nabla^2 f(0)\) and \(\nabla f(0)\) are bounded by \(\theta\), then 2. and 3. are implied by 1., but we state the assumption this way for convenience.

4.2.1 A motivating example

Before we state our main theorem, it will help to motivate some of our assumptions by considering an application to the stochastic gradient algorithm.

Consider a classification problem where one tries to learn the parameters \(x\) of a model. One is given \(S\) datapoints \((z_1, y_1), \ldots, (z_s, y_s)\), and a likelihood function \(\ell(x, (z_i, y_i))\), and one tries to minimize \(U(x)\) for

\[
U(x) := \frac{1}{S} \sum_{i=1}^{S} U_i(x), \quad \text{with} \quad U_i(x) := \ell(x, (z_i, y_i)).
\]

The stochastic gradient algorithm proceeds as follows:

\[
x_{k+1} = x_k - \delta \nabla U_{\eta_k}(x_k) \\
\phantom{x_{k+1}} = x_k - \delta \nabla U(x_k) + \sqrt{\delta} T_{\eta_k}(x_k),
\]

where for each \(k\), \(\eta_k\) is an integer sampled uniformly from \(\{1, \ldots, S\}\), and we define \(T_{\eta_k}(x) := \sqrt{\delta/2} (\nabla U(x) - \nabla U_{\eta_k}(x))\).

Notice that \((\ref{eq:3})\) is identical to \((\ref{eq:4})\).

The mean and variance of \(T_{\eta}\) are

\[
\begin{align*}
\mathbb{E}_\eta [T_{\eta}(x)] &= 0 \\
\mathbb{E}_\eta [T_{\eta}(x) T_{\eta}(x)^T] &= \frac{\sqrt{\delta/2}}{S} \mathbb{E}_{\eta \sim \text{Unif}\{1, \ldots, S\}} \left[ (\nabla U(x) - \nabla U_i(x)) (\nabla U(x) - \nabla U_i(x))^T \right]
\end{align*}
\]

Assuming \(\delta \leq 1\), Assumption \((\ref{eq:2})\) is true if for some constant \(c_\varnothing\),

\[
\mathbb{E}_{\eta \sim \text{Unif}\{1, \ldots, S\}} \left[ (\nabla U(x) - \nabla U_i(x)) (\nabla U(x) - \nabla U_i(x))^T \right] \leq c_\varnothing^2 \cdot I
\]

(\ref{eq:3})

Furthermore, \(T_{\eta}(x), G_{\eta}(x), M_{\eta}(x), N_{\eta}(x)\) are respectively \(\sqrt{\delta/2} \nabla (U(x) - U_{\eta_k}(x)) + \eta_k^2\), \(\sqrt{\delta/2} \Sigma (U(x) - U_{\eta_k}(x))\), \(\sqrt{\delta/2} \Sigma^2 (U(x) - U_{\eta_k}(x))\), \(\sqrt{\delta/2} \Sigma^3 (U(x) - U_{\eta_k}(x))\), so Assumption \((\ref{eq:4})\) is guaranteed by the loss function \(\ell\) having Lipschitz derivatives (in \(x\)) up to fourth order.

If \(\nabla U_i(x)\) is \(m\)-strongly convex and has \(L\)-Lipschitz gradients for all \(i\), then Assumption \((\ref{eq:6})\) is true for \(\lambda = m\) for all \(\delta \leq 1/(2L)\), by a synchronous coupling argument (see Lemma \((\ref{eq:3})\) in Appendix \((\ref{eq:3})\).

Finally, we remark that the upper bound for Assumption \((\ref{eq:2})\) implied by \((\ref{eq:3})\) is in fact quite loose when \(\delta \ll 1\).

We will now state our main theorem for this section:
Theorem 3 Let \( p_0 \) be an arbitrary initial distribution, and let \( p_k \) be defined as in \( 3 \) with step size \( \delta \). Recall the definition of \( \bar{p} \) as the invariant distribution of \( 3 \) and \( p^* \) as the invariant distribution of \( 5 \). For \( \delta \leq \frac{e^2}{d^2} \cdot \text{poly} \left( \frac{1}{m}, L, \theta \right)^{-1} \),

\[
W_2(\bar{p}, p^*) \leq \epsilon.
\] (15)

If, in addition, \( k \geq \frac{d^2}{\epsilon^2} \log \frac{W_2(p_0, p^*)}{\epsilon} \cdot \text{poly} \left( L, \theta, \frac{1}{m}, c, \theta \right) \), then

\[
W_2(p_k, p^*) \leq \epsilon.
\] (16)

Remark 2 Like Theorem 3, this also gives a \( 1/\sqrt{k} \) rate, which is optimal. (see Section 4.1.1).

5 Proof of Main Theorems

In this section, we sketch the proofs of Theorems 2 and 3.

5.1 Proof of Results for Homogeneous Diffusion

Proof of Theorem 2 We first prove \((11)\).

By Theorem \( 4 \) below, for \( \delta \leq \frac{\min\{m^2, 1\}}{2^{\log \left( \frac{1}{m}, L, \theta \right)}}, \)

\[
W_2(p_k, p^*) \leq e^{-m\delta k/8}W_2(p_0, p^*) + 2^d \delta^{1/2}d^{3/2} (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^7.
\] (17)

Thus if \( \delta \leq \frac{e^2}{d^2} \cdot \left( \frac{1}{2^{166}d^3 (L + 1)^9} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^{14} \right)^{-1} \), then

\[
2^d \delta^{1/2}d^{3/2} (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^7 \leq \frac{\epsilon}{2}.
\]

Additionally, if \( k \geq \frac{8}{m \delta} \log \frac{W_2(p_0, p^*)}{\epsilon} = \frac{d^3}{\epsilon^2} \cdot \log \frac{W_2(p_0, p^*)}{\epsilon} \cdot \text{poly} \left( \frac{1}{m}, L \right) \).

This proves \((11)\). To prove \((10)\), use our above assumption on \( \delta \), and take the limit of \((17)\) as \( k \to \infty \).

Theorem 4 Let \( p_0 \) be an arbitrary initial distribution, and let \( p_k \) be defined as in \( 5 \).

Let \( \epsilon > 0 \) be some arbitrary constant. For any step size \( \delta \) satisfying \( \delta \leq \frac{\min\{m^2, 1\}}{2^{\log \left( \frac{1}{m}, L, \theta \right)}}, \) the Wasserstein distance between \( p_k \) and \( p^* \) is upper bounded as

\[
W_2(p_k, p^*) \leq e^{-m\delta k/8}W_2(p_0, p^*) + 2^d \delta^{1/2}d^{3/2} (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^7.
\]

Proof of Theorem 4 Recall our definition of \( \Phi_k \) in \( 7 \). Let \( \Phi_k^k \) denote \( k \) repeated applications of \( \Phi_k \), so \( p_k = \Phi_k^k(p_0) \). Our objective is thus to bound \( W_2(\Phi_k^k(p_0), p^*) \).

We first use triangle inequality to split the objective into two terms:

\[
W_2(\Phi_k^k(p_0), p^*) \leq W_2(\Phi_k^k(p_0), \Phi_k^k(p^*)) + W_2(\Phi_k^k(p^*), p^*)
\] (18)

The first term is easy to bound. We can apply Lemma 14 (in Appendix A) to get

\[
W_2(\Phi_k^k(p^*), p^*) \leq e^{-m\delta k/8}W_2(p_0, p^*)
\] (19)

To bound the second term of \( (18) \), we use an argument adapted from (Zhai 2016):

\[
W_2(\Phi_k^k(p^*), p^*) = W_2(\Phi_k(\Phi_k^{k-1}(p^*)), p^*)
\]
\[
\leq W_2(\Phi_k(\Phi_k^{k-1}(p^*)), \Phi_k(p^*)) + W_2(\Phi_k(p^*), p^*)
\]
\[
\leq e^{-m\delta/8}W_2(\Phi_k^{k-1}(p^*), p^*) + W_2(\Phi_k(p^*), p^*)
\]

\[
\vdots
\]
\[
\leq \sum_{i=0}^{k-1} e^{-m\delta/8}W_2(\Phi_k(p^*), p^*)
\]
\[
\leq \frac{8}{m \delta}W_2(\Phi_k(p^*), p^*).
\]
Here the third inequality is by induction. This reduces our problem to bounding the expression $W_2(\Psi_3(p^*), p^*)$, which can be thought of as the one-step divergence between (3) and (5) when $p_0 = p^*$. We apply Lemma 1 below to get
\[ W_2(\Psi_3(p^*), p^*) \leq 2^{78} \delta^{3/2} \log (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^6. \]

Thus, substituting (19) and (20) into (18), we get
\[ W_2(\Phi_3(p_0), p^*) \leq e^{-m \delta k/8} W_2(p_0, p^*) + 2^{82} \delta^{1/2} \log (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^7. \] (21)

**Lemma 1** Let $p_0 := \Phi_3(p^*)$. Then for any $\delta \leq \min \left\{ \frac{n^2}{2}, \frac{1}{2} \right\}$,
\[ W_2(p_0, p^*) \leq 2^{78} \delta^{3/2} \log (L + 1)^{9/2} \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^6. \]

(This lemma is similar in spirit to Lemma 1.6 in Zhai (2018).)

**Proof of Lemma 1**
Using Talagrand’s inequality and the fact that $U(x)$ is strongly convex, we can upper bound $W_2^2(q, p^*)$ by $\chi^2(q, p^*)$ for any distribution $q$ which has density wrt $p^*$, i.e.:
\[ W_2^2(q, p^*) \leq \frac{2}{m} \int_0^1 \left( \frac{p_0(x)}{p^*(x)} - 1 \right)^2 p^*(x) \, dx. \] (22)

See Lemma 12 in Appendix B for a rigorous proof of (22).

Under our assumptions on $\delta$, we can apply Lemma 2 below, giving
\[ \int_{B_R} \left( \frac{p_0(x)}{p^*(x)} - 1 \right)^2 p^*(x) \, dx 
\leq 2^{23} \delta^3 d^2 (L + 1)^9 \int \exp \left( \frac{m}{16} \|x\|_2^2 \right) \left( \|x\|_2^2 + 1 \right) p^*(x) \, dx 
\leq 2^{24} \delta^3 d^2 (L + 1)^9 \left( \int \exp \left( \frac{m}{16} \|x\|_2^2 \right) p^*(x) \, dx + \int \left( \|x\|_2^4 + 1 \right) p^*(x) \, dx \right) 
\leq 2^{24} \delta^3 d^2 (L + 1)^9 \left( 8d + \max \left\{ \left( \frac{m}{16} \log \left( \frac{2^8}{m} \right) \right)^{11}, \left( \frac{1}{m} \log \left( \frac{1}{m} \right) \right)^{11} \right\} \right) 
\leq 2^{150} \delta^3 d^3 (L + 1)^9 \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^{11}, \]

where the first inequality is by Lemma 2, the second inequality is by Young’s inequality, the third inequality is by Lemma 34 and Lemma 38, with $c_v = 1$. Plugging the above into (22),
\[ W_2^2(p_0, p^*) \leq 2^{156} \delta^3 d^3 (L + 1)^9 \max \left\{ \frac{1}{m} \log \left( \frac{1}{m} \right), 1 \right\}^{12}. \] (23)

The following lemma studies the “discretization error” between the SDE (3) and one step of (5).

**Lemma 2** Let $p_3 := \Phi_3(p^*)$. For any $R \geq 0, x \in B_R$, and $\delta \leq \min \left\{ \frac{n^2}{2}, \frac{1}{2} \right\}$,
\[ \left| \frac{p_3(x)}{p^*(x)} - 1 \right| \leq 512 \delta^{3/2} d (L + 1)^{9/2} \exp \left( \frac{m}{32} \|x\|_2^2 \right) \left( \|x\|_2^6 + 1 \right). \]

**Proof of Lemma 2** Recall that $p_3 = (F_n)_\# p^*(x)$. Thus by the change of variable formula, we have
\[ p_3(x) = \int p^*(F_n^{-1}(x)) \det (\nabla F_n (F_n^{-1}(x)))^{-1} q(y) \, dy 
= \mathbb{E}_{q(y)} \left[ p^*(F_n^{-1}(x)) \det (\nabla F_n (F_n^{-1}(x)))^{-1} \right], \] (23)
where $\nabla F_0(y)$ denotes the Jacobian matrix of $F_0$ at $y$. The invertibility of $F_0$ is shown in Lemma 40. We rewrite (1) as its Taylor expansion about $x$:

$$p^* (F_0^{-1}(x)) = p^*(x) + \langle \nabla p^*(x), F_0^{-1}(x) - x \rangle + \frac{1}{2} \left\langle \nabla^2 p^*(x), (F_0^{-1}(x) - x) (F_0^{-1}(x) - x)^T \right\rangle$$

$$+ \int_0^1 \int_0^r \int_0^s \left\langle \nabla^3 p^* \left( ((1 - r)x + rF_0^{-1}(x)) , (F_0^{-1}(x) - x)^3 \right) drdsdt \right.$$

Substituting the above into (23) and applying Lemmas 3, 4, 5, and 6, we get

$$p_k(x) = E_{\eta(x)} \left[ (\delta + \beta + \gamma + \varphi) \cdot \omega \right]$$

$$= p^*(x) + p^*(x) \left( \delta \Tr (\nabla^2 U(x)) + \beta \langle \nabla p^*(x), \nabla U(x) \rangle + \gamma \Tr (\nabla^2 p^*(x)) + \varphi \right)$$

for some $\Delta$ satisfying

$$|\Delta| \leq p^*(x) \cdot 85^3/2dL^{3/2} (\|x\|_2 + 1) + p^*(x) \cdot 165^3/2dL^{3/2} (\|x\|_2 + 1)$$

$$+ p^*(x) \cdot 64d^{3/2} (L + 1)^{3/2} (\|x\|_2 + 1)$$

$$+ p^*(x) \cdot 256d^{3/2} (L + 1)^{3/2} (\|x\|_2 + 1)$$

$$\leq p^*(x) \cdot 512d^{3/2} (L + 1)^{3/2} (\|x\|_2 + 1).$$

Furthermore, by using the expression $p^*(x) \propto e^{-U(x)}$ and some algebra, we see that

$$p^*(x) \left( \delta \Tr (\nabla^2 U(x)) + \beta \langle \nabla p^*(x), \nabla U(x) \rangle + \gamma \Tr (\nabla^2 p^*(x)) \right)$$

$$= \delta p^*(x) \left( \Tr (\nabla^2 U(x)) - \|\nabla U(x)\|_2^2 - \Tr (\nabla^2 U(x)) + \Tr (\nabla U(x) \nabla U(x)^T) \right)$$

$$= 0.$$

Substituting the above into (23) gives $p_k(x) = p^*(x) + \Delta$, which implies that

$$\left| \frac{p_k(x)}{p^*(x)} - 1 \right| \leq 512d^{3/2} (L + 1)^{3/2} (\|x\|_2^2 + 1).$$

\[\square\]

5.2 Proof of Results for Inhomogeneous Diffusion

The proof of Theorem 3 is quite similar to the proof of Theorem 2 and can be found in the Appendix (Section B). We will highlight some additional difficulties in the proof compared to Theorem 2.

The heart of the proof lies in Lemma 13, which bounds the discretization error between the SDE (43) and the Fokker-Planck equation (see (43)). This allows us to, somewhat remarkably, prove that $p^*$ is the invariant distribution of (3).

Substituting the above into (24) gives $p_k(x) = p^*(x) + \Delta$, which implies that

$$\left| \frac{p_k(x)}{p^*(x)} - 1 \right| \leq 512d^{3/2} (L + 1)^{3/2} (\|x\|_2^2 + 1).$$

5 Conclusion and Future Directions

The main result of this paper is a generalization of the classical Central Limit Theorem to discrete-time stochastic processes of the form (43). Our results assume that $U(x)$ is strongly convex (Assumption 13). This is not strictly necessary. We use strong convexity in two ways:

1. We use it for proving contraction of (43), as in Lemma 14 and Lemma 33. Assuming that the noise $T_n$ contains an independent symmetric component (e.g., Gaussian noise), and assuming that $U(x)$ is nonconvex inside but strongly convex outside a ball, then we can use a reflection coupling argument to show that Assumption 5 holds.
2. We use it for proving that \( p^* \) is subgaussian, as in Lemma \( \text{[35]} \). For this lemma, it suffices that \( U(x) \) is \( m \)-dissipative.

Another assumption that can be relaxed is Assumption \( \text{[2.2]} \), which is used to show that \( p^* \) is subgaussian. We can replace this assumption by the weaker condition

\[
E_{q(x)} [T_0(x) T_0(x)^T] \prec c_2 \sigma_2 \|x\|_2^2 I.
\]

We only need to make an additional assumption that \( U(x) \) is \( M \)-dissipative for some radius \( D \), with \( M \geq 8c_2 \). We do not prove this here to keep the proofs simple; a proof will be included in the full version of this paper.

Finally, we remark that \( \text{[5]} \) suggests that \( x_t \) moves quickly through regions of large \( \sigma_x \). This seems to suggest that in the stochastic gradient algorithm, the iterates will, with higher probability, end up in minima where the covariance of the gradient is small. This may in turn suggest that the noise SGD tends to select “stable” solutions, where stability is defined as the determinant of the covariance of the gradient. This property would not be present with a different noise such as Gaussian noise in Langevin diffusion. A rigorous investigation of this possibility is beyond the scope of this paper.
References

Thomas Bonis. Rates in the central limit theorem and diffusion approximation via Stein’s method. arXiv preprint arXiv:1506.06966, 2015.

Ioannis Chatzigeorgiou. Bounds on the Lambert function and their application to the outage analysis of user cooperation. IEEE Communications Letters, 17(8):1505–1508, 2013.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv preprint arXiv:1805.01648, 2018.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):651–676, 2017.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. arXiv preprint arXiv:1605.01559, 2016.

