# Diffusion Generative Models meet Differential Privacy: A Theoretical Insight

**Ziyu Huang, Wenpin Tang** *
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027
{zh2532, wt2319}@columbia.edu

## Abstract

Score-based diffusion models have emerged as popular generative models trained on increasingly large datasets, yet they are often susceptible to attacks that can disclose sensitive information. To offer Differential Privacy (DP) guarantees, training these models for score-matching with DP-SGD has become a common solution. In this work, we study Differentially Private Diffusion Models (DPDM) both theoretically and empirically. We provide a quantitative $L^2$ rate of DP-SGD to its global optimum, leading to the first error analysis of diffusion models trained with DP-SGD. Our theoretical framework contributes to uncertainty quantification in generative AI systems, providing essential convergence guarantees for trustworthy decision-making applications that require both privacy preservation and reliability.

## 1 Introduction

Diffusion models have emerged as a powerful class of generative models, demonstrating remarkable success in generating high-quality synthetic data across various domains including images, text, and time series data [8, 11]. However, as these models are increasingly trained on large-scale datasets that often contain sensitive information, privacy concerns have become paramount. The ability of generative models to memorize and potentially leak training data has been well-documented [2, 3], raising fundamental questions about the privacy guarantees of diffusion models in practical applications.

To address these privacy vulnerabilities, differentially private training has emerged as the gold standard for providing formal privacy guarantees. Differential Privacy (DP) offers a mathematically rigorous framework that quantifies the maximum information leakage about any individual training example [5]. The canonical approach for training machine learning models with DP guarantees is DP-SGD [1], which adds calibrated noise to gradients during training. While empirical studies have shown the feasibility of training differentially private diffusion models [4, 6], theoretical understanding of the convergence properties and privacy-utility tradeoffs remains limited.

This paper provides theoretical analysis of diffusion models trained with differential privacy. We focus on the score-matching objective central to denoising diffusion probabilistic models (DDPMs) and analyze the convergence properties of DP-SGD applied to this objective. Our analysis reveals fundamental tradeoffs between privacy, accuracy, and computational complexity, providing theoretical insights that complement existing empirical studies and inform the design of privacy-preserving generative models.

*Authors ordered alphabetically.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

### 1.1 Our Contributions

We provide the first rigorous convergence analysis of DP-SGD applied to the score-matching objective used in diffusion models. Specifically, we establish upper bounds on the Wasserstein-2 distance between the output of DP-SGD and the global optimum, revealing how the convergence rate depends on the privacy parameters $(\epsilon, \delta)$, dimensionality $d$, and step size $\eta$. Our analysis shows that the convergence rate scales as $O(1+\eta^{1/2}d^{1/2}+\eta^{3/4}d^{3/4}+\eta d+\eta^{3/2}d^{3/2}+\eta^2 d^2+\eta^3 d^{9/4}+\eta^4 d^3+\eta^6 d^4)$, when $\eta$ is small and $d$ is large. When $\eta = O(d^{-\theta})$, the phase transition of the convergence rate is described in Table 1. We observe that in order to achieve less than constant convergence rate, the step size needs to be at most $O(d^{-1})$. This demonstrates the fundamental privacy-utility-dimensionality tradeoff in differentially private diffusion models. See Appendix A for the comparison of our work with the related works.

| range of $\theta$ | $\theta \geq 1$ | $\frac{1}{2} \leq \theta < 1$ | $0 < \theta < \frac{1}{2}$ |
|---|---|---|---|
| convergence rate | $O(1)$ | $O(d^{2-2\theta})$ | $O(d^{4-6\theta})$ |

Table 1: Convergence rates in terms of $\eta = O(d^{-\theta})$.

## 2 Preliminaries

### 2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) are latent variable models of the form $p_\theta(x_0) := \int p_\theta(x_{0:T})\, dx_{1:T}$, where $x_1, \ldots, x_T$ are latents of the same dimensionality as the data $x_0 \sim q(x_0)$. The model consists of two complementary Markov processes: a fixed forward process that gradually corrupts data by adding Gaussian noise, and a learned reverse process that generates data by iteratively denoising. The forward process $q(x_{1:T}|x_0)$ is defined as a Markov chain that adds Gaussian noise according to a variance schedule $\beta_1, \ldots, \beta_T$:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{1}$$

The reverse process $p_\theta(x_{0:T})$ is parameterized as a Markov chain with learned Gaussian transitions where $p(x_T) = \mathcal{N}(x_T; 0, I)$ is the prior distribution. Training is performed by optimizing the variational bound on negative log likelihood. However, [8] showed that superior empirical results are obtained using the **simplified training objective**:

$$\mathcal{L}_{\text{simple}}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2\right], \tag{2}$$

where $t \sim \text{Uniform}(1, \ldots, T)$ and $\epsilon \sim \mathcal{N}(0, I)$. In this parameterization, the neural network $\epsilon_\theta(x_t, t)$ is trained to predict the noise $\epsilon$ that was added to the original data $x_0$ to obtain the noisy observation $x_t$.