Jianqing Fan, Wenyuan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In International Conference on Artificial Intelligence and Statistics, pages 1017–1026, 2018.

Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In International Conference on Machine Learning, pages 354–363, 2016.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate Bayesian inference. The Journal of Machine Learning Research, 18(1):4873–4907, 2017.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. Journal of Functional Analysis, 173(2):361–400, 2000.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization, 30(4):838–855, 1992.

Emmanuel Rio et al. Upper bounds for minimal distances in the central limit theorem. In Annales de l’Institut Henri Poincaré, Probabilités et Statistiques, volume 45, pages 802–817. Institut Henri Poincaré, 2009.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Alex Zhai. A high-dimensional CLT in $W_2$ distance with near optimal convergence rate. Probability Theory and Related Fields, 170(3-4):821–845, 2018.

9
A Auxiliary Lemmas for Section 4.1

In this subsection, we present the proof of Lemma 1 as well as some auxiliary lemmas.

**Lemma 3** For \( \delta \leq \frac{1}{2\delta^2L} \),
\[
E_{q(n)} \left[ p^*(x) \cdot \det (\nabla F_{\eta} (F_{\eta}^{-1}(x))) \right] = p^*(x) + p^*(x) (\delta \text{tr} (\nabla^2 U(x))) + \Delta,
\]
for some \( |\Delta| \leq p^*(x) \cdot 8\delta^{3/2}L^{3/2} (\|x\|_2 + 1) \).

**Proof of Lemma 3** Let us define
\[
\Delta' := \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} - (1 + \delta \text{tr} (\nabla^2 U(x))).
\]
By Lemma 1 \( |\Delta'| \leq 8\delta^{3/2}L^{3/2} (\|x\|_2 + 1) \), so
\[
E_{q(n)} \left[ p^*(x) \cdot \det (\nabla F_{\eta} (F_{\eta}^{-1}(x))) \right] = E_{q(n)} \left[ p^*(x) \cdot (1 + \delta \text{tr} (\nabla^2 U(x))) \right] + E_{q(n)} [p^*(x) \cdot \Delta']
= p^*(x) (1 + \delta \text{tr} (\nabla^2 U(x))) + p^*(x) \cdot \Delta'.
\]
We complete the proof by taking \( \Delta := p^*(x)\Delta' \).

**Lemma 4** For \( \delta \leq \frac{1}{64\delta^2L} \),
\[
E_{q(n)} \left[ (\nabla p^*(x), F_{\eta}^{-1}(x) - x) \cdot \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} \right] = \delta \langle \nabla p^*(x), \nabla U(x) \rangle + \Delta
\]
for some \( |\Delta| \leq p^*(x) \cdot 16\delta^{3/2}L^{5/2} (\|x\|_2^3 + 1) \).

**Proof of Lemma 4** Let
\[
\Delta_1 := F_{\eta}^{-1}(x) - x - \left( -\sqrt{2\delta} T_{\eta} + \delta \nabla U(x) \right), \quad \Delta_2 := \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} - 1.
\]
By Lemma 2 and Corollary 11
\[
\|\Delta_1\|_2 \leq 4\delta^{3/2}L^{3/2} (\|x\|_2 + 1), \quad |\Delta_2| \leq 2\delta dL (\|x\|_2 + 1).
\]
Moving terms around,
\[
E_{q(n)} \left[ (\nabla p^*(x), F_{\eta}^{-1}(x) - x) \cdot \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} \right]
= E_{q(n)} \left[ (\nabla p^*(x), -\sqrt{2\delta} T_{\eta} + \delta \nabla U(x)) \right] + E_{q(n)} [\langle \nabla p^*(x), \delta \nabla U(x) \rangle]
+ E_{q(n)} \left[ (\nabla p^*(x), \sqrt{2\delta} T_{\eta} + \delta \nabla U(x)) \cdot \Delta_2 \right]
+ E_{q(n)} \left[ (\nabla p^*(x), \Delta_1) \cdot \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} \right].
\]
The main term of interest is (26), which evaluates to
\[
E_{q(n)} \left[ (\nabla p^*(x), -\sqrt{2\delta} T_{\eta}) \right] = E_{q(n)} \left[ (\nabla p^*(x), \delta \nabla U(x)) \right]
= \delta \langle \nabla p^*(x), \nabla U(x) \rangle,
\]
where the first equality is by Assumption 2.1.
We now consider the terms in (27) and (28):
\[
|27| = \left| E_{q(n)} \left[ (\nabla p^*(x), \sqrt{2\delta} T_{\eta} + \delta \nabla U(x)) \cdot \Delta_2 \right] \right|
\leq \|
\nabla p^*(x)\|_2 E_{q(n)} \left[ \left\| \sqrt{2\delta} T_{\eta} + \delta \nabla U(x) \right\|_2 \|\Delta_2\| \right]
\leq p^*(x) L\|x\|_2 \cdot \sqrt{\delta} \left( \sqrt{L} + \sqrt{\delta L}\|x\|_2 \right) \cdot 2\delta dL (\|x\|_2 + 1)
\leq 8p^*(x)\delta^{3/2}L^{5/2} (\|x\|_2^3 + 1),
\]
where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 13, our upperbound on \( |\Delta_2| \) at the start of the proof, and Assumptions 11.2 and 11.2.

\[
|28| = \left| E_{q(n)} \left[ (\nabla p^*(x), \Delta_1) \cdot \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} \right] \right|
\leq \|
\nabla p^*(x)\|_2 E_{q(n)} \left[ \|\Delta_1\|_2 \cdot \left| \text{det} (\nabla F_{\eta} (F_{\eta}^{-1}(x)))^{-1} \right| \right]
\leq p^*(x) L\|x\|_2 \cdot \delta^{3/2}L^{5/2} (\|x\|_2 + 1) \cdot (1 + 2\delta dL (\|x\|_2 + 1))
\leq 8p^*(x)\delta^{3/2}L^{5/2} (\|x\|_2^3 + 1),
\]
where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 31, our upperbound on $\|\Delta_1\|_2$ and $|\Delta_2|$ at the start of the proof, and our assumption on $\delta$.

Letting $\Delta := (27) + (28)$, we have
\[
|\Delta| \leq 8p^*(x)\delta^{3/2}dL^{5/2} (\|x\|_3^3 + 1) + 8p^*(x)\delta^{3/2}dL^{5/2} (\|x\|_3^3 + 1) \\
\leq p^*(x) \cdot 16\delta^{3/2}dL^{5/2} (\|x\|_3^3 + 1).
\]

Lemma 5

For $\delta \leq \frac{1}{64\delta^2(L+1)}$, \[
\frac{1}{2}E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), (F_\eta^{-1}(x) - x) (F_\eta^{-1}(x) - x)^T \right\rangle \cdot \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} \right] = \delta tr \left( \nabla^2 p^*(x) \right) + \Delta
\]

for some $|\Delta| \leq p^*(x) \cdot 64\delta^{3/2}d(L + 1)^{5/2} (\|x\|_2^5 + 1)$.

Proof of Lemma 5

Define
\[
\Delta_1 := F_\eta^{-1}(x) - x - \left( -\sqrt{2\delta} F_\eta \right), \quad \Delta_2 := \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} - 1.
\]

By Lemma 7 and Corollary 11
\[
|\Delta_1| \leq 2\delta L (\|x\|_2 + 1), \quad |\Delta_2| \leq 2\delta dL (\|x\|_2 + 1).
\]

Then
\[
E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), (F_\eta^{-1}(x) - x) (F_\eta^{-1}(x) - x)^T \right\rangle \cdot \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} \right] = 2\delta E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), T_\eta T_\eta^T \right\rangle \right]
\]
\[
= 2\delta E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), T_\eta T_\eta^T \right\rangle \cdot \Delta_2 \right] + E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), \Delta_1 \Delta_1^T - \sqrt{2\delta} T_\eta \Delta_1^T - \sqrt{2\delta} \Delta_1 T_\eta^T \right\rangle \cdot \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} \right].
\]

We are mainly interested in (29), which evaluates to
\[
2\delta E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), T_\eta T_\eta^T \right\rangle = 2\delta \left\langle \nabla^2 p^*(x), E_{\eta(q)} \left[ T_\eta T_\eta^T \right] \right\rangle = 2\delta tr (\nabla^2 p^*(x)), \right.
\]

where the last equality is by Assumption 41.

We now bound the magnitudes of (29) and (31).

\[
|29| = \left| 24E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), T_\eta T_\eta^T \right\rangle \cdot \Delta_2 \right] \right|
\]
\[
\leq 2\delta \left\| \nabla^2 p^*(x) \right\|_2 E_{\eta(q)} \left[ \left\| T_\eta \right\|_2 \left\| \Delta_2 \right\| \right]
\]
\[
\leq 26p^*(x) (L + L^2 (\|x\|_2^5) \cdot L \cdot 2\delta dL (\|x\|_2 + 1)
\]
\[
\leq 88p^*(x) \cdot d (L + 1)^2 (\|x\|_2^3 + 1),
\]

where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 32 and our upper bound on $|\Delta_2|$ at the start of the proof.

\[
|31| = E_{\eta(q)} \left[ \left\langle \nabla^2 p^*(x), \Delta_1 \Delta_1^T - \sqrt{2\delta} T_\eta \Delta_1^T - \sqrt{2\delta} \Delta_1 T_\eta^T \right\rangle \cdot \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} \right]
\]
\[
\leq \left\| \nabla^2 p^*(x) \right\|_2 E_{\eta(q)} \left[ \left( \left\| \Delta_1 \right\|_2 \right)^2 + 2\sqrt{2\delta} \left\| T_\eta \right\|_2 \left\| \Delta_1 \right\|_2 \right] \det (\nabla F_\eta(F_\eta^{-1}(x)))^{-1} \right]
\]
\[
\leq p^*(x) (L + L^2 (\|x\|_2^5) \cdot \left( (25L (\|x\|_2 + 1))^2 + 4\sqrt{5}L^{1/2} (25L (\|x\|_2 + 1))^2 \right) \\
\cdot (1 + 2\delta dL (\|x\|_2 + 1))
\]
\[
\leq 32\delta^{3/2}p^*(x) (L + 1)^{5/2} (\|x\|_2^3 + 1),
\]

where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 32 and our upper bound on $|\Delta_1|$ at the start of the proof. Defining $\Delta := (29) + (31)$, we have
\[
|\Delta| \leq 88p^*(x) d (L + 1)^3 (\|x\|_3^3 + 1) + 32\delta^{3/2}p^*(x) (L + 1)^{5/2} (\|x\|_2^5) \\
\leq p^*(x) \cdot 64\delta^{3/2} d (L + 1)^{5/2} (\|x\|_2^3 + 1).
\]
Lemma 6 For \( \delta \leq \frac{\text{min}(m^2, 1)}{2 \eta^3 (L + 1)^2} \),
\[
\left\| \mathbb{E}(\eta) \left[ \left( \int_0^1 \int_0^t \left( \nabla^3 p^* \left( (1 - t) x + t F_{-1}^{-1}(x) \right), (F_{-1}^{-1}(x) - x)^3 \right) \, dr \, ds \, dt \right) \cdot \det \left( \nabla F_{-1}^{-1}(x) \right)^{-1} \right] \right\|
\leq p^*(x) \cdot 2 \delta^{3/2} \exp \left( \frac{m}{2 \bar{\delta}^3} \|x\|^2 \right) (L + 1)^{3/2} \|x\|^2 + 1.
\]

Proof of Lemma 6 Using Lemma 4 and our choice of \( \delta \), \( \|x - F_{-1}^{-1}(x)\|_2 \leq \frac{3}{4} (\|x\|_2 + 1) \), and so \( \|F_{-1}^{-1}(x)\| \leq 2\|x\|_2 + 1 \). Thus for all \( t \in [0, 1] \),
\[
\|(1 - t)x + t F_{-1}^{-1}(x)\|_2 \leq 2\|x\|_2 + 1.
\]
Thus,
\[
\left\| \mathbb{E}(\eta) \left[ \left( \int_0^1 \int_0^t \left( \nabla^3 p^* \left( (1 - t) x + t F_{-1}^{-1}(x) \right), (F_{-1}^{-1}(x) - x)^3 \right) \, dr \, ds \, dt \right) \cdot \det \left( \nabla F_{-1}^{-1}(x) \right)^{-1} \right] \right\|
\leq p^*(x) \exp \left( 2L (\|x\|_2 + 1) \right) \cdot (L + 4L^2 (\|x\|_2 + 1) + 8L^3 \|x\|^2 + 1)
\leq p^*(x) \exp \left( 2\delta^{3/2} (\|x\|^2 + 1) \right) \cdot (L + 2 \delta L (\|x\|^2 + 1)) \cdot (L + 4 \delta^2 L^2 (\|x\|^2 + 1) + 8 \delta^3 \|x\|^2 + 1).
\]

where the first inequality is by Jensen’s inequality, the triangle inequality and the Cauchy-Schwarz inequality, the second inequality is by Lemmas 3, 4, and 4, the third inequality is by the fact that \( p^*(x) \propto \exp \left( -U(x) \right) \), by Assumption 3, and by (32) (we perform a first order Taylor expansion on \( U(x) \)), the fourth inequality is by our assumption on \( \sigma \) and some algebra, and the fifth inequality is by our assumption on \( \delta \). 

Lemma 7 For any \( \delta \leq \frac{\text{min}(m^2, 1)}{2 \eta^3 (L + 1)^2} \), for any \( x, y \) such that \( x = y - \delta \nabla U(y) + \sqrt{2 \delta} \mathcal{T}_n \) and for \( \eta \) a.s.,

1. \( \|y - x\|_2 \leq 4 \delta^{3/2} L^{1/2} (\|x\|_2 + 1) \),
2. \( \|y - x - \left( -\sqrt{2 \delta} \mathcal{T}_n + \delta \nabla U(x) \right)\|_2 \leq 4 \delta^{3/2} L^{1/2} (\|x\|_2 + 1) \),
3. \( \|y - x - \left( -\sqrt{2 \delta} \mathcal{T}_n(x) \right)\|_2 \leq 2 \delta L (\|x\|_2 + 1) \).

Proof of Lemma 7

1.
\[
\|y - x\|_2 = \left\| \delta \nabla U(y) + \sqrt{2 \delta} \mathcal{T}_n \right\|_2
\leq \|\delta \nabla U(x) + \sqrt{2 \delta} \mathcal{T}_n \|_2 + \delta \|\nabla U(y) - \nabla U(x)\|_2
\leq \|\nabla U(x) + \sqrt{2 \delta} \mathcal{T}_n\|_2 + \delta L \|y - x\|_2,
\]

where the first inequality is by triangle inequality, the second inequality is by Assumption 3,2. Moving terms around,
\[
(1 - \delta L)\|y - x\|_2 \leq \|\delta \nabla U(x) + \sqrt{2 \delta} \mathcal{T}_n\|_2
\leq \delta L \|x\|_2 + \sqrt{2 \delta} \mathcal{L}
\Rightarrow
\|y - x\|_2 \leq \left( \delta L + \sqrt{2 \delta} \mathcal{L} \right) (\|x\|_2 + 1)
\leq 2 \delta^{3/2} L^{1/2} (\|x\|^2 + 1),
\]

where the second inequality is by Assumptions 3,2 and 3,2, and the third inequality is by our assumption on \( \delta \).
where the first line is by definition of $x$ and $y$, the second line is by Assumption 1.2, and the third line is by Lemma 7.1. 

3. 
\[ \left\| y - x - \left( -\sqrt{2B}T_\eta + \delta \nabla U(x) \right) \right\|_2 \leq 2 \delta \sqrt{L} \left( \|x\|_2 + 1 \right), \]

where the first line is by triangle inequality, the second line is by Lemma 7.2 and Assumption 1.2, and the third line is by our assumption on $\delta$. 

Lemma 8 For any $\delta \leq \frac{1}{4dL}$, for any $x, y$ such that $x = y - \delta \nabla U(y) + \sqrt{2B}T_\eta(y)$ and for $\eta$ a.s.,
\[ |\text{tr} (\nabla^2 U(y)) - \text{tr} (\nabla^2 U(x))| \leq 4 \delta \sqrt{L} d \frac{3}{2} (\|x\|_2 + 1). \]

Proof of Lemma 8 
\[ |\text{tr} (\nabla^2 U(y)) - \text{tr} (\nabla^2 U(x))| = |\text{tr} (\nabla^2 U(y) - \nabla^2 U(x))| \leq d \left\| \nabla^2 U(y) - \nabla^2 U(x) \right\|_2 \leq dL \|x - y\|_2 \leq 4 \delta \sqrt{L} d \frac{3}{2} (\|x\|_2 + 1), \]

where the first inequality is by Lemma 11, the second inequality is by Assumption 1.4, and the third inequality is by Lemma 7.1. 

Lemma 9 For any $\delta \leq \frac{1}{4dL}$, for any $x$ and for $\eta$ a.s.,
\[ |\det (I - (\delta \nabla^2 U(x)))^{-1} - (1 + \delta \text{tr} (\nabla^2 U(x)))| \leq 64 \delta d^2 L^2. \]

Proof of Lemma 9 First, let’s consider an arbitrary symmetric matrix $A \in \mathbb{R}^{2d}$, let $c$ be a constant such that $\|A\|_2 \leq c$ and let $\epsilon$ be a constant satisfying $\epsilon \leq 1/(2cd)$. By Lemma 12, we have
\[ \det (I + \epsilon A) = 1 + \text{ctr} (A) + \frac{\epsilon^2}{2} (\text{tr} (A^2) - \text{tr} (A^2)) + \Delta \]
for some $|\Delta| \leq 2c^3 c^3 d^3$. Using a Taylor expansion, we can verify that for any $a \in [-1/2, 1/2]$
\[ |(1 + a)^{-1} - (1 - a + a^2)| \leq 2|a|^3. \]

By our assumption on $\epsilon$, we have $\text{ctr} (A) + \frac{\epsilon^2}{2} (\text{tr} (A^2) - \text{tr} (A^2)) + \Delta \in [-1/2, 1/2]$, therefore
\[
(1 + \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) + \Delta)^{-1} \\
\leq 1 - \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) - \Delta \\
+ \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) + \Delta^3 \\
+ 2 (\epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) + \Delta)^3 \\
\leq 1 - \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) + \epsilon^2 \text{tr} (A^2) + \epsilon^2 \text{tr} (A^2)^3 \\
+ 4 \left( \epsilon^2 c^2 d^2 + \epsilon^2 c^2 d^4 \right) + 16 \epsilon c d^3 + 16 \epsilon c^2 d^2 + \epsilon^3 c^2 d^3 \\
\leq 1 - \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) - \text{tr} (A^2)) + \epsilon^2 \text{tr} (A^2)^2 + 32 \epsilon c d^3 \\
= 1 - \epsilon \text{tr} (A) + \epsilon^2 / 2 (\text{tr} (A^2) + \text{tr} (A^2)) + 32 \epsilon c d^3. 
\]
where the first inequality is by Lemma 9, the second inequality is by the triangle inequality, the third inequality is by our assumption that \( \|A\|_2 \leq c \), by our assumption that \( |\Delta| \leq 2e^c d^3 \), and by Lemma 11 and the last two lines are by collecting terms. Conversely, one can show that
\[
(1 + \varepsilon \text{tr} (A) + e^2/2 \text{tr} (A^2)) + \Delta)^{-1}
\geq 1 - \varepsilon \text{tr} (A) + e^2/2 \text{tr} (A^2) - 32 (ecd)^3.
\]
The proof is similar and is omitted.

Therefore
\[
|\det (I + \varepsilon A) - (1 - \varepsilon \text{tr} (A) + e^2/2 \text{tr} (A^2))| \leq 32 (ecd)^3.
\] (34)

Now, we consider the case that \( A := -\nabla^2 U(x), \varepsilon := \delta \) and \( c := L \). Recall our assumption that \( \delta \leq \frac{1}{2cdL} \). Combined with Assumption 12, we get
\[
1. \|A\|_2 \leq c,
2. \varepsilon = \delta \leq 1/(2Ld) = 1/(2cd).
\]

Using (34),
\[
\det (I - \delta (\nabla^2 U(x)))^{-1} =: \det (I + \varepsilon A)^{-1} = 1 - \varepsilon \text{tr} (A) + e^2/2 \text{tr} (A^2) + 32 (ecd)^3
\leq 1 + \delta \text{tr} (\nabla^2 U(x))
+ \frac{\delta^2}{2} \text{tr} (\nabla^2 U(x))^2 + \frac{\delta^2}{2} \text{tr} \left((\nabla^2 U(x))^2\right)
+ 32\delta^3 d^3 L^3
\leq 1 + \delta \text{tr} (\nabla^2 U(x)) + 64\delta^5 d^2 L^2,
\]
where the first inequality is by (34), the first inequality is by definition of \( A \) and \( \varepsilon \), and the second inequality is by Assumption 12 and moving terms around.

Conversely, one can show that
\[
\det (I - \delta (\nabla^2 U(x)))^{-1} \geq 1 + \delta \text{tr} (\nabla^2 U(x)) - 64\delta^5 d^2 L^2
\]
The proof is similar and is omitted.

**Lemma 10** For any \( \delta \leq \frac{1}{2cdL} \), for any \( x \) and for \( \eta \) a.s.,
\[
\det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} = 1 + \delta \text{tr} (\nabla^2 U(x)) + \Delta
\]
for some \( |\Delta| \leq 8\delta^{3/2} dL^{3/2} (\|x\|_2 + 1) \).

**Proof of Lemma 10** Consider the Jacobian matrix inside the determinant. By definition of \( F_{\eta} \), we know that
\[
\nabla F_{\eta}(F_{\eta}^{-1}(x)) = I - \delta \nabla^3 U(F_{\eta}^{-1}(x)).
\]
Thus,
\[
\det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1}
= \det (I - \delta \nabla^2 U(F_{\eta}^{-1}(x)))^{-1}
\leq 1 + \delta \text{tr} (\nabla^2 U(F_{\eta}^{-1}(x))) + 64\delta^5 d^2 L^2
\leq 1 + \delta \text{tr} (\nabla^2 U(x)) + \delta \left| \text{tr} (\nabla^2 U(F_{\eta}^{-1}(x)) - \text{tr} (\nabla^2 U(x)) \right| + 64\delta^5 d^2 L^2
\leq 1 + \delta \text{tr} (\nabla^2 U(x)) + 48\delta^{3/2} dL^{3/2} (\|x\|_2 + 1) + 64\delta^5 d^2 L^2
\leq 1 + \delta \text{tr} (\nabla^2 U(x)) + 88\delta^{3/2} dL^{3/2} (\|x\|_2 + 1),
\]
where the first inequality is by Lemma 9, the second inequality is by the triangle inequality, the third inequality is by Lemma 8 and the fourth inequality is by our assumption that \( \delta \leq \frac{1}{2cdL} \). Conversely, one can show that
\[
\det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} \geq 1 + \delta \text{tr} (\nabla^2 U(x)) - 88\delta^{3/2} dL^{3/2} (\|x\|_2 + 1).
\]
The proof is similar and is omitted.

**Corollary 11** For any \( \delta \leq \frac{1}{2cdL} \), for any \( x \), and for \( \eta \) a.s.,
\[
\left| \det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} - 1 \right| \leq 2\delta dL (\|x\|_2 + 1).
\]
Proof of Corollary 11 From Lemma 10 we get
\[
\left| \det \left( \nabla F_0(F^{-1}_n(x)) \right) \right|^{-1} - 1 \leq \left| \det \left( \nabla F_0(F^{-1}_n(x)) \right) - (1 + \delta_1 tr(\nabla^2 U(x))) \right| + |\delta_1 tr(\nabla^2 U(x))| \\
\leq 8\delta^{1/2}dL^{1/2}(\|x\|_2 + 1) + \delta_1 dL\|x\|_2 \\
\leq 2\delta dL(\|x\|_2 + 1),
\]
where the first inequality is by the triangle inequality, the second inequality is by Lemma 10 and Assumption 1.2, and the third inequality is by our assumption on \(\delta\).

Lemma 12 Let \(p^*(x) \propto e^{-U(x)}\), for any \(q\) which is absolutely continuous \(p^*(x)\),
\[
W_2^2(p^*, q) \leq \frac{2}{m} \int \left( \frac{q(x)}{p^*(x)} - 1 \right)^2 p^*(x) \, dx.
\]

Proof of Lemma 12 By Theorems 1 and 2 (Talagrand’s Inequality) from Otto and Villani (2000), we see that if \(p^*(x) \propto e^{-U(x)}\) for an \(m\)-strongly-convex \(U(x)\) (Assumption 1), then for all \(q\) absolutely continuous \(p^*, \)
\[
W_2^2(q, p^*) \leq \frac{2}{m} KL(q||p^*).
\]
By the inequality \(t \log t \leq t^2 - t\), we get
\[
KL(q||p^*) = \int \frac{q(x)}{p^*(x)} \log \frac{q(x)}{p^*(x)} p^*(x) \, dx \\
\leq \int \left( \frac{q(x)}{p^*(x)} - \frac{q(x)}{p^*(x)} \right) p^*(x) \, dx \\
= \int \left( \frac{q(x)}{p^*(x)} - 1 \right)^2 p^*(x) \, dx.
\]
Combining the two inequalities, we get that
\[
W_2^2(q, p^*) \leq \frac{2}{m} \int \left( \frac{q(x)}{p^*(x)} - 1 \right)^2 p^*(x) \, dx.
\]

Lemma 13 For \(p^*(x) \propto e^{-U(x)}\), and for any \(x\),
1. \(\|\nabla p^*(x)\|_2 \leq p^*(x) \cdot (L\|x\|_2)\),
2. \(\|\nabla^2 p^*(x)\|_2 \leq p^*(x) \cdot (L + L^2\|x\|_2)\),
3. \(\|\nabla^3 p^*(x)\|_2 \leq (L + 2L^2\|x\|_2 + L^3\|x\|_2)\).

Proof of Lemma 13
1.
\[
\|\nabla p^*(x)\|_2 = \left\| e^{-U(x)}(-\nabla U(x)) \right\|_2 \\
\leq p^*(x) \cdot (L\|x\|_2),
\]
where the inequality is by Assumption 1.2.
2.
\[
\|\nabla^2 p^*(x)\|_2 = \left\| e^{-U(x)}(-\nabla^2 U(x) + \nabla U(x)\nabla U(x)^T) \right\|_2 \\
\leq p^*(x) \cdot (L + L^2\|x\|_2),
\]
where the inequality is by Assumption 1.2.
3.
\[
\|\nabla^3 p^*(x)\|_2 = p^*(x) \cdot \|\nabla^3 U(x) + \nabla^2 U(x) \otimes \nabla U(x) + \nabla U \otimes \nabla^2 U(x) - \nabla U(x) \otimes \nabla U(x) \otimes \nabla U(x) \|_2 \\
\leq p^*(x) \cdot (L + 2L^2\|x\|_2 + L^3\|x\|_2),
\]
where the inequality is by Assumptions 1.2 and 1.4.

Lemma 14 For any \(\delta \leq \frac{1}{16}\) and for any distributions \(p\) and \(q\), under the assumptions of Section 1.1,
\[
W_2(\Phi_{\delta}(p); \Phi_{\delta}(q)) \leq e^{-m^2/8}W_2(p, q).
\]

15
Proof of Lemma 14 Let $\gamma^*$ be an optimal coupling between $p$ and $q$, i.e.

$$W_2^2(p, q) = E_{\gamma^*(x,y)} \left[ \| x - y \|^2 \right].$$

We define a coupling $\gamma'$ as follows:

$$\gamma'(x, y) := (F_p, F_q) \# \gamma^*,$$

where # denotes the push-forward operator. (See (6) for the definition of $F_p$.) It is thus true by definition that $\gamma'$ is a valid coupling between $\Phi_3(p)$ and $\Phi_3(q)$. Thus,

$$W_2(\Phi_3(p), \Phi_3(q)) \leq E_{\gamma^*(x,y)} \left[ \| x - y \|^2 \right] = E_{\gamma^*(x,y)} \left[ \| F_p(x) - F_q(y) \|^2 \right] = E_{\gamma^*(x,y)} \left[ \| x - \delta \nabla U(x) + \sqrt{2\delta T_n} - (y - \delta \nabla U(y) + \sqrt{2\delta T_n}) \|^2 \right] \leq E_{\gamma^*(x,y)} \left[ (1 - m\delta/2) \| x - y \|^2 \right] \leq e^{-m\delta/4} W^2(p, q),$$

where the second inequality follows from Assumptions 1.2 and 1.3 and our assumption that $\delta \leq \frac{1}{2\sqrt{2L}}$, and the third inequality is by the fact that $m\delta/2 \leq m/(2L) \leq 1/2$.

B Auxiliary Lemmas for Section 4.2

Proof of Theorem 5 By Theorem 5 for

$$1 \leq \delta \leq \max \left\{ \frac{2^9 d^2 L}{2^{17} L^2} \frac{2^{27} L^2 \left( \frac{2}{m} \log \frac{2}{m}\right)^3}{2^{72} L^3 \theta^2 \frac{\log \left( \frac{2^2 L^2}{m}\right)}{m^3}} \frac{d^7 \epsilon^{-2} \cdot 2^{142} L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \lambda^{-2} \cdot \left( \frac{2}{m} \log \frac{2}{m}\right)^{12}}{d^7 \epsilon^{-2} \cdot 2^{142} L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \lambda^{-2} \cdot \left( \frac{2}{m} \log \left( 2^{224} d^7 L \left( \theta^3 + \theta^2 + \theta \right) \lambda^{-6} \epsilon^{-6} \right)^{12}} \right\},$$

we can bound

$$W_2(p_k, p^*) \leq e^{-4Lk} W_2(p_0, p^*) + \frac{\epsilon}{2}.$$ (35)

To prove (35), take the limit of (35) as $k \to \infty$.

Next, if

$$k \geq \frac{1}{\lambda \delta} \log \frac{2 W_2(p_0, p^*)}{\epsilon} = \frac{d^7}{\epsilon^2} \cdot \log \frac{2 W_2(p_0, p^*)}{\epsilon} \cdot \log \left( L, \theta, \frac{1}{m}, \frac{1}{\lambda}\right),$$

then $e^{-m\delta k/8} W_2(p_0, p^*) \leq \frac{\epsilon}{2}$, so we get

$$W_2(p_k, p^*) \leq \epsilon.$$ This proves (35).

Theorem 5 Let $p_0$ be an arbitrary initial distribution, and let $p_{k\delta}$ be defined as in (4). Let $\epsilon > 0$ be some arbitrary constant. For any stepsize $\delta$ satisfying

$$1 \leq \delta \leq \max \left\{ \frac{2^9 d^2 L}{2^{17} L^2} \frac{2^{27} L^2 \left( \frac{2}{m} \log \frac{2}{m}\right)^3}{2^{72} L^3 \theta^2 \frac{\log \left( \frac{2^2 L^2}{m}\right)}{m^3}} \frac{d^7 \epsilon^{-2} \cdot 2^{142} L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \lambda^{-2} \cdot \left( \frac{2}{m} \log \frac{2}{m}\right)^{12}}{d^7 \epsilon^{-2} \cdot 2^{142} L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \lambda^{-2} \cdot \left( \frac{2}{m} \log \left( 2^{224} d^7 L \left( \theta^3 + \theta^2 + \theta \right) \lambda^{-6} \epsilon^{-6} \right)^{12}} \right\},$$

we have

$$W_2(p_k, p^*) \leq \epsilon.$$ (36)
the Wasserstein distance between \( p_k \) and \( p^* \) is upper bounded as
\[
W_2(p_k, p^*) \leq e^{-\lambda \delta k} W_2(p_0, p^*) + \frac{\epsilon}{2}.
\]

**Proof of Theorem** We first use the triangle inequality to split the objective into two terms:
\[
W_2(\Phi_3^k(p_0), p^*) \leq W_2(\Phi_3^k(p_0), \Phi_3(p^*)) + W_2(\Phi_3(p^*), p^*)
\]
(37)

The first term is easy to bound. We use Assumption 5 to get
\[
W_2(\Phi_3^k(p_0), \Phi_3(p^*)) \leq e^{-\lambda \delta k} W_2(p_0, p^*)
\]
We now bound the second term of (37):
\[
W_2(\Phi_3(p^*), p^*) = W_2(\Phi_3(\Phi_3^{-1}(p^*)), p^*)
\leq W_2(\Phi_3(\Phi_3^{-1}(p^*)), \Phi_3(p^*)) + W_2(\Phi_3(p^*), p^*)
\leq e^{-\lambda \delta} W_2(\Phi_3^{-1}(p^*), p^*) + W_2(\Phi_3(p^*), p^*)
\leq \sum_{i=0}^{k-1} e^{-\lambda \delta i} W_2(\Phi_3(p^*), p^*)
\leq \frac{1}{\lambda \delta} W_2(\Phi_3(p^*), p^*).
\]
(38)

where the first inequality is by triangle inequality, the second inequality is by Assumption 5.

Next, we apply Lemma 15 to get
\[
W_2(\Phi_3(p^*), p^*) \leq 2^{70\delta^{1/2} \delta^{7/2} L (\theta^3 + \theta^2 + \theta)} \max \left\{ \frac{\epsilon^2}{m} \log \frac{\epsilon^2}{m}, \frac{\epsilon^2}{m} \log \left( \frac{1}{2^{124} d^2 L^2 (\theta^3 + \theta^2 + \theta) \delta^3} \right), 1 \right\}.^6
\]

Note that the first four clauses under (36) satisfy the requirement of Lemma 15.

There is a little trickiness due to the \( \log \frac{1}{\lambda \delta} \) term in the above upper bound. The calculations to get rid of the \( \log \frac{1}{\lambda \delta} \) term are packed away in Lemma 32. We verify that \( \delta \) satisfies the conditions (63) of Lemma 32 as the last 3 clauses of (60) implies,
\[
\frac{1}{\lambda \delta} \geq \frac{d^7 \lambda^2}{\epsilon^2} \cdot \frac{2^{78} L^2 (\theta^3 + \theta^2 + \theta)}{\lambda^2} \cdot \max \left\{ \frac{\epsilon^2}{m} \log \frac{\epsilon^2}{m}, \frac{\epsilon^2}{m} \log \left( \frac{1}{2^{124} d^2 L^2 (\theta^3 + \theta^2 + \theta) \delta^3} \right), 1 \right\}.^6
\]

Thus we can apply Lemma 32 to get
\[
\frac{1}{\lambda \delta} W_2(\Phi_3(p^*), p^*) \leq 2^{70\delta^{1/2} \delta^{7/2} L (\theta^3 + \theta^2 + \theta)} \max \left\{ \frac{\epsilon^2}{m} \log \frac{\epsilon^2}{m}, \frac{\epsilon^2}{m} \log \left( \frac{1}{2^{124} d^2 L^2 (\theta^3 + \theta^2 + \theta) \delta^3} \right), 1 \right\}.^6 \lambda^{-1}
\]
\[
\leq \frac{\epsilon}{2}.
\]
(39)

The conclusion follows by substituting (38) and (39) into (37).
Proof of Lemma 16
Let us define the radius
\[ R := 2\sqrt{\max\left\{ \frac{c^2}{m} \log \frac{c^2}{m}, \frac{c^2}{m} \log \left( \frac{1}{2^{124}R^6L^2(\theta^3 + \theta^2 + \theta)^2} \right), 1 \right\} } \]
We can verify that by the definition of \( R \) and our assumptions on \( \delta \),
\[ R \geq \sqrt{\max\left\{ 2^{13} \frac{c^2}{m} \left( \log \left( \frac{2^{13}c^2}{m} \right) \right), 1 \right\} } \]
and \( \delta \leq \frac{1}{\sqrt{L}} \), so we can apply Corollary 40 to give
\[
W_2^2(p^*, p_\delta) \leq 4R^2 \int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x)dx + 84d \exp \left( \frac{mR^2}{64c^2} \right)
\]
where the second inequality follows from the definition of \( R \), which implies that \( R \geq \frac{c^2}{m} \log \left( \frac{1}{2^{124}R^6L^2(\theta^3 + \theta^2 + \theta)^2} \right) \).