### 2.2 Differential Privacy and Algorithms

Differential privacy (DP) provides a rigorous mathematical framework for quantifying privacy guarantees in machine learning algorithms [1]. Formally, a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies $(\epsilon, \delta)$-differential privacy if for any two adjacent datasets $D, D' \in \mathcal{D}$ differing by a single record and any subset of outputs $S \subseteq \mathcal{R}$:

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S] + \delta,$$

where $\epsilon$ controls the privacy budget and $\delta$ accounts for the probability of privacy failure.

In our analysis, we compare standard Stochastic Gradient Descent (SGD) with Differentially Private SGD with Normalization (DP-NSGD) in the empirical risk minimization setting.

**Definition 1** (SGD with singleton batch)**.** *Starting with an arbitrary initial point $x_0$, for a fixed step size $\eta > 0$, the Stochastic Gradient Descent algorithm (SGD) iteratively updates*

$$x_{t+1} = x_t - \eta \nabla_x g_t(x_t), \tag{3}$$

*where $g_t$ is sampled uniformly from $\{f_1, \ldots, f_n\}$ independently.*

To provide differential privacy guarantees, we use DP-NSGD [14], which normalizes gradients instead of clipping them:

**Definition 2** (DP-NSGD with singleton batch)**.** *Starting with an arbitrary initial point $x_0$, for a fixed step size $\eta > 0$, a normalization parameter $r \geq 1$, and a noise parameter $\sigma > 0$, the Differentially Private SGD with Normalization (DP-NSGD) algorithm iteratively updates*

$$\tilde{x}_{t+1} = \tilde{x}_t - \eta \left( \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||} + z_t \right) \tag{4}$$

*where $\tilde{g}_t$ is sampled uniformly from $\{f_1, \ldots, f_n\}$ independently and $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$.*

The privacy guarantee for DP-NSGD is established by the following result:

**Lemma 3** (Privacy Guarantee [1])**.** *If the noise parameter $\sigma \geq \frac{\sqrt{2\log(1.25/\delta)}}{\epsilon}$, then DP-NSGD is $(O(\frac{\sqrt{T}\epsilon}{|X|}), \delta)$-differentially private, where $|X|$ is the size of the training set and $T$ is the number of iterations.*

A key advantage of DP-NSGD over traditional DP-SGD (which uses gradient clipping) is the ease of parameter tuning: the regularizer $r$ provides a more robust and intuitive hyperparameter that can be tuned without extensive knowledge of gradient magnitudes, unlike clipping thresholds which require careful dataset-specific calibration.

## 3 Main Result

Our theoretical analysis focuses on the optimization of the simplified diffusion objective introduced in Section 2.1. Since the objective function samples clean data from the training set, we consider the empirical risk minimization (ERM) formulation for optimization purposes. Given a training dataset $\mathcal{I} = \{\xi_i\}_{i=1}^n$ where each $\xi_i$ represents a training sample (which could be a noisy data point and timestep pair), we formulate the ERM problem as:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x) = \min_{x \in \mathbb{R}^d} f(x),$$

where $x$ is the parameter to be optimized. The individual loss functions $f_i(x)$ correspond to the samples of the simplified objective function from the training set. This formulation allows us to analyze the convergence properties of DP-NSGD applied to diffusion model training while working with the mathematically tractable objective that underlies the simplified diffusion loss.

Let $f^* := \inf_x f(x)$. We measure how close the output of DP-NSGD is to $f^*$ using the Wasserstein-2 distance $W_2(X, Y) = \inf_{\gamma \in \text{Coup}(X,Y)} \sqrt{\mathbb{E}_\gamma[||X - Y||^2]}$, where $\text{Coup}(X, Y)$ is the collection of all couplings between the distributions of $X$ and $Y$.