Next, we apply Lemma 41 which shows that under our assumptions on \( \delta \) and our definition of \( R \),
\[ \delta \leq \min \left\{ \frac{1}{2^9L^2}, \frac{1}{2^{11}L^2(\rho^2 + 1)} \right\} \]
We can thus apply Lemma 16 to get
\[
\int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x)dx \leq 2^{10} \delta^3d^6L^2 (\theta^3 + \theta^2 + \theta)^2 \int_{B_R} (\|x\|_2^2 + 1) p^*(x)dx
\]
where the second inequality follows from Lemma 38.

Plugging the above into 40, we get
\[ = 4R^2 \left( 2^{124} \delta^3d^6L^2 (\theta^3 + \theta^2 + \theta)^2 \max \left\{ \frac{c^2}{m} \log \frac{c^2}{m}, 1 \right\}^{10} \right) + \left( 2^{124} \delta^3d^6L^2 (\theta^3 + \theta^2 + \theta)^2 \right) \]
where the first line is by 41 and 40, the second line is because \( R \geq 1 \), the third line is again by definition of \( R \) and some algebra.

Lemma 16 Let \( p_\delta := \Phi_\delta(p^*) \). For any \( R \geq 0 \), for all \( x \in B_R \), and for all \( \delta \leq \min \left\{ \frac{1}{2^9L^2}, \frac{1}{2^{11}L^2(\rho^2 + 1)} \right\} \)
\[ \left| \frac{p_\delta(x)}{p^*(x)} - 1 \right| \leq 2^{15} \delta^{3/2}d^3L^{3/2} (\theta^3 + \theta^2 + \theta) (\|x\|_2^3 + 1) . \]

Proof of Lemma 16 By the definition 7, \( p_\delta(p^*) = (F_\eta)_\# p^* \). The change of variable formula gives
\[
 p_\delta(x) = \int p^* (F^{-1}_\eta(x)) \det(\nabla F_\eta (F^{-1}_\eta(x)))^{-1} q(\eta)d\eta
\]
\[ \begin{aligned}
\Delta \left\{ p^* (F^{-1}_\eta(x)) \det(\nabla F_\eta (F^{-1}_\eta(x)))^{-1} \right\} ,
\quad (42)
\end{aligned}
\]
where in the above, $\nabla F_\eta(y)$ denotes the Jacobian matrix of $F_\eta$ at $y$. The invertibility of $F_\eta$ is proven in Lemma \[1\] We now rewrite (1) as its Taylor expansion:

$$p^*(F_\eta^{-1}(x)) = p^*(x) + \langle \nabla p^*(x), F_\eta^{-1}(x) - x \rangle + \frac{1}{2} \langle \nabla^2 p^*(x), (F_\eta^{-1}(x) - x) (F_\eta^{-1}(x) - x)^T \rangle + \int_0^1 \int_0^1 \int_0^1 \langle \nabla^2 p^* \left( (1 - r)x + r F_\eta^{-1}(x) \right), (F_\eta^{-1}(x) - x)^2 \rangle \, dr \, ds \, dt .$$

Putting everything together, we get

$$p^*_\delta(x) = \mathbb{E}_\eta \left[ (1 + 3 + 10 + 43) \cdot 2 \right] = p^*(x) + \delta p^*(x) \left( \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} + \delta \text{tr} \left( \nabla^2 U(x) \right) \right)
+ \delta \left( \sum_{i=1}^{d} \frac{\partial}{\partial x_i} p^*(x) \cdot \frac{\partial}{\partial x_i} U(x) \right)
+ 2 \delta \sum_{i=1}^{d} \left( \frac{\partial}{\partial x_i} p^*(x) \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} \right)
+ \delta \left( \nabla^2 p^*(x), \sigma_x \sigma_x^T \right) + \Delta
$$

$$= p^*(x) + \Delta$$

(43)

The third equality is by Lemma \[10\]. The second equality is by Lemmas \[17, 18, 19\] and \[20\]. Note that by our assumption that $x \in B_R$ and $\delta \leq \min \left\{ \frac{1}{\sqrt{d}}, \frac{7}{256 \sqrt{d} \log(d+1) \mu} \right\}$, $\delta$ satisfies the condition for Lemmas \[17, 18, 19\] and \[20\]. Also by these four lemmas, we have

$$|\Delta| \leq p^*(x) \cdot 128 \lambda^{d/2} d^3 L^{3/2} (\|x\|_2^2 + 1)
+ p^*(x) \cdot 256 \lambda^{d/2} d^2 L^{3/2} \theta (\|x\|_2^2 + 1)
+ p^*(x) \cdot 256 \lambda^{d/2} dL^{d/2} \theta^2 + \theta (\|x\|_2^{10} + 1)
+ p^*(x) \cdot 2^{14} \lambda^{d/2} L^{3/2} (\theta^3 + \theta^2 + \theta) (\|x\|_2^{12} + 1)
\leq p^*(x) \cdot 2^{15} \lambda^{d/2} d^4 L^{3/2} (\theta^3 + \theta^2 + \theta) (\|x\|_2^{11} + 1).$$

As a consequence,

$$\frac{|p^*_\delta(x)|}{p^*(x)} - 1 \leq 2^{15} \lambda^{d/2} d^4 L^{3/2} (\theta^3 + \theta^2 + \theta) (\|x\|_2^{11} + 1).$$

**Lemma 17**: For $\delta \leq \frac{1}{\sqrt{d} \lambda^{d/2}}$,

$$\mathbb{E}_\eta \left[ \frac{p^*(x) \cdot \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right) }{p^*(x)} \right] = p^*(x) + \delta p^*(x) \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} + \delta \text{tr} \left( \nabla^2 U(x) \right) + \Delta,$$

for some $|\Delta| \leq p^*(x) \cdot 128 \lambda^{d/2} d^3 L^{3/2} (\|x\|_2^2 + 1)$.

**Proof of Lemma 17**: Let us define

$$\Delta' := \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right)^{-1}
- \left( 1 - \sqrt{2 \delta \text{tr} \left( G_\eta(x) \right) } \right) + 2 \delta \text{tr} \left( M_\eta(x), T_\eta(x) \right)
+ \delta \text{tr} \left( \nabla^2 U(x) \right) + \delta \text{tr} \left( (G_\eta(x))^2 \right).$$

19
By Lemma 22, $|\Delta'| \leq 128\delta^{3/2}d^3L^{3/2}(\|x\|_2^2 + 1)$. Hence,

$$E_{q(x)} \left[ p^*(x) \cdot \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) \right]$$

$$= E_{q(x)} \left[ p^*(x) \cdot \left( 1 - \sqrt{2\delta} \text{tr} \left( G_q(x) \right) + 2\delta \text{tr} \left( \{M_q(x), T_q(x)\}_{\ell} \right) + \delta \text{tr} \left( \nabla^2 U(x) + 2 \text{tr} \left( G_q(x)^2 \right) + 2 \text{tr} \left( (G_q(x))^2 \right) \right) \right]$$

$$+ E_{q(x)} \left[ p^*(x) \cdot \Delta' \right]$$

$$= p^*(x) + p^*(x) E_{q(x)} \left[ \left( -\sqrt{2\delta} \text{tr} \left( G_q(x) \right) + 2\delta \text{tr} \left( \{M_q(x), T_q(x)\}_{\ell} \right) + \delta \text{tr} \left( G_q(x)^2 \right) + 2 \text{tr} \left( (G_q(x))^2 \right) \right) \right]$$

$$+ p^*(x) \delta \text{tr} \left( \nabla^2 U(x) \right) + p^*(x) \cdot \Delta'.$$

We complete the proof by taking $\Delta := p^*(x)\Delta'$.

**Lemma 18** Let $\delta \leq \frac{1}{2^{2d}L}$,

$$E_{q(x)} \left[ \left( \nabla p^*(x), F_{q(x)}^{-1}(x) \cdot x \right) \cdot \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) \right]$$

$$= \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} p^*(x) \cdot \left( \frac{\partial}{\partial x_i} U(x) + 2 \sum_{j=1}^d \frac{\partial}{\partial x_j} [\sigma_x \sigma_y^T]_{i,j} \right) \right) + \Delta$$

for some $|\Delta| \leq p^*(x) \cdot 256\delta^{3/2}d^3L^{3/2}(\|x\|_2^2 + 1)$.

**Proof of Lemma 18** Let

$$\Delta_1 := F_{q(x)}^{-1}(x) - x - \left( -\sqrt{2\delta} T_q(x) + \sqrt{\delta} U(x) + 2 \delta G_q(x) T_q(x) \right),$$

$$\Delta_2 := \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) - 1,$$

$$\Delta_3 := \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) - \left( 1 - \sqrt{2\delta} \text{tr} \left( G_q(x) \right) \right).$$

By Lemma 22, Corollary 24 and Corollary 26,

$$\|\Delta_1\|_2 \leq 16\delta^{3/2}L^{3/2}(\|x\|_2^2 + 1),$$

$$|\Delta_2| \leq 2\delta^{3/2}dL^{1/2}(\|x\|_2^2 + 1),$$

$$|\Delta_3| \leq 8\delta d^2L(\|x\|_2^2 + 1).$$

Moving terms around,

$$E_{q(x)} \left[ \left( \nabla p^*(x), F_{q(x)}^{-1}(x) \cdot x \right) \cdot \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) \right]$$

$$= E_{q(x)} \left[ \left( \nabla p^*(x), -\sqrt{2\delta} T_q(x) \right) \cdot \left( 1 - \sqrt{2\delta} \text{tr} \left( G_q(x) \right) \right) \right]$$

$$+ E_{q(x)} \left[ \left( \nabla p^*(x), \sqrt{\delta} U(x) + 2 \delta G_q(x) T_q(x) \right) \right]$$

$$+ E_{q(x)} \left[ \left( \nabla p^*(x), -\sqrt{2\delta} T_q(x) \right) \cdot \Delta_3 \right]$$

$$+ E_{q(x)} \left[ \left( \nabla p^*(x), \sqrt{\delta} U(x) + 2 \delta G_q(x) T_q(x) \right) \cdot \Delta_2 \right]$$

$$+ E_{q(x)} \left[ \Delta_1 \cdot \text{det} \left( \nabla F_{q(x)}(F_{q(x)}^{-1}(x)) \right) \right].$$

The main term of interest are (44) and (45), which evaluate to

$$E_{q(x)} \left[ \left( \nabla p^*(x), -\sqrt{2\delta} T_q(x) \right) \cdot \left( 1 - \sqrt{2\delta} \text{tr} \left( G_q(x) \right) \right) \right]$$

$$+ E_{q(x)} \left[ \left( \nabla p^*(x), \sqrt{\delta} U(x) + 2 \delta G_q(x) T_q(x) \right) \right]$$

$$= \left( \nabla p^*(x), \sqrt{\delta} U(x) + 2 \delta G_q(x) T_q(x) \right)$$

$$+ 2\delta E_{q(x)} \left[ \text{tr} \left( G_q(x) T_q(x) \right) \right]$$

$$= \delta \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} p^*(x) \cdot \left( \frac{\partial}{\partial x_i} U(x) + 2 \sum_{j=1}^d \frac{\partial}{\partial x_j} [\sigma_x \sigma_y^T]_{i,j} \right) \right),$$

where the first equality is by Assumption 21, and the last equality is by Lemma 15. We now consider the terms (44), (47) and (48):

$$\|\Delta_1\| \leq \|\nabla p^*(x)\| \sqrt{2\delta} E_{q(x)} \left[ |T_q(x)| \cdot |\Delta_3| \right]$$

$$\leq p^*(x) \delta (\|x\|_2^2 + 1) \cdot \sqrt{2\delta} L(\|x\|_2^2 + 1) \cdot 8\delta d^2L(\|x\|_2^2 + 1)$$

$$\leq 16\delta^{3/2}p^*(x)d^3L^{3/2}(\|x\|_2^2 + 1).$$
where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 28.1 and our upperbound on $|\Delta_3|$ at the start of the proof.

$$\leq p^*(x) \theta \cdot (||x||^2 + 1) \cdot 2\Delta L \cdot (||x||^2 + 1) \cdot 2\Delta L^1/2 \cdot (||x||^2 + 1) \leq 32\Delta^3/2 \cdot p^*(x) \cdot L \cdot \theta (||x||^2 + 1)$$

where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 28.1 and our upperbound on $|\Delta_3|$ at the start of the proof.

$$\leq 16\Delta^3/2 \cdot p^*(x) \cdot L \cdot \theta (||x||^2 + 1) + 16\Delta^3/2 \cdot p^*(x) \cdot dL \cdot \theta (||x||^2 + 1) \leq p^*(x) \cdot 256\Delta^3/2 \cdot dL \cdot \theta (||x||^2 + 1)$$

where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 28.1 and our upperbound on $|\Delta_3|$ and $|\Delta_4|$ at the start of the proof.

Defining $\Delta := (47) + (48) + (45)$, we have

$$|\Delta| \leq 16\Delta^3/2 \cdot p^*(x) \cdot dL \cdot \theta (||x||^2 + 1) + 16\Delta^3/2 \cdot p^*(x) \cdot dL \cdot \theta (||x||^2 + 1) + 128\Delta^3/2 \cdot p^*(x) \cdot dL \cdot \theta (||x||^2 + 1) \leq p^*(x) \cdot 256\Delta^3/2 \cdot dL \cdot \theta (||x||^2 + 1)$$

Lemma 19 For $\delta \leq \frac{1}{2\sqrt{\theta L}}$,

$$\frac{1}{2} \mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), (F_{\eta}^{-1}(x) - x) (F_{\eta}^{-1}(x) - x)^T \cdot \det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} \right) \right] = \delta \left( \nabla^2 p^*(x), \sigma_x \sigma_x^T \right) + \Delta$$

for some $|\Delta| \leq p^*(x) \cdot 256\Delta^3/2 \cdot dL \cdot \theta (||x||^3 + 1)$. Proof of Lemma 19 Define

$$\Delta_1 := F_{\eta}^{-1}(x) - x - \sqrt{2\delta} T_{\eta}(x), \quad \Delta_2 := \det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} - 1.$$

By Lemma 28.3 and Corollary 27,

$$|\Delta_1| \leq 16\delta L \cdot (||x||^2 + 1) \quad \text{and} \quad |\Delta_2| \leq 2\delta L^1/2 \cdot (||x||^2 + 1)$$

Then

$$\mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), (F_{\eta}^{-1}(x) - x) (F_{\eta}^{-1}(x) - x)^T \cdot \det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} \right) \right] = 2\delta \mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), T_{\eta}(x) T_{\eta}(x)^T \right) \right] + 2\delta \mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), T_{\eta}(x) T_{\eta}(x)^T \right) \cdot \Delta_2 \right] + \mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), \Delta_1 \Delta_1^T - \sqrt{2\delta} T_{\eta}(x) \Delta_1^T - \sqrt{2\delta} \Delta_1 (x)^T \cdot \det (\nabla F_{\eta}(F_{\eta}^{-1}(x)))^{-1} \right) \right].$$

We are mainly interested in (49), which evaluates to

$$2\delta \mathbb{E}_{v(x)} \left[ \left( \nabla^2 p^*(x), T_{\eta}(x) T_{\eta}(x)^T \right) \right] = 2\delta \left( \nabla^2 p^*(x), \mathbb{E}_{v(x)} \left[ T_{\eta}(x)^T \right] \right) = 2\delta \left( \nabla^2 p^*(x), \sigma_x \sigma_x^T \right),$$

where the last equality is by definition of $T_{\eta}(x)$ and $\sigma_x$. We now bound the magnitude of (50) and (51).
where the first inequality is by Cauchy-Schwarz, and the second inequality is by Lemma 28.2 and our upper bound on $|\Delta_2|$ at the start of the proof.

\[
\begin{aligned}
|\Delta_1| &\leq \mathbb{E}_{\eta}(\mathbb{D}_3) \left[ \left( \nabla^2 p^\ast(x), \Delta_1 \Delta_1^T + \sqrt{2a} \eta(x) \Delta_1^T + \sqrt{2a} \Delta_1 \eta(x) \right)^T \cdot \det \left( \nabla F_\eta(F_\eta^{-1}(x)) \right)^{-1} \right] \\
\leq &\left\| \nabla^2 p^\ast(x) \right\| \mathbb{E}_{\eta}(\mathbb{D}_3) \left[ \left( \left| \Delta_1 \right|^2 + 2 \sqrt{2a} \left\| \eta(x) \right\|_2 \left\| \Delta_1 \right\|_2 \right) \cdot \det \left( \nabla F_\eta(F_\eta^{-1}(x)) \right)^{-1} \right] \\
\leq &p^\ast(x) \left( \theta^2 + \theta \right) \left( \left\| x \right\|_2^2 + 1 \right) \cdot \left[ \left( 16\delta L \left( \left\| x \right\|_2^2 + 1 \right) \right)^2 + 2 \sqrt{2a} \left( L^{1/2} \left( \left\| x \right\|_2^2 + 1 \right) \right) \right] \\
\leq &25\delta^3/2 p^\ast(x) dL^{1/2} \left( \theta^2 + \theta \right) \left( \left\| x \right\|_2^2 + 1 \right) ,
\end{aligned}
\]

where the first inequality is by Jensen’s inequality, the triangle inequality and Cauchy-Schwarz, and the second inequality is by Lemma 28.2 and our upper bound on $|\Delta_1|$ at the start of the proof.

Defining $\Delta := \left( \frac{30}{2} + \frac{31}{2} \right)$, we have

\[
\left\| \Delta \right\| \leq 32\delta^3/2 p^\ast(x) dL^{1/2} \left( \theta^2 + \theta \right) \left( \left\| x \right\|_2^2 + 1 \right) + 256\delta^3/2 p^\ast(x) dL^{1/2} \left( \theta^2 + \theta \right) \left( \left\| x \right\|_2^2 + 1 \right) .
\]

Lemma 20 For $\delta \leq \min \left\{ \frac{2\delta L \left( \theta^2 + \theta \right)}{25 \left( \left\| x \right\|_2^2 + 1 \right)} , \frac{1}{10} \right\}$,

\[
\left\| \mathbb{E}_{\eta}(\mathbb{D}_3) \left[ \int_0^1 \int_0^1 \int_0^{\infty} \left\| \nabla^3 p^\ast \left( \left( 1-t \right) x + t F_\eta^{-1}(x) \right) \left( F_\eta^{-1}(x) - x \right)^3 \right\| \cdot \det \left( \nabla F_\eta(F_\eta^{-1}(x)) \right)^{-1} \right] \right\| \left\| \left( 1-t \right) x + t F_\eta^{-1}(x) \right\|_2 \leq 2 \left\| x \right\|_2 + 1 ,
\]

and

\[
\left\| \mathbb{E}_{\eta}(\mathbb{D}_3) \left[ \int_0^1 \int_0^1 \int_0^{\infty} \left\| \nabla^3 p^\ast \left( \left( 1-t \right) x + t F_\eta^{-1}(x) \right) \right\| \cdot \det \left( \nabla F_\eta(F_\eta^{-1}(x)) \right)^{-1} \right] \right\| \left\| \left( 1-t \right) x + t F_\eta^{-1}(x) \right\|_2 \leq 2 \left\| x \right\|_2 + 1 ,
\]

Proof of Lemma 20 Using Lemma 22.1, by our choice of $\delta$, $\left\| x - F_\eta^{-1}(x) \right\|_2 \leq 1/3 \left( \left\| x \right\|_2 + 1 \right)$, thus $\left\| F_\eta^{-1}(x) \right\| \leq 2 \left\| x \right\|_2 + 1$. Hence, for all $t \in [0, 1]$,

\[
\left\| \left( 1-t \right) x + t F_\eta^{-1}(x) \right\|_2 \leq 2 \left\| x \right\|_2 + 1 ,
\]

Lemma 21 For any $\delta$, for any $x, y$, and for $\eta$ a.s.,

1. $\left\| T_\eta(x) - T_\eta(y) \right\| \leq L^{1/2} \left\| x - y \right\|_2$,
2. $\left\| G_\eta(x) - G_\eta(y) \right\| \leq L^{1/2} \left\| x - y \right\|_2$,
3. $\left\| T_\eta(y) - T_\eta(x) - G_\eta(x)(y - x) \right\| \leq L^{1/2} \left\| y - x \right\|_2$,
4. $\left\| G_\eta(x) - G_\eta(y) - \left( M(x), y - x \right) \right\| \leq \left\| x - y \right\|_2$.

Proof of Lemma 21
1. We use Assumption 4.3 and a Taylor expansion:

\[
\|T_\eta(x) - T_\eta(y)\|_2 = \left\| \int_0^1 G_\eta(t(x) + (1-t)(x-y)) \, dt \right\|_2 \\
\leq L^{1/2} \|x - y\|_2.
\]

2. We use Assumption 4.4 and a Taylor expansion:

\[
\|G_\eta(x) - G_\eta(y)\|_2 = \left\| \int_0^1 M_\eta((1-t)x + ty) \, dt \right\|_2 \\
\leq L^{1/2} \|x - y\|_2.
\]

3. Using Taylor’s theorem and the definitions of \(T_\eta\), \(G_\eta\) and \(M_\eta\) from Assumption 4,

\[
T_\eta(y) = T_\eta(x) + \int_0^1 \langle G_\eta((1-t)x + ty), (y-x) \rangle \, dt \\
= T_\eta(x) + \int_0^1 \left( \langle G_\eta(x) + \int_0^t \langle M_\eta((1-s)x + sy), (y-x) \rangle \, ds, (y-x) \rangle \right) \, dt \\
= T_\eta(x) + \langle G_\eta(x), y - x \rangle + \int_0^1 \int_0^t \langle \langle M_\eta((1-s)x + sy), y - x \rangle, y - x \rangle \, ds \, dt.
\]

Therefore,

\[
\|T_\eta(y) - T_\eta(x) - G_\eta(x)(y-x)\|_2 \\
\leq \int_0^1 \int_0^t \left\| \langle \langle M_\eta((1-s)x + sy), y - x \rangle, y - x \rangle \right\|_2 \, ds \, dt \\
\leq \int_0^1 \int_0^t \|M_\eta((1-s)x + sy)\|_2 \|y - x\|_2^2 \, ds \, dt \\
\leq L^{1/2} \|y - x\|_2^2,
\]

where the first inequality is by the triangle inequality, the second inequality is by definition of the \(\| \cdot \|_2\) norm in \(\mathbb{S}\), and the last inequality is by Assumption 4.4.

4. Using Taylor’s theorem and the definitions of \(T_\eta\), \(G_\eta\), \(M_\eta\) and \(N_\eta\),

\[
G_\eta(y) = G_\eta(x) + \int_0^1 M_\eta((1-t)x + ty)(y-x) \, dt \\
= G_\eta(x) + \int_0^1 \left( M_\eta(x) + \int_0^t \langle N_\eta((1-s)x + sy), (y-x) \rangle \, ds \right) (y-x) \, dt \\
= G_\eta(x) + M_\eta(x)(y-x) + \int_0^1 \int_0^t \langle \langle N_\eta((1-s)x + sy), y - x \rangle, y - x \rangle \, ds \, dt.
\]

Therefore,

\[
\|G_\eta(y) - G_\eta(x) - M_\eta(x)(y-x)\|_2 \\
\leq \int_0^1 \int_0^t \left\| \langle \langle N_\eta((1-s)x + sy), y - x \rangle, y - x \rangle \right\|_2 \, ds \, dt \\
\leq \int_0^1 \int_0^t \|N_\eta((1-s)x + sy)\|_2 \|y - x\|_2^2 \, ds \, dt \\
\leq L^{1/2} \|y - x\|_2^2,
\]

where the first inequality is by our expansion above and Jensen’s inequality, the second inequality is by definition of \(\| \cdot \|_2\) in \(\mathbb{S}\), and the third inequality is by Assumption 4.5.

**Lemma 22** For any \(\delta \leq \frac{1}{2L^2}\), for any \(x, y\) such that \(x = F_\eta(y)\) and for \(\eta\) a.s.

1. \(\|y - x\|_2 \leq 2\delta^{1/2}L^{1/2}(\|x\|_2 + 1)\),
2. \(\|y - x - \left( -\sqrt{2\delta}T_\eta(x) + \delta \nabla U(x) + 2\delta G_\eta(x)T_\eta(x) \right)\|_2 \leq 16\delta^{3/2}L^{3/2}(\|x\|_2^3 + 1)\),
3. \(\|y - x - \left( -\sqrt{2\delta}T_\eta(x) \right)\|_2 \leq 16\delta L(\|x\|_2^3 + 1)\).

**Proof of Lemma 22**
1. We first bound the expression
\[ \|y - x\|_2 = \left\| \delta \nabla U(y) + \sqrt{2\delta} T_n(y) \right\|_2 \]
\[ \leq \left\| \delta \nabla U(x) + \sqrt{2\delta} T_n(x) \right\|_2 + \delta \| \nabla U(y) - \nabla U(x) \|_2 + \sqrt{2\delta} \| T_n(y) - T_n(x) \|_2 \]
\[ \leq \left\| \delta \nabla U(x) + \sqrt{2\delta} T_n(x) \right\|_2 + \delta L \| y - x \|_2 + \sqrt{2\delta L} \left( \| y - x \|_2 + 1 \right), \]
where the first inequality is by the triangle inequality, the second inequality is by Assumptions\(^1\)2 and \(^3\)3.

Moving terms around,
\[ (1 - \delta L - \sqrt{2\delta L}) \| y - x \|_2 \leq \left\| \delta \nabla U(x) + \sqrt{2\delta} T_n(x) \right\|_2 \]
\[ \leq \delta L \| x \|_2 + \sqrt{2\delta L} (\| x \|_2 + 1) \]
\[ \Rightarrow \]
\[ \| y - x \|_2 \leq \frac{2}{\delta L + \sqrt{2\delta L}} \left( \| x \|_2 + 1 \right), \]
where the second inequality is by Assumptions\(^1\)1, \(^1\)2 and \(^1\)2, and the third inequality is by our assumption that \( \delta \leq 1/(32L). \)

2. We first bound the expression \( T_n(y) - T_n(x) + \sqrt{2\delta G_n(x)} T_n(x). \)
Plugging in \( x = F_n(y) := y - \delta \nabla U(y) + \sqrt{2\delta} T_n(y), \) we get
\[ \| T_n(y) - T_n(x) - G_n(x)(y - x) \|_2 \]
\[ = \left\| T_n(y) - T_n(x) - G_n(x) \left( \delta \nabla U(y) - \sqrt{2\delta} T_n(y) \right) \right\|_2 \]
\[ \geq \left\| T_n(y) - T_n(x) - G_n(x) \left( -\sqrt{2\delta} T_n(x) \right) \right\|_2 \]
\[ - \| G_n(x) (\delta \nabla U(x) - \delta \nabla U(y)) \|_2 \]
\[ - \left\| G_n(x) \left( \sqrt{2\delta} T_n(x) - \sqrt{2\delta} T_n(y) \right) \right\|_2 \]
\[ - \| G_n(x) \delta \nabla U(x) \|_2 \]
\[ \geq \left\| T_n(y) - T_n(x) - G_n(x) \left( -\sqrt{2\delta} T_n(x) \right) \right\|_2 \]
\[ - \delta L^{3/2} \| x - y \|_2 - \sqrt{2\delta L} \| x - y \|_2 - \delta L^{3/2} \| x \|_2, \]
where the first inequality is by triangle inequality, and the second inequality is by Assumptions \(^1\)1 and \(^1\)1 and Lemma \(^2\). Moving terms around, we get
\[ \left\| T_n(y) - T_n(x) + \sqrt{2\delta G_n(x)} T_n(x) \right\|_2 \]
\[ \leq \| T_n(y) - T_n(x) - G_n(x)(y - x) \|_2 + \delta L^{3/2} \| x - y \|_2 + \sqrt{2\delta L} \| x - y \|_2 + \delta L^{3/2} \| x \|_2 \]
\[ \leq L^{1/2} \| x - y \|_2^2 + \delta L^{3/2} \| x - y \|_2 + \sqrt{2\delta L} \| x - y \|_2 + \delta L^{3/2} \| x \|_2 \]
\[ \leq 8\delta L^{1/2} (\| x \|_2^2 + 1), \] (53)
where the second inequality is by Lemma \(^2\)\(^1\)3, Lemma \(^2\)\(^1\)1, and Young’s Inequality, and the third inequality is by our assumption that \( \delta \leq 1/(32L). \) Finally, by definition of \( F_n(x), \)
\[ x = y - \delta \nabla U(y) + \sqrt{2\delta} T_n(y) \]
\[ \Rightarrow \]
\[ y - x = \left( -\sqrt{2\delta} T_n(x) + \sqrt{2\delta} T_n(y) \right) \]
\[ \Rightarrow \]
\[ \left\| \delta \nabla U(y) - \sqrt{2\delta} T_n(y) \right\|_2 \]
\[ = \left\| \delta \nabla U(y) - \nabla U(x) + \left( \nabla U(x) - \sqrt{2\delta} T_n(x) \right) \right\|_2 \]
\[ = \left\| \delta \nabla U(y) - \nabla U(x) \right\|_2 + \sqrt{2\delta} \left( T_n(x) - T_n(y) + \sqrt{2\delta} G_n(x) T_n(x) \right) \]
\[ \leq \delta \| \nabla U(x) - \nabla U(y) \|_2 + \sqrt{2\delta} \left( T_n(y) - T_n(x) + \sqrt{2\delta} G_n(x) T_n(x) \right) \]
\[ \leq \delta L \| x - y \|_2 + 8\sqrt{2\delta^{3/2} L^{3/2}} (\| x \|_2^2 + 1) \]
\[ \leq 2\delta^{3/2} L^{3/2} (\| x \|_2 + 1) + 8\sqrt{2\delta^{3/2} L^{3/2}} (\| x \|_2^2 + 1) \]
\[ \leq 16\delta^{3/2} L^{3/2} (\| x \|_2^2 + 1), \]
where the first inequality is by triangle inequality, the second inequality is by Assumptions \(^1\)1 and \(^5\)3, and the third inequality is by Lemma \(^2\).
Lemma 23

For any $x \neq y$, where the second inequality is by Lemma 21.4, and the third inequality is by Lemma 22.1 and our assumption that $\delta \leq 1/(32L)$. The last inequality is by our assumption that $\delta \leq 1/(32L)$.

Proof of Lemma 23

1. By our definition of $x$ and $y$,

\[
\left\| G_{\eta}(y) - G_{\eta}(x) - (M_{\eta}(x), y - x) \right\|_2 \\
\geq \left\| G_{\eta}(y) - G_{\eta}(x) - \left( M_{\eta}(x), \sqrt{2\delta}T_{\eta}(x) \right) \right\|_2
\]

where the first equality is by definition of $x$ and $y$, the first inequality is by the triangle inequality, and the second inequality is by Assumptions [1] and [4] and Lemmas 21.1 and 22.1. Moving terms around, we get

\[
\left\| G_{\eta}(y) - G_{\eta}(x) + \left( M_{\eta}(x), \sqrt{2\delta}T_{\eta}(x) \right) \right\|_2 \\
\leq L^{1/2} \|x - y\|_2 + \sqrt{2\delta} L \|x - y\|_2 + \delta L^{3/2} \|x\|_2 \\
\leq 8\delta L^{3/2} (\|x\|_2^2 + 1),
\]

where the second inequality is by Lemma 21.4, and the third inequality is by Lemma 22.1 and our assumption that $\delta \leq 1/(32L)$. Finally, using the inequality $\|A\|_2 \leq \|A\|_2$ from Lemma 1, we get

\[
\left\| \text{tr} \left( G_{\eta}(y) - G_{\eta}(x) + \left( M_{\eta}(x), \sqrt{2\delta}T_{\eta}(x) \right) \right) \right\|_2 \\
\leq d \left\| G_{\eta}(y) - G_{\eta}(x) + \left( M_{\eta}(x), \sqrt{2\delta}T_{\eta}(x) \right) \right\|_2 \\
\leq 8\delta d L^{3/2} (\|x\|_2^2 + 1).
\]
Lemma 24
For any $\parallel A \parallel_2 \leq c$ and let $\epsilon$ be a constant satisfying $\epsilon \leq 1/(2cd)$.

By Lemma 22, we have

$$\det (I + \epsilon A) = 1 + \epsilon \text{tr} (A) + \frac{\epsilon^2}{2} (\text{tr} (A)^2 - \text{tr} (A^2)) + \Delta$$

for some $|\Delta| \leq c^3 d^3$.

On the other hand, using Taylor expansion of $1/(1+x)$ about $x = 0$, we can verify that for any $a \in [-1/2, 1/2]$

$$|(1 + a)^{-1} - (1 - a + a^2)| \leq |2a|^3.$$  (54)
By our assumption on $\epsilon$, we have $\text{ctr}(A) + \frac{\epsilon^2}{2} (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta \in [-1/2, 1/2]$, therefore

\[
(\det(I + \epsilon A))^{-1} = (1 + \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)^{-1}
\]

\[
\leq 1 - \text{ctr}(A) - \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) - \Delta
+ (\text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)^2
+ 2 (\text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)^3
= 1 - \text{ctr}(A) - \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \epsilon^2 \text{tr}(A^2)
+ (\epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta) (\text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)
+ 2 (\text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)^3
\leq 1 - \text{ctr}(A) - \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \epsilon^2 \text{tr}(A^2) + 10 (\epsilon \text{cd})^3
= 1 - \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 + \text{tr}(A^2)) + 10 (\epsilon \text{cd})^3,
\]

where the first inequality is by (55), the first inequality is by moving terms around, the second inequality is by our assumption that $\|A\|_2 \leq c$ and the fact that $|\Delta| \leq \epsilon^3 d^3$ and by Lemma 11 and the last two lines are by collecting terms. Conversely, one can show that

\[
(1 + \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 - \text{tr}(A^2)) + \Delta)^{-1}
\geq 1 - \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 + \text{tr}(A^2)) - 10 (\epsilon \text{cd})^3.
\]

The proof is similar and is omitted.

Therefore

\[
|\det(I + \epsilon A)^{-1} - (1 - \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 + \text{tr}(A^2)))| \leq 10 (\epsilon \text{cd})^3. \tag{55}
\]

Now, we consider the case that $A := -\sqrt{\nabla^2 U(x)} + \sqrt{G_n(x)}$, $\epsilon := \sqrt{\delta}$ and $c := 2L^{1/2}$. Recall our assumption that $\delta \leq \frac{1}{2\pi^2 L^2}$. Combined with Assumption 11 and 12, we get

1. $\|A\|_2 \leq c$,
2. $\epsilon = \sqrt{\delta} \leq 1/\left(2^4 d L^{1/2}\right) \leq 1/(2 cd)$.