Our goal is to give an upper bound for the minimal $W_2$ distance between $f(\tilde{x}_t)$ and $f^*$ as $t$ goes to infinity:

$$\limsup_{t \to \infty} W_2(f(\tilde{x}_t), f^*).$$

We need several assumptions on the class of functions $\{f_i\}_{i=1}^n$.

**Assumption 4** (Lipschitz continuity)**.** *There exists a constant $C_1$ such that for any $i \in [n]$, $x, y \in \mathbb{R}^n$, $||\nabla f_i(x) - \nabla f_i(y)|| \leq C_1 ||x - y||$.*

This assumption ensures that gradients don't change too rapidly, which is crucial for controlling the accumulation of errors introduced by privacy noise.

**Assumption 5** (Bounded gradient variance)**.** *There exists a constant $C_2$ such that for any $i \in [n],\ x \in \mathbb{R}^n$, when $g$ is uniformly sampled from $\{f_i\}_{i=1}^n$, $\mathbb{E}[||\nabla g(x)||^2] = C_2$.*

This controls the inherent stochasticity in gradient estimates.

**Assumption 6** (Dissipation)**.** *There exists a constant $C_3$ such that for any $i \in [n], x, y \in \mathbb{R}^n$, $(x-y)^T(\nabla f_i(x) - \nabla f_i(y)) \geq C_3||x-y||^2$.*

This is a strong convexity-like condition that ensures the optimization landscape is well-conditioned. It guarantees that the algorithm makes consistent progress toward the optimum and prevents oscillatory behavior.

**Assumption 7** (Bounded gradient fourth moment)**.** *There exists a constant $C_6$ such that for any $i \in [n],\ x \in \mathbb{R}^n$, when $g$ is uniformly sampled from $\{f_i\}_{i=1}^n$, $\mathbb{E}[||\nabla g(x)||^4] = C_6$.*

This higher-order moment condition is needed to control the variance of our coupling analysis. It ensures that extreme gradient values don't dominate the convergence analysis.

**Assumption 8** (Polyak-Lojasiewicz (PL) Condition)**.** *There exists a constant $\mu$ such that, for any $x$ in the domain of $f$, $||\nabla f(x)||^2 \geq 2C_8(f(x) - f^*)$, where $f^*$ is the global infinimum of $f$.*

This condition ensures that whenever we are far from the optimum (large $f(x) - f^*$), the gradients are sufficiently large to make progress. This is weaker than strong convexity but still ensures convergence.

**Assumption 9** (Expected Residual)**.** *Each $f_i$ satisfies the expected residual condition. That is, there exists $C_7 > 0$ such that*

$$\mathbb{E}[||f_i(x) - f_i(x^*) - (\nabla f(x) - \nabla f(x^*))||^2] \leq 2C_7(f(x) - f(x^*)),\ \forall x \in \mathbb{R}^n$$

This technical condition bounds the variance between individual functions $f_i$ and the population function $f$. It ensures that the finite-sample approximation doesn't deviate too much from the population objective.

To give an upper bound for $\limsup_{t\to\infty} W_2(f(\tilde{x}_t), f^*)$, we separate $W_2(f(\tilde{x}_t), f^*)$ into two terms $W_2(f(\tilde{x}_t), f(x_t))$ and $W_2(f(x_t), f^*)$ and give an upper bound on each of them. See Appendix B for the complete proof and methodology.

**Lemma 10** (Main Coupling Lemma)**.** *Let $\tilde{x}_t, x_t$ be the t-iterate of DP-NSGD, SGD out of $T$ iterations respectively. Assume each $f_i$ satisfies Assumptions 4–7. Let $\pi_{t-1}$ be a coupling of $x_t, \tilde{x}_t$ such that $g_l = \tilde{g}_l$ for any $1 \leq l \leq t-1$. Under appropriate step size conditions,*

$$\limsup_{t\to\infty} W_2(f(\tilde{x}_t), f(x_t)) = O(1+\eta^{1/2}d^{1/2}+\eta^{3/4}d^{3/4}+\eta d+\eta^{3/2}d^{3/2}+\eta^2 d^2+\eta^3 d^{9/4}+\eta^4 d^3+\eta^6 d^4).$$

Combining this lemma with an upper bound on $W_2(f(x_t), f^*)$ [7], we have our main theorem.