Using (55),

\[
\det \left(I - \sqrt{\delta} \left(\sqrt{\nabla^2 U(x)} - \sqrt{G_n(x)}\right)\right)^{-1}
=: \det(I + \epsilon A)^{-1}
\leq 1 - \text{ctr}(A) + \epsilon^2/2 (\text{tr}(A)^2 + \text{tr}(A^2)) + 10 (\epsilon \text{cd})^3
= 1 + \sqrt{\delta} \text{tr} \left(\sqrt{\nabla^2 U(x)} - \sqrt{G_n(x)}\right)
+ \delta/2 \left(\text{tr} \left(\sqrt{\nabla^2 U(x)} - \sqrt{G_n(x)}\right)^2 + \text{tr} \left(\sqrt{\nabla^2 U(x)} - \sqrt{G_n(x)}\right)^2\right)
+ 80 \delta^{3/2} d^2 L^{3/2}
= 1 + \delta \text{tr} \left(\nabla^2 U(x)\right) - \sqrt{\delta} \text{tr} \left(G_n(x)\right)
+ \delta \text{tr} \left(G_n(x)^2\right) - \delta \text{tr} \left((G_n(x))^2\right)
+ \delta^2/2 \text{tr} \left(\nabla^2 U(x)^2\right)^2 + 2 \delta^{3/2} \text{tr} \left(\nabla^2 U(x)\right) \text{tr} \left(G_n(x)\right)
+ \delta^2/2 \text{tr} \left(\nabla^2 U(x)^2\right) + 2 \delta^{3/2} \text{tr} \left(\nabla^2 U(x)G_n(x)\right)
+ 80 \delta^{3/2} d^2 L^{3/2}
\leq 1 + \delta \text{tr} \left(\nabla^2 U(x)\right) - \sqrt{\delta} \text{tr} \left(G_n(x)\right)
+ \delta \text{tr} \left(G_n(x)^2\right) + \delta \text{tr} \left((G_n(x))^2\right)
+ \delta^{3/2} d L^{3/2} + 2 \delta^{3/2} d^2 L^{3/2}
+ \delta^{3/2} L^{3/2} + 2 \delta^{3/2} d L^{3/2}
+ 80 \delta^{3/2} d^2 L^{3/2}
\leq 1 + \delta \text{tr} \left(\nabla^2 U(x)\right) - \sqrt{\delta} \text{tr} \left(G_n(x)\right) + \delta \text{tr} \left(G_n(x)^2\right) + \delta \text{tr} \left((G_n(x))^2\right) + 90 \delta^{3/2} d L^{3/2},
\]

where the first inequality is by (55), the second equality is by definition of $A$ and $\epsilon$, the third equality is by moving terms around, the second inequality is by Assumption 11 and 12, the third inequality is again by moving terms around.
Conversely, one can show that
\[
\det \left( I - \sqrt{3} \left( \sqrt{\delta} \nabla^2 U(x) - \sqrt{2} G_\eta(x) \right) \right)^{-1} \\
\geq_1 1 + \delta \text{tr} \left( \nabla^2 U(x) \right) - \sqrt{23} \text{tr} \left( G_\eta(x) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) - 9\delta^{3/2} d^3 L^{3/2}.
\]
The proof is similar and is omitted.

**Lemma 25** For any \( \delta \leq \frac{1}{2\sqrt{d} L} \), for any \( x \), and for \( \eta \) a.s.,
\[
\det \left( \nabla F_\eta \left( F_{\eta}^{-1}(x) \right) \right)^{-1} \\
= 1 - \sqrt{28} \text{tr} \left( G_\eta(x) \right) + 2\delta \text{tr} \left( \left( M_\eta(x), \nabla T_\eta(x) \right) \right) + \delta \text{tr} \left( \nabla^2 U(x) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \Delta
\]
for some \( |\Delta| \leq 128\delta^{3/2} d^3 L^{3/2} \left( \|x\|_2^2 + 1 \right) \).

**Proof of Lemma 25** Consider the Jacobian matrix inside the determinant. By definition of \( F_\eta \), we know that
\[
\nabla F_\eta \left( F_{\eta}^{-1}(x) \right) = I - \delta \nabla^2 U \left( F_{\eta}^{-1}(x) \right) + \sqrt{28} G_\eta \left( F_{\eta}^{-1}(x) \right).
\]
Thus,
\[
\det \left( \nabla F_\eta \left( F_{\eta}^{-1}(x) \right) \right)^{-1} \\
= \det \left( I - \sqrt{3} \left( \sqrt{\delta} \nabla^2 U \left( F_{\eta}^{-1}(x) \right) - \sqrt{2} G_\eta \left( F_{\eta}^{-1}(x) \right) \right) \right)^{-1} \\
\leq 1 - \sqrt{28} \text{tr} \left( G_\eta \left( F_{\eta}^{-1}(x) \right) \right) + \delta \text{tr} \left( \nabla^2 U \left( F_{\eta}^{-1}(x) \right) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \frac{80}{3}\delta^{3/2} d^3 L^{3/2} \\
\leq 1 - \sqrt{28} \text{tr} \left( G_\eta(x) \right) + 2\delta \text{tr} \left( \left( M_\eta(x), \nabla T_\eta(x) \right) \right) + \delta \text{tr} \left( \nabla^2 U(x) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \frac{80}{3}\delta^{3/2} d^3 L^{3/2} \\
\leq 1 - \sqrt{28} \text{tr} \left( G_\eta(x) \right) + 2\delta \text{tr} \left( \left( M_\eta(x), \nabla T_\eta(x) \right) \right) + \delta \text{tr} \left( \nabla^2 U(x) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \frac{80}{3}\delta^{3/2} d^3 L^{3/2} \left( \|x\|_2^2 + 1 \right),
\]
where the first inequality is by Lemma 24, the second inequality is by triangle inequality, the third inequality is by Lemma 23, the fourth inequality is by collecting terms.

Conversely, one can show that
\[
\det \left( \nabla F_\eta \left( F_{\eta}^{-1}(x) \right) \right)^{-1} \\
\geq 1 - \sqrt{28} \text{tr} \left( G_\eta(x) \right) + 2\delta \text{tr} \left( \left( M_\eta(x), \nabla T_\eta(x) \right) \right) + \delta \text{tr} \left( \nabla^2 U(x) \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \delta \text{tr} \left( \left( G_\eta(x) \right)^2 \right) + \frac{90}{3}\delta^{3/2} d^3 L^{3/2} \left( \|x\|_2^2 + 1 \right).
\]
The proof is similar and is omitted.

**Corollary 26** For any \( \delta \leq \frac{1}{2\sqrt{d} L} \), for any \( x \), and for \( \eta \) a.s.,
\[
\left| \det \left( \nabla F_\eta \left( F_{\eta}^{-1}(x) \right) \right)^{-1} - \left( 1 - \sqrt{28} \text{tr} \left( G_\eta(x) \right) \right) \right| \\
\leq 8\delta^2 L \left( \|x\|_2^2 + 1 \right).
\]
Proof of Corollary 26
Let
\[ \Delta := \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right)^{-1} \]
\[ - \left( 1 - \sqrt{2} \text{tr} (G_\eta(x)) + 2 \text{tr} \left( \{ M_\eta(x), T_\eta(x) \} \right) + \text{tr} (\nabla^2 U(x)) + \text{tr} (G_\eta(x))^2 + \text{tr} \left( (G_\eta(x))^2 \right) \right). \]
By Lemma 25
\[ |\Delta| \leq 128 \delta^{3/2} d^3 L^{3/2} (\|x\|^2 + 1). \]
Thus
\[ \left| \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right)^{-1} - 1 \right| \]
\[ \leq \sqrt{2} \delta d L^{1/2} \left| \Delta \right| \leq 128 \delta^{3/2} d^3 L^{3/2} (\|x\|^2 + 1) \]
\[ \leq 128 \delta^{3/2} d^3 L^{3/2} (\|x\|^2 + 1) + 2 \delta d L (\|x\|^2 + 1) + \delta d L + \delta d^2 L + \delta d \]
\[ \leq 6 \delta^2 L (\|x\|^2 + 1), \]
where the first line is by our definition of \( \Delta \), the second line is by our bound on \( |\Delta| \) above and by Assumptions 11 and 4, the third inequality is by moving terms around.

Corollary 27 For any \( \delta \leq \frac{1}{2 \sqrt{d} L} \), for any \( x \), and for \( \eta \) a.s.,
\[ \left| \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right)^{-1} - 1 \right| \]
\[ \leq 2^{5/2} d^{1/2} L^{1/2}. \]

Proof of Corollary 27 From Lemma 26 we get
\[ \left| \det \left( \nabla F_\eta (F_\eta^{-1}(x)) \right)^{-1} - 1 \right| \]
\[ \leq \sqrt{2} \delta d L^{1/2} \left| \Delta \right| \]
\[ \leq 2 \delta^2 L (\|x\|^2 + 1) \]
where the first inequality is by triangle inequality, and the second inequality is by Corollary 26.

Lemma 28 Under Assumption 4 for all \( x \),
\[ 1. \| \nabla p^\ast (x) \|_2 \leq p^\ast (x) \theta \left( \|x\|^2 + 1 \right) \]
\[ 2. \| \nabla^2 p^\ast (x) \|_2 \leq p^\ast (x) \left( \theta^2 + \theta \right) \left( \|x\|^2 + 1 \right) \]
\[ 3. \| \nabla^3 p^\ast (x) \|_2 \leq 2 p^\ast (x) \left( \theta^3 + \theta^2 + \theta \right) \left( \|x\|^6 + 1 \right) \]

Proof of Lemma 28 To prove the first claim:
\[ \| \nabla p^\ast (x) \|_2 = \| p^\ast (x) \nabla \log p^\ast (x) \|_2 \]
\[ \leq p^\ast (x) \theta \left( \|x\|^2 + 1 \right). \]
To prove the second claim:
\[ \| \nabla^2 p^\ast (x) \|_2 = p^\ast (x) \left( \| \nabla \log p^\ast (x) \|_2 + \| \nabla \log p^\ast (x) \|_2^2 \right) \]
\[ \leq p^\ast (x) \left( \theta^2 + \theta \right) \left( \|x\|^2 + 1 \right), \]
where the second and third inequalities are by Assumption 6.
To prove the third claim:
\[ \| \nabla^3 p^\ast (x) \|_2 \]
\[ = p^\ast (x) \left( \| \nabla^2 \log p^\ast (x) \|_2 + \| \nabla \log p^\ast (x) \|_2 \right) \]
\[ \leq 2 p^\ast (x) \left( \theta^3 + \theta^2 + \theta \right) \left( \|x\|^6 + 1 \right), \]
where \( \otimes \) denotes the tensor outer product.
Lemma 29 Under Assumption 4 for all $x, y$,

\[ p^*(y) \leq p^*(x) \cdot \exp \left( \theta \left( \|x\|_2^2 + \|y\|_2^2 \right) \|y - x\|_2 \right). \]

Proof Under Assumption 4

\[
\begin{aligned}
|\log p^*(y) - \log p^*(x)| & = \left| \int_0^1 \nabla \log p^*((1-t)x + ty), y - x \right| dt \\
& \leq \int_0^1 \|\nabla \log p^*((1-t)x + ty)\|_2 \cdot \|y - x\|_2 dt \\
& \leq \int_0^1 \theta \left( \|1-t)x + ty\|_2^2 + 1 \right) \|y - x\|_2 dt \\
& \leq 2\theta \left( \|x\|_2^2 + \|y\|_2^2 \right) \|y - x\|_2.
\end{aligned}
\]

Therefore,

\[
\exp \left( -\theta \left( \|x\|_2^2 + \|y\|_2^2 \right) \|y - x\|_2 \right) \leq \frac{p^*(y)}{p^*(x)} \leq \exp \left( \theta \left( \|x\|_2^2 + \|y\|_2^2 \right) \|y - x\|_2 \right).
\]

Lemma 30 The stationary distribution $p^*$ of (5) satisfies the equality (for all $x$)

\[
0 = p^*(x) \left( \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma \sigma^T \right]_{i,j} + \text{tr} \left( \nabla^2 U(x) \right) \right) + \sum_{i=1}^d \frac{\partial}{\partial x_i} p^*(x) \cdot \frac{\partial}{\partial x_i} U(x) + 2 \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} p^*(x) \left( \sum_{j=1}^d \frac{\partial}{\partial x_j} \left[ \sigma \sigma^T \right]_{i,j} \right) \right)
\]

Proof of Lemma 30 For a distribution $p_t$, the Fokker Planck equation under (5) is

\[
\frac{d}{dt} p_t(x) = \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} (\nabla U(x))_i \cdot p_t(x) \right) + \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma \sigma^T \right]_{i,j} \cdot p_t(x)
\]

\[
= \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} (\nabla U(x))_i \cdot p_t(x) \right) + \sum_{i=1}^d \sum_{j=1}^d \left( \frac{\partial}{\partial x_j} \left[ \sigma \sigma^T \right]_{i,j} \cdot p_t(x) \right)
\]

\[
+ 2 \sum_{i=1}^d \sum_{j=1}^d \left( \frac{\partial}{\partial x_j} \left[ \sigma \sigma^T \right]_{i,j} \cdot \frac{\partial}{\partial x_i} p_t(x) \right) + \text{tr} \left( \nabla^2 U(x) \right)
\]

\[
= p_t(x) \left( \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma \sigma^T \right]_{i,j} + \text{tr} \left( \nabla^2 U(x) \right) \right)
\]

\[
+ \sum_{i=1}^d \frac{\partial}{\partial x_i} p_t(x) \cdot \frac{\partial}{\partial x_i} U(x) + 2 \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} p_t(x) \left( \sum_{j=1}^d \frac{\partial}{\partial x_j} \left[ \sigma \sigma^T \right]_{i,j} \right) \right).
\]
Observe that by definition of $p^*$ being the stationary distribution of (53), \( \frac{d}{dt} p_t(x) \big|_{p_t=p^*} = 0 \). Thus, we have

\[
p^*(x) \left( \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2}{\partial x_i \partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} + \text{tr} \left( \nabla^2 U(x) \right) \right)
\]
\[
+ \sum_{i=1}^d \frac{\partial}{\partial x_i} p^*(x) \cdot \frac{\partial}{\partial x_i} U(x)
\]
\[
+ \sum_{i=1}^d \left( \frac{\partial}{\partial x_i} p^*(x) \left( \sum_{j=1}^d \frac{\partial}{\partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} \right) \right)
\]
\[
+ \left( \nabla^2 p^*(x), \sigma_x \sigma_x^T \right) = 0. \]

\[\square\]

**Lemma 31** For any $\delta$ satisfying

\[
\frac{1}{\delta} \geq \max \left\{ \frac{2^8 \delta^2 L}{2^3 L \theta^2}, \frac{2^7 \delta R^2 \left( \frac{c_2}{m} \log \frac{c_2}{m} \right)^3}{2^7 \delta \log \left( \frac{c_2 R^2}{m} \right)} \right\}.
\]

and for

\[ R := 2^7 \sqrt{\max \left\{ \frac{c_2}{m} \log \frac{c_2}{m}, \frac{1}{2^{124} d^6 L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \delta^3} \right\}}, \]

$\delta$ and $R$ satisfy

\[
\delta \leq \min \left\{ \frac{1}{2^8 \delta^2 L \theta^2 \left( R^6 + 1 \right)}, \frac{1}{2^{15} L \theta^2 \left( R^6 + 1 \right)} \right\}.
\]

**Proof of Lemma 31** Our first assumption in (60) immediately implies that $\delta \leq (2^8 d^2 L)^{-1}$, so we only need to verify that

\[
\delta \leq \frac{1}{2^{15} L \theta^2 \left( R^6 + 1 \right)}.
\]

Since $R$ is a max of three terms, we will consider 2 cases:

**Case 1:** $R = 2^7 \sqrt{\max \left\{ \frac{c_2}{m} \log \frac{c_2}{m}, 1 \right\}}$

In this case, (61) follows immediately from our second and third assumption in (60).

**Case 2:** $R = 2^7 \sqrt{\frac{c_2}{m} \log \left( \frac{1}{2^{124} d^6 L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \delta^3} \right)}$

Recall that we would like to prove that

\[
\delta \leq \left( 2^{15} L \theta^2 \left( R^6 + 1 \right) \right)^{-1}
\]

Since $R^6 + 1 \leq \max \left\{ 2 R^6, 2 \right\}$, it suffices to prove that

\[
\frac{1}{\delta} \geq 2^{16} L \theta^2
\]

and

\[
\frac{1}{\delta} \geq 2^{16} L \theta^2 R^6.
\]

The first inequality follows immediately from our second assumption in (60). The second inequality expands to be

\[
\frac{1}{\delta} \geq 2^{15} L \theta^2 \frac{c_2}{m^3} \left( \log \left( \frac{1}{2^{124} d^6 L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \delta^3} \right) \right)^3.
\]

Moving terms around, we see that it is sufficient to prove

\[
\delta^{-1/3} \geq \left( 2^{20} L^{1/3} \theta^{2/3} \frac{c_2}{m} \right) \log \left( \delta^{-1/3} \left( 2^{124} d^6 L^2 \left( \theta^3 + \theta^2 + \theta \right)^2 \right)^{-1/9} \right). \tag{62}
\]

31
We define \( a := \left(2^{24} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)\right)^{-1/9}, \ c := \left(2^{20} L^{1/3} \theta^{2/3} \frac{e^2}{m}\right)^{-1} \) and \( \delta := \delta^{-1/3} \). We verify that \( a \) and \( c \) are both strictly positive quantities. By the fourth assumption in (60),

\[
\delta^{-1/3} \geq 2^{24} L^{1/3} \theta^{2/3} \frac{e^2}{m} \log \left(2^{62} c^2 \frac{e^2}{m}\right)
\]

\[
\geq 3 \cdot 2^{20} L^{1/3} \theta^{2/3} \frac{e^2}{m} \cdot \log \left(\frac{2^{20} L^{1/3} \theta^{2/3} \frac{e^2}{m}}{(2^{124} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2) \delta^3}\right)^{1/7}
\]

\[
= 3 \cdot \frac{1}{c} \log \frac{a}{c}
\]

We can thus apply Corollary 44 with the given \( a, c, x \), to prove (62) \( (\delta^{1/3} > 0 \) is guaranteed by the first assumption of (60).

We have concluded the proof of Case 2, and hence (61).

\[\Box\]

**Lemma 32** For any \( \epsilon > 0 \), and for any stepsize \( \delta \) satisfying

\[
\frac{1}{\delta} \geq \frac{d^2}{\epsilon^2} \cdot \frac{2142 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2}{\lambda^2} \cdot \max \left\{ \left(\frac{c^2}{m} \log \frac{e^2}{m}\right)^{12}, 1\right\}
\]

then

\[
2^{70} \delta^{1/2} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \max \left\{ \frac{c^2}{m} \log \frac{e^2}{m}, \frac{c^2}{m} \log \left(\frac{1}{2^{124} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2} \delta^3\right)ight\} \frac{6}{\lambda^1} \leq \frac{\epsilon}{2},
\]

**Proof of Lemma 32** By the first two cases in the max in (63) and moving terms around, one can immediately verify that

\[
2^{70} \delta^{1/2} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \max \left\{ \frac{c^2}{m} \log \frac{e^2}{m}, 1\right\} \frac{6}{\lambda^1} \leq \frac{\epsilon}{2}.
\]

Thus we only need to prove that

\[
2^{70} \delta^{1/2} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \left(\frac{c^2}{m} \log \left(\frac{1}{2^{124} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2} \delta^3\right)\right)^{1/6} \lambda^{-1} \leq \frac{\epsilon}{2}.
\]

The above is equivalent to

\[
\delta^{-1/12} \geq \log \left(\frac{\left(2^{24} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2\right)^{-1/36} \delta^{-1/12}}{\left(2^{124} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2 \delta^3\right)^{1/6}}\right) \cdot \left(2^{71} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \frac{c^2}{m} \lambda^1 \lambda^{-1} \epsilon^{-1}\right)^{1/6}.
\]

Let us define

\[
a := \left(2^{24} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2\right)^{-1/36},
\]

\[
c := \left(2^{71} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \frac{c^2}{m} \lambda^1 \lambda^{-1} \epsilon^{-1}\right)^{-1/6},
\]

\[
x := \delta^{-1/12}.
\]

Then by the third case in our max in (63),

\[
\delta^{-1/12} \geq \left(2^{71} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \frac{c^2}{m} \lambda^1 \lambda^{-1} \epsilon^{-1}\right)^{-1/6} \cdot \log \left(2^{24} d^6 L \left(\theta^3 + \theta^2 + \theta\right) \lambda^{-1} \epsilon^{-6}\right) \cdot \left(2^{71} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \frac{c^2}{m} \lambda^1 \lambda^{-1} \epsilon^{-1}\right)^{-1/6} \cdot \log \left(\frac{2^{71} d^{7/2} L \left(\theta^3 + \theta^2 + \theta\right) \frac{c^2}{m} \lambda^1 \lambda^{-1} \epsilon^{-1}}{(2^{124} d^6 L^2 \left(\theta^3 + \theta^2 + \theta\right)^2 \delta^3)^{1/36}}\right)
\]

\[
\geq 3 \cdot \frac{1}{c} \log \frac{a}{c}.
\]

Thus (61) follows immediately from Corollary 44 with the \( a, c, x \) as defined above.

\[\Box\]

**Lemma 33** For any \( \delta \leq \frac{1}{27} \) and for \( x_0 \) with dynamics defined in (33). If \( U_i(x) \) is \( m \) strongly convex and has \( L \) lipschitz gradients for all \( i \in \{1...S\} \), then Assumption 3 holds with \( \lambda = m \), i.e. for any two distributions \( p \) and \( q \),

\[
W_2(F_\delta(p), F_\delta(q)) \leq e^{-m \delta} W_2(p, q).
\]
Proof of Lemma 33: Let $\gamma^*$ be an optimal coupling between $p$ and $q$, i.e.

$$W_2^2(p, q) = \mathbb{E}_{\gamma^*}(x, y) \left[ \|x - y\|_2^2 \right]$$

We define a coupling $\gamma'$ as follows:

$$\gamma'(x, y) := (F_{\eta}, F_{\eta})_{\#} \gamma^*$$

Where $\#$ denotes the push-forward operator. (See (6) for the definition of $F_{\eta}$.) It is thus true by definition that $\gamma'$ is a valid coupling between $\Phi_{\delta}(p)$ and $\Phi_{\delta}(q)$.

Thus

$$W_2(\Phi_{\delta}(p), \Phi_{\delta}(q)) \leq \mathbb{E}_{\gamma'(x, y)} \left[ \|x - y\|_2^2 \right]$$

$$= \mathbb{E}_{\gamma^*}(x, y) \left[ \|F_{\eta}(x) - F_{\eta}(y)\|_2^2 \right]$$

$$= \mathbb{E}_{\gamma^*}(x, y) \left[ \|x - \delta \nabla U(x) + \sqrt{2\delta} T_{\eta}(x) - (y - \delta \nabla U(y) + \sqrt{2\delta} T_{\eta}(y))\|_2^2 \right]$$

$$= \mathbb{E}_{\gamma^*}(x, y) \left[ \|x - \delta \nabla U(x) - (y - \nabla U(y))\|_2^2 \right]$$

$$\leq \mathbb{E}_{\gamma^*}(x, y) \left[ (1 - m\delta/2) \|x - y\|_2^2 \right]$$

$$\leq e^{-m\delta/4} \mathbb{E}_{\gamma^*}(x, y) \left[ \|x - y\|_2^2 \right]$$

$$= e^{-m\delta/4} W_2^2(p, q)$$

Where the second inequality follows from our assumption that $U_i(x)$ is $m$ strongly convex and has $L$ lipschitz gradients, and our assumption that $\delta \leq \frac{1}{2L}$, and the third inequality is by the fact that $m\delta/2 \leq m/(2L) \leq 1/2$. ■
C Subgaussian Bounds

Lemma 34 Let $p^*$ be the invariant distribution to $[3]$. Under the assumptions of Section $4.3$, $p^*$ satisfies

$$E_{p^*}(x) \left[ \exp \left( \frac{m}{8c_s^2} \| x \|^2 \right) \right] \leq 8d$$

Proof Let $p_0$ be an initial distribution for which the above expectation is finite. Let $x_t$ be as defined in $[5]$ (we use $x_t$ instead of $x(t)$ to reduce clutter). For convenience of notation, let $s := \frac{m}{8c_s^2}$.

$$\frac{d}{dt} E \left[ \exp \left( s \| x_t \|^2 \right) \right]$$

$$= E \left[ \exp \left( s \| x_t \|^2 \right) \cdot \left( -\nabla U(x_t), 2sx_t \right) + \left( 2sl + 4s^2x_t, 2\sigma_s \right) \right]$$

$$\leq E \left[ \exp \left( s \| x_t \|^2 \right) \cdot \left( -2ms \| x_t \|^2 + 4sc_s^2 + 8s^2c_s^2 \| x_t \|^2 \right) \right]$$

$$E \left[ \exp \left( s \| x_t \|^2 \right) \cdot \left( -ms \| x_t \|^2 + 4sc_s^2 \right) \right]$$

$$\leq 4sc_s^2 E \left[ \exp \left( s \| x_t \|^2 \right) \cdot 1 \left\{ \| x_t \|^2 \geq \frac{8c_s^2}{m} \right\} \right] + 4dsc_s e$$

$$\leq 4sc_s^2 E \left[ \exp \left( s \| x_t \|^2 \right) \right] + 8dsc_s e,$$

where the first line is by Ito’s lemma, the second line is by Assumption $1$ and Assumption $2.2$, the third line is by definition of $s$, the fifth line is again by definition of $s$.

Since $p_t \rightarrow p^*$, the above implies that

$$E_{p^*} \left[ \exp \left( s \| x_t \|^2 \right) \right] < \infty$$

Furthermore, by invariance of $p^*$ under $[3]$, we have that if $p_0 = p^*$ then $\frac{d}{dt} E \left[ \exp \left( s \| x_t \|^2 \right) \right] = 0$, so

$$0 = \frac{d}{dt} E \left[ \exp \left( s \| x_t \|^2 \right) \right] \leq -4sc_s^2 E \left[ \exp \left( s \| x_t \|^2 \right) \right] + 8dsc_s e$$

$$\Rightarrow 4sc_s^2 E \left[ \exp \left( s \| x_t \|^2 \right) \right] \leq 8dsc_s e$$

$$\Rightarrow E_{p^*}(x) \left[ \exp \left( \frac{m}{8c_s^2} \| x \|^2 \right) \right] \leq 8d$$

(65)

Lemma 35 Let $p^*$ be the invariant distribution to $[3]$. Under the assumptions of Section $4.3$, $p^*$ satisfies

$$p^* \left( \| x \|^2 \geq t \right) \leq 8 \exp \left( -\frac{mt}{8c_s^2} \right),$$

where $m$ and $c_s$ are as defined in Section $[3]$

Proof of Lemma 35 From Lemma 34

$$E \left[ \exp \left( \frac{m}{8c_s^2} \| x \|^2 \right) \right] \leq 8d$$

By Markov’s inequality:

$$P \left( \| x \|^2 \geq t \right) = P \left( \exp \left( \frac{m}{8c_s^2} \| x \|^2 \right) \geq \exp \left( \frac{m}{8c_s^2} \right) \right)$$

$$\leq \frac{E \left[ \exp \left( \frac{m}{8c_s^2} \| x \|^2 \right) \right]}{\exp \left( \frac{m}{8c_s^2} \right)}$$

$$\leq 8d \exp \left( -\frac{mt}{8c_s^2} \right)$$

As a Corollary to Lemma 35, we can bound $E \left[ \| x \|^2 1\left\{ \| x \|^2 \geq t \right\} \right]$ for all $t$:

Corollary 36 Let $p^*$ be the invariant distribution to $[3]$. Under the assumptions of Section $4.3$, for any $S \geq \frac{4sc_s^2}{m} \text{max} \left\{ \log \left( \frac{4c_s^2}{m} \right), 1 \right\}$, $p^*$ satisfies

$$E_{p^*} \left[ \| x \|^2 1\left\{ \| x \|^2 \geq S \right\} \right] \leq 12d \exp \left( \frac{mS}{16c_s^2} \right)$$
Proof of Corollary 36 Let $y$ be a real valued random variable that is always positive. We use the equality

$$E[y] = \int_0^\infty P(y \geq s) ds$$

Let $y := \|x\|_2^2 \cdot 1 \{\|x\|_2^2 \geq t\}$. Then

$$P(y \geq s) = \begin{cases} 
1 & \text{if } s = 0 \\
1 - \Phi_3((\|x\|_2^2 - t)/\delta) & \text{if } s \in (0, t) \\
0 & \text{if } s \geq t 
\end{cases}$$

Therefore,

$$E_p\left[\|x\|_2^2 \cdot 1 \{\|x\|_2^2 \geq S\}\right] = E[y] = \int_0^\infty P(y \geq s) ds = \int_0^t P(\|x\|_2^2 \geq s) ds + \int_t^\infty P(\|x\|_2^2 \geq s) ds$$

$$= 8dS \exp\left(-\frac{mS}{8\sigma^2}\right) + \int_S^\infty 8d \exp\left(-\frac{ms}{8\sigma^2}\right) ds$$

$$\leq 8d \exp\left(-\frac{mS}{8\sigma^2}\right) + \int_S^\infty \frac{64dE^2}{m} \exp\left(-\frac{mS}{8\sigma^2}\right)$$

$$= \left(8d + \frac{64dE^2}{m}\right) \exp\left(-\frac{mS}{8\sigma^2}\right)$$

$$\leq 12d \exp\left(-\frac{mS}{16\sigma^2}\right),$$

where the first inequality above uses Lemma 35, the second inequality uses our assumption on $S$, and the third inequality is by our assumption on $S$ combined with Lemma 35. \[\square\]

**Corollary 37** Let $p_s := \Phi_3(p^*)$, then for all $t \geq 1$ and $\delta \leq \frac{1}{16L}$

1. $p_s(\|x\|_2^2 \geq t) \leq 8d \exp\left(-\frac{mt}{32\sigma^2}\right)$
2. $E_{p_s}[\|x\|_2^2 \cdot 1 \{\|x\|_2^2 \geq t\}] \leq 12d \exp\left(-\frac{mt}{16\sigma^2}\right)$

**Proof of Corollary 37** By Lemma 22 and our assumption that $\delta \leq 1/(16L)$ and Triangle inequality, we get

$$\|F_n^{-1}(x)_2\|_2 \geq \|x\|_2 - 2\delta^{1/2}L^{1/2}\|x\|_2 + 1) \geq 1/2\|x\|_2 - 1/8$$

Thus for $t \geq 1$ and $\delta \leq \frac{1}{16L}$

$$\|x\|_2 \geq \sqrt{t} \Rightarrow \|F_n^{-1}(x)_2\|_2 \geq 1/4\|x\|_2 - 1/8 \geq 1/4\|x\|_2 \geq \sqrt{t}/2$$

Thus

$$p_s(\|x\|_2^2 \geq t) \leq \Phi_3((t^2/4) \leq \Phi_3((t/4)) \leq 8d \exp\left(-\frac{mt}{32\sigma^2}\right)$$

This proves the first claim.

Using the first claim, and an identical proof as Corollary 36 we can prove the second claim. \[\square\]

**Lemma 38** For any $k$, we have the bound

$$E_{p^*}[\|x\|_2^{2k}] \leq \max\left\{\left(2^k(k - 1)\frac{c^2}{m} \log\left(\frac{16(k - 1)\sigma^2}{m}\right)\right)^{k-1}, 128kd\frac{c^2}{m}\right\}$$
Proof of Lemma 39 Let us define the fixed radius $S := \max \left\{ \frac{48(k-1)c^2}{m} \log \left( \frac{16(k-1)c^2}{m} \right), 0 \right\}$

$$
\mathbb{E}_{p^*} \left[ \|x\|_2^2 \right] = \int_0^\infty p^*(\|x\|_2^2 \geq t) \, dt
$$

where the first inequality is by Lemma 35 and the second inequality is by Lemma 43 and our choice of $S$.

**Proof of Lemma 39**

For any two densities $p, q$ over $\mathbb{R}^d$, let $c = \max \{ p(\|x\|_2 > R), q(\|x\|_2 > R) \}$, then

$$W_2^2(p, q) \leq 4R^2 \int_{\mathcal{B}_R} \left( \frac{p(x)}{q(x)} - 1 \right)^2 \, dx + 32c^2R^2 + 2cR + 2\mathbb{E}_p \left[ \|x\|_2^2 \mathbb{1} \{ \|x\|_2 > R \} \right] + 2\mathbb{E}_q \left[ \|x\|_2^2 \mathbb{1} \{ \|x\|_2 > R \} \right]
$$

**Proof of Lemma 39**

Let $p$ and $q$ be two distributions.

Let $a := p(\|x\|_2 > R)$ and $b := q(\|x\|_2 > R)$, let $c := \max \{ a, b \}$. To simplify the proof, assume that $a \leq b$. The proof for the case $b \leq a$ is almost identical and omitted.

For a radius $R$, let

$$p_R(x) := \frac{1}{1-a} \cdot \mathbb{1} \{ \|x\|_2 \leq R \} \cdot p(x)$$

$$q_R(x) := \frac{1}{1-b} \cdot \mathbb{1} \{ \|x\|_2 \leq R \} \cdot q(x)
$$

I.e. $p$ and $q$ conditioned on $\|x\|_2 \leq R$.

(The proof for when $b \leq a$ is almost identical and is omitted)

We will also define

$$p_R^*(x) := \frac{1}{b} \left( \frac{b-a}{1-a} \cdot \mathbb{1} \{ \|x\|_2 \leq R \} \cdot p(x) \right) + \frac{1}{b} (\mathbb{1} \{ \|x\|_2 > R \} \cdot p(x))$$

$$q_R^*(x) := \frac{1}{b} \mathbb{1} \{ \|x\|_2^2 > R \} \cdot q(x)
$$

One can verify that

$$p(x) = (1-b) \cdot p_R(x) + b \cdot p_R^*(x)$$

$$q(x) = (1-b) \cdot q_R(x) + b \cdot q_R^*(x)
$$

Suppose that we have a coupling $\gamma_R$ between $p_R$ and $q_R$ (i.e. $\gamma_R$ is a density over $\mathbb{R}^{2d}$). Then one can verify that $(1-b)\gamma_R + b\gamma_R^*$ is a valid coupling for $p$ and $q$.

Thus

$$W_2^2(p, q) \leq \mathbb{E}_{(x,y)} \cdot \mathbb{1} \{ x \neq y \} \cdot (1-b)\gamma_R + b\gamma_R^* \left[ \|x-y\|^2 \right]$$

$$= (1-b) \cdot \mathbb{E}_{(x,y)} \cdot \mathbb{1} \{ x \neq y \} \cdot \mathbb{E}_{(x,y)} \cdot \gamma_R \left[ \|x-y\|^2 \right] + b \cdot \mathbb{E}_{(x,y)} \cdot \gamma_R^* \left[ \|x-y\|^2 \right]$$

$$\leq (1-b) \cdot \mathbb{E}_{(x,y)} \cdot \gamma_R \left[ \|x-y\|^2 \right] + b \cdot \left( 2\mathbb{E}_{p_R^*} \left[ \|x\|^2 \right] + 2\mathbb{E}_{q_R^*} \left[ \|y\|^2 \right] \right)
$$

Since the above holds for all valid $\gamma_R$, it holds for the optimal $\gamma_R^*$, thus

$$W_2^2(p, q) \leq (1-b) \cdot W_2^2(p_R, q_R) + 2b \cdot \left( \mathbb{E}_{p_R^*} \left[ \|x\|^2 \right] + \mathbb{E}_{q_R^*} \left[ \|y\|^2 \right] \right)
$$
Since $p_R$ and $q_R$ are constrained to the ball of radius $R$, we can upper bound $W_2$ by $TV$:

$$W_2^2(p_R, q_R) \leq TV(p_R, q_R)^2 R^2$$

$$\leq \mathcal{K} \mathcal{L}(\|p\|_R, q_R) R^2$$

$$\leq \chi^2(p_R, q_R) R^2$$

We can upper bound $\chi^2(p_R, q_R)$ as

$$\chi^2(p_R, q_R) := \int q_R(x) \left( \frac{p_R(x)}{q_R(x)} - 1 \right)^2 dx$$

$$= \int_{B_R} \frac{1}{1 - b^2} q(x) \left( \frac{1 - b}{1 - a} \frac{p(x)}{q(x)} - 1 \right)^2 dx$$

$$\leq (1 + 2c) \int_{B \setminus R} q(x) \left( \frac{1 + 4c}{q(x)} - 1 \right)^2 dx$$

$$= (1 + 2c) \int_{B_R} q(x) \left( \frac{1 + 4c}{q(x)} - 1 \right)^2 dx + 64c^2,$$
Let \( p := p^* \) and \( q := p_\delta \), by the above results, we have
\[
\max \{ p(\|x\|_2 \geq R), q(\|x\|_2 \geq R) \} \leq 8 d \exp \left( \frac{mR^2}{32c^2} \right)
\]
(note that \( c_\sigma \) is defined in Assumption 2.2 and is unrelated to the \( c \) we defined in this proof).

Therefore, we apply Lemma 39 to get
\[
W_2^2(p_\delta, p^*) \leq 4R^2 \int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x) dx + 32R^2 \exp \left( -\frac{mR^2}{32c^2} \right) + 4R^2 \exp \left( -\frac{mR^2}{64c^2} \right) + 48d \exp \left( -\frac{mR^2}{64c^2} \right)
\]
\[
+ 2E_{p_\delta} \{ ||x||_2^2 \{ ||x||_2 > R \} \} + 2E_{p^*} \{ ||x||_2^2 \{ ||x||_2 > R \} \}
\]
\[
\leq 4R^2 \int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x) dx + 32R^2 \exp \left( -\frac{mR^2}{32c^2} \right) + 4R^2 \exp \left( -\frac{mR^2}{64c^2} \right) + 48d \exp \left( -\frac{mR^2}{64c^2} \right)
\]
\[
\leq 4R^2 \int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x) dx + 36 \exp \left( -\frac{mR^2}{128c^2} \right) + 48d \exp \left( -\frac{mR^2}{64c^2} \right)
\]
\[
\leq 4R^2 \int_{B_R} \left( \frac{p_\delta(x)}{p^*(x)} - 1 \right)^2 p^*(x) dx + 84d \exp \left( -\frac{mR^2}{64c^2} \right),
\]
where the third inequality is by Lemma 43 and our assumption that
\[
R^2 \geq \max \left\{ \frac{2^{13} c^2}{m} \left( \log \left( \frac{2^{11} c^2}{m} \right) \right), 0 \right\}
\]
D Miscellaneous Lemmas

Lemma 41 For any matrix $A \in \mathbb{R}^{2d}$,

$$\text{tr} A \leq d \|A\|_2$$

Proof of Lemma 41 For any matrices $A \in \mathbb{R}^{2d}$ and $B \in \mathbb{R}^{2d}$, we use the fact that

$$\langle A, B \rangle_F := \text{tr} (AB^T)$$

is an inner product.

Let $A = UDV$ where $U$ and $V$ are two orthonormal matrices and $D$ is a diagonal of positive singular values. Let $\lambda := \max_i D_{ii}$. It is known that $\lambda = \|A\|_2$.

Then

$$\text{tr} (A) = \text{tr} (UDV) = \text{tr} (DVU)$$

$$= \sqrt{\langle D, (VU)^T \rangle_F} \sqrt{\langle (VU)^T, VU \rangle_F}$$

$$\leq \sqrt{\|D\|^2 \|U^T V^T VU\|}$$

$$= d\lambda$$

$$= d \|A\|_2,$$

where the first inequality is by Cauchy Schwierz, and the second inequality uses the fact that $U^T V^T VU = I$.

Lemma 42 Let $A \in \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a symmetric matrix such that $\|A\|_2 \leq c$. Let $\epsilon \leq \frac{1}{2cd}$ then

$$\left| \det (I + \epsilon A) - \left( 1 + \epsilon \text{tr} (A) + \frac{\epsilon^2}{2} (\text{tr} (A^2) - \text{tr} (A^2)) \right) \right| \leq \epsilon^3 c^3 d^3$$

Proof of Lemma 42 Let the eigenvalue decomposition of $A$ be $A = UDU^T$, where $U$ is orthogonal, and $D$ is the diagonal matrix of $A$’s eigenvalues. Let $\lambda_i := D_{ii}$, and let $D$ be chosen such that

$$|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_d|$$

It is known that $|\lambda_1| = \|A\|_2 \leq c$.

The matrix $I + \epsilon A$ can be written as

$$U (I + \epsilon D) U^T$$

Since the determinant of products is the product of determinants,

$$\det (I + \epsilon A) = \det (I + \epsilon D) \cdot \det (U) \det (U)$$

$$= \det (I + \epsilon D) \cdot \det (U^T U)$$

$$= \det (I + \epsilon D)$$

$$= \prod_{i=1}^{d} (1 + \epsilon \lambda_i)$$

$$= 1 + \epsilon \sum_{i=1}^{d} \lambda_i + \frac{\epsilon^2}{2} \sum_{i=1}^{d} \sum_{j \neq i} \lambda_i \lambda_j + ...$$

Thus

$$\left| \det (I + \epsilon A) - \left( 1 + \epsilon \sum_{i=1}^{d} \lambda_i + \frac{\epsilon^2}{2} \sum_{i=1}^{d} \sum_{j \neq i} \lambda_i \lambda_j \right) \right| \leq \epsilon^3 c^3 d^3$$

(67)

where the last inequality is by the assumption that $\epsilon \leq \frac{1}{2cd}$.

It can be verified that

$$\epsilon^3 c^3 d^3.$$
1. \( \text{tr} (A) = \sum_{i=1}^{d} \lambda_i \)
2. \( \text{tr} (A^2) = \sum_{i=1}^{d} \lambda_i^2 \)
3. \( \text{tr} (A)^2 = \left( \sum_{i=1}^{d} \lambda_i \right)^2 = \sum_{i=1}^{d} \lambda_i^2 + \sum_{i \neq j} \lambda_i \lambda_j \)

Thus, we can rewrite (67) as

\[
\left| \det (I + \epsilon A) - \left( 1 + \epsilon \text{tr} (A) + \frac{\epsilon^2}{2} (\text{tr} (A)^2 - \text{tr} (A^2)) \right) \right| \leq \epsilon^3 c^3 d^3
\]

Lemma 43  For any \( c > 0, \ x > 3 \max \left\{ \frac{1}{c} \log \frac{1}{c}, 0 \right\} \), the inequality

\[
\frac{1}{c} \log(x) \leq x
\]

holds.

Proof  We will consider two cases:

Case 1: If \( c \geq \frac{1}{3} \), then the inequality

\[
\log(x) \leq cx
\]

is true for all \( x \).

Case 2: \( c < \frac{1}{3} \).

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):

\[
\forall u > 0, \ W_{-1}(-e^{-u-1}) > -u - \sqrt{2u - 1}
\]

equivalently \( \forall c \in (0,1/e), \ W_{-1}(-c) < -c + \frac{3}{2} \log \left( \frac{1}{c} \right) - \frac{1}{2} \)

\[
= \log \left( \frac{1}{c} \right) + \frac{3}{2} \log \left( \frac{1}{c} \right) - 1
\]

\[
\leq 3 \log \left( \frac{1}{c} \right)
\]

Thus by our assumption,

\[
x \geq 3 \cdot \frac{1}{c} \log \left( \frac{1}{c} \right)
\]

\[
\Rightarrow x \geq \frac{1}{c} (-W_{-1}(-c))
\]

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

\[
x \geq \frac{1}{c} \max \left\{ -W_{-1}(-c), 1 \right\}
\]

\[
\Rightarrow (-cx) e^{-cx} \geq -c
\]

\[
x e^{-cx} \leq 1
\]

\[
\Rightarrow \log(x) \leq cx
\]

The first implication is justified as follows: \( W_{-1}^{-1}: \left[ -\frac{1}{e}, \infty \right) \to (-\infty, -1) \) 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. ■

Corollary 44  For any \( a > 0 \), and for any \( c > 0, \ x > 3 \max \left\{ \frac{1}{a} \log \frac{1}{a}, 0 \right\} \), the inequality

\[
\frac{1}{c} \log(a \cdot x) \leq x
\]

holds.

Proof of Corollary 44  Let \( c := \frac{1}{a} \). Then for any \( x' > 3 \max \left\{ \frac{1}{c} \log \frac{1}{c}, 0 \right\} \), Lemma 43 gives

\[
\log(x') \leq c \cdot x' - \frac{c}{a} x'
\]

Thus with a change of variables \( x' = ax \), we get that for any \( x > \frac{3}{a} \max \left\{ \frac{1}{c} \log \frac{1}{c}, 0 \right\} = 3 \max \left( \frac{1}{a} \log \frac{1}{a}, 0 \right) \),

\[
\log(ax) \leq cx
\]

■
Lemma 45

\[
\sum_{j=1}^{d} \frac{\partial}{\partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} = \left[ \mathbb{E}_{q(\eta)} \left[ G_\eta(x) T_\eta(x) + \text{tr}(G_\eta(x)) T_\eta(x) \right] \right]_{i,j}
\]

Proof of Lemma 45

\[
\sum_{j=1}^{d} \frac{\partial}{\partial x_j} \left[ \sigma_x \sigma_x^T \right]_{i,j} = \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \mathbb{E}_{q(\eta)} \left[ T_\eta(x) T_\eta(x)^T \right]_{i,j}
\]

\[
= \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \mathbb{E}_{q(\eta)} \left[ [T_\eta(x)]_{i,j} [T_\eta(x)]_{j,i} \right]
\]

\[
= \sum_{j=1}^{d} \mathbb{E}_{q(\eta)} \left[ G_\eta(x)_{i,j} [T_\eta(x)]_{j,i} + [T_\eta(x)]_{i,j} G_\eta(x)_{j,i} \right]
\]

\[
= \left[ \mathbb{E}_{q(\eta)} \left[ G_\eta(x) T_\eta(x) + \text{tr}(G_\eta(x)) T_\eta(x) \right] \right]_{i,j}
\]

Lemma 46 Let \( \delta \leq \frac{1}{8L} \), then the function \( F_\eta(y) \) as defined in (\ref{eq:50}) is invertible for all \( y \) and for \( \eta \) a.s.

Proof of Lemma 46 To prove the invertibility of \( F_\eta(x) \), we only need to show that the Jacobian of \( F_\eta(x) \) is invertible. The Jacobian of \( F_\eta(x) \) is

\[
I - \delta \nabla^2 U(x) + \sqrt{2\delta} G_\eta(x) \succ (1 - \delta L - \sqrt{2\delta L}) I \succ \frac{1}{2} I
\]

Where we used Assumption 1 and Assumption 2. The existence of \( F_\eta^{-1} \) thus follows immediately from Inverse Function Theorem.

\[\square\]
E Relation to Classical CLT

Lemma 47 Let $\eta_1, \ldots, \eta_k$ be iid random variables such that $\mathbb{E} [\eta_i] = 0$, $\mathbb{E} [\eta_i \eta_i^T] = I$, and $\|\eta_i\|_2$ is a.s. bounded by some constant. Let $\delta_k := \sqrt{\frac{k-1}{k}} \leq \frac{1}{2(k+1)}$ be a sequence of stepsizes. Let $x_{k+1} = x_k - \delta_k x_k + \sqrt{2\delta_k} \eta_k$, and let $p_k$ be the distribution of $x_k$. Let $p^* = N(0, I)$, then

$$W_2 (p_k, p^*) = O \left( \frac{d^{3/2}}{\sqrt{k}} \right)$$

Proof of Lemma 47 First, we establish some properties of $\delta_k$.

By performing Taylor expansion of $\sqrt{x+1}$, we get that for $k \geq 2$,

$$\left| \delta_k - \frac{1}{2(k+1)} \right| \leq \frac{1}{k^2} \tag{68}$$

We also show that for integers $a \leq b$,

$$\sum_{i=a}^{b} \delta_i \leq \sum_{i=1}^{k} \frac{1}{2(i+1)} + \sum_{i=1}^{k} \frac{1}{k^2} \leq \frac{1}{2} \log \frac{b}{a} + 1$$

A similar argument proves a lower bound, so we have

$$\left| \sum_{i=a}^{b} \delta_i - \frac{1}{2} \log \frac{b}{a} \right| \leq 2 \tag{69}$$

Let $K$ be a sufficiently large integer such that

$$\delta_K = \frac{1}{2(K+1)} \leq \min \{ m^2, 1 \}$$

For any $k \geq K$, we can show that

$$W_2 (p_k, p^*) \leq W_2 (\Phi_{\delta_k} (p_{k-1}), \Phi_{\delta_k} (p^*)) + W_2 (\Phi_{\delta_k} (p^*), p^*)$$

$$\leq e^{-\delta_k} W_2 (p_{k-1}, p^*) + C \cdot d^{3/2} \cdot k^{-3/2}$$

$$\leq \exp \left( - \sum_{i=K}^{k} \delta_i \right) W_2 (p_K, p^*) + \sum_{i=K}^{k} \left( \exp \left( - \sum_{j=i}^{k} \delta_k \right) \cdot C \cdot d^{3/2} \cdot i^{-3/2} \right)$$

$$\leq 8 \exp \left( \frac{1}{2} \log \frac{k}{K} \right) W_2 (p_K, p^*) + 8C \cdot d^{3/2} \cdot \sum_{i=1}^{k} \left( \frac{1}{2} \log \frac{k}{i} \right) \cdot i^{-3/2}$$

$$\leq 8 \sqrt{\frac{K}{k}} + 8C \cdot d^{3/2} \cdot \sum_{i=1}^{k} \frac{1}{\sqrt{i}} \cdot \frac{1}{i}$$

$$\leq 8 \sqrt{\frac{K}{k}} + 8C \cdot d^{3/2} \frac{1}{\sqrt{k}} \log k$$

$$\leq C' \cdot d^{3/2} \frac{\log k}{\sqrt{k}}$$

where the first inequality is by triangle inequality, the second inequality is by Theorem 4 (with $k = 1$), and our assumption on $\delta_K$ and the fact that $\delta_k \leq \delta_K$, the third and fourth inequalities are by some algebra, the fifth inequality is by (69), the second last inequality is by harmonic sum.

In applying Theorem 4 in (70), we crucially used the fact that $p^* := N(0, I)$ is the invariant distribution to the SDE

$$dx(t) = -\nabla U(x(t)) dt + \sqrt{2} dB_t$$

for $U(x) = \frac{1}{2} \|x\|_2^2$, and the fact that

$$x_{k+1} = x_k - \delta_k \nabla U(x_k) + \sqrt{2\delta_k+1} \eta_k$$
Note also that the contraction term in (70), $e^{-2\delta_k}$ is tighter than is proven in Theorem 4, but this tighter contraction can easily be verified using synchronous coupling, i.e. for any two random variables $x_k$ and $y_k$,

$$\|x_k - \delta_k x_k - (y_k - \delta_k y_k)\|_2^2 \leq (1 - \delta_k)^2 \|x_k - y_k\|_2^2 \leq e^{-2\delta_k} \|x_k - y_k\|_2^2$$

■

**Corollary 48** Let $S_k := \sum_{i=1}^k \eta_i \sqrt{k}$. Let $q_k$ be the distribution of $S_k$ and let $p^* = N(0, I)$. Then $W_2(q_k, p^*) = \tilde{O}\left(\frac{d^{3/2}}{\sqrt{k}}\right)$

**Proof of Corollary 48** Let $\delta_k$, $x_k$ be as defined in Lemma 47, with initial $x_0 = 0$. It can be verified that

$$S_{k+1} = S_k - \delta_k S_k + \frac{1}{2} (\sqrt{k+1}) \eta_{k+1}$$

Thus

$$E[\|x_{k+1} - S_{k+1}\|_2^2] = E[\| (1 - \delta_k) (x_k - S_k) + \left( \delta_k - \frac{1}{2(k+1)} \right) \eta_{k+1} \|_2^2]$$

$$= E[\| (1 - \delta_k) (x_k - S_k) \|_2^2] + E\left[ \left( \delta_k - \frac{1}{2(k+1)} \right) \eta_{k+1} \right]^2$$

$$\leq \exp(-2\delta_k) E[\|x_k - S_k\|_2^2] + \frac{1}{k^2} d,$$

where the second last inequality is by the independence of $\eta_k$ and $E[\eta_k] = 0$, and the last inequality is by (69) and the fact that $E[\eta_k^T] = I$.

Applying the above inequality recursively, we get

$$E[\|x_k - S_k\|_2^2] \leq \sum_{i=1}^k \exp\left( - \sum_{j=i}^k 2\delta_j \right) \frac{d}{j^2} + \exp\left( - \sum_{i=1}^k \delta_k \right) E[\|x_0 - S_0\|_2^2]$$

$$\leq \sum_{i=1}^k \exp\left( - \log \frac{k}{i} \right) \frac{d}{i^2}$$

$$\leq \sum_{i=1}^k \frac{d}{i^2} \frac{d}{k}$$

$$\leq \frac{16d}{k} \log k,$$

where the second inequality is by (69), and the fact that $x_0 = S_0 = 0$.

Thus

$$W_2(x_k, S_k) = \tilde{O}\left(\frac{d^{3/2}}{\sqrt{k}}\right)$$

Together with the result from Lemma 17, we conclude our proof. ■