**Theorem 11** (Main Result)**.** *Let $\{f_i\}$ satisfy Assumptions 4–9. Under appropriate step size conditions:*

$$\limsup_{t\to\infty} W_2(f(\tilde{x}_t), f^*) = O(1+\eta^{1/2}d^{1/2}+\eta^{3/4}d^{3/4}+\eta d+\eta^{3/2}d^{3/2}+\eta^2 d^2+\eta^3 d^{9/4}+\eta^4 d^3+\eta^6 d^4). \tag{5}$$

This theorem establishes the fundamental privacy-utility-dimensionality tradeoffs in differentially private diffusion models, showing how the convergence rate degrades with increasing dimensionality and privacy constraints.

## 4 Conclusion

In this work, we provided a theoretical analysis of diffusion models trained with differential privacy, establishing rigorous convergence bounds for DP-SGD applied to score-matching objectives and revealing fundamental privacy-utility-dimensionality tradeoffs. Our analysis demonstrates that convergence rates scale polynomially with dimensionality, highlighting the curse of dimensionality in privacy-preserving generative modeling, while providing theoretical foundations that complement empirical studies. Future work includes extending analysis to sophisticated diffusion variants, investigating alternative privacy mechanisms, and experiments on financial time series, and supply chain logistics data.

## References


[1] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[2] N. Carlini, D. Paleka, K. D. Dvijotham, T. Steinke, J. Hayase, A. F. Cooper, K. Lee, M. Jagielski, M. Nasr, A. Conmy, E. Wallace, D. Rolnick, and F. Tramèr. Stealing part of a production language model. *ArXiv*, abs/2403.06634, 2024.

[3] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. X. Song, Ú. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2020.

[4] T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis. Differentially private diffusion models. *ArXiv*, abs/2210.09929, 2022.

[5] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confidentiality*, 7:17–51, 2006.

[6] S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle. Differentially private diffusion models generate useful synthetic images. *ArXiv*, abs/2302.13861, 2023.

[7] R. M. Gower, O. Sebbouh, and N. Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[9] J. Liu, A. Lowy, T. Koike-Akino, K. Parsons, and Y. Wang. Efficient differentially private fine-tuning of diffusion models. *ArXiv*, abs/2406.05257, 2024.

[10] S. Lyu, M. Vinaroz, M. F. Liu, and M. Park. Differentially private latent diffusion models. *Trans. Mach. Learn. Res.*, 2024, 2023.

[11] Y. Song, J. N. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020.

[12] Y.-L. Tsai, Y. Li, Z. Chen, P.-Y. Chen, C.-M. Yu, X. Ren, and F. Buet-Golfouse. Differentially private fine-tuning of diffusion models. *ArXiv*, abs/2406.01355, 2024.

[13] H. Wang, S. Pang, Z. Lu, Y. Rao, Y. Zhou, and M. Xue. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *USENIX Security Symposium*, 2024.

[14] X. Yang, H. Zhang, W. Chen, and T.-Y. Liu. Normalized/clipped sgd with perturbation for differentially private non-convex optimization. *ArXiv*, abs/2206.13033, 2022.

[15] Y. Yang, R. Gao, X. Wang, N. Xu, and Q. Xu. MMA-diffusion: Multimodal attack on diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7737–7746, 2023.

# A Literature Review

The intersection of differential privacy and diffusion models emerged as a critical research area following demonstrated privacy vulnerabilities in generative systems. [3] first demonstrated that large language models leak training data, while [2] extended this to production model extraction attacks, and [15] revealed multimodal vulnerabilities specific to diffusion models. In response to these privacy threats, [4] introduced the first comprehensive framework for differentially private diffusion models (DPDM), establishing the theoretical and practical foundations for privacy-preserving generative modeling. Building on this pioneering work, [6] provided empirical validation that DPDM can generate high-utility synthetic images with formal privacy guarantees, while [10] improved computational efficiency through latent space operations. Recent advances have focused on practical deployment with [13] presenting comprehensive production-scale evaluation and efficient fine-tuning approaches developed by [9] and [12], collectively establishing differentially private training as the primary defense against privacy attacks on diffusion models. Despite these empirical successes, theoretical understanding of convergence properties and privacy-utility tradeoffs in DPDM remains limited, motivating the need for rigorous analysis of DP-SGD applied to score-based generative models.

# B Detailed Proofs

## B.1 Proof of Lemma 10

We denote $a_t := \mathbb{E}_{\pi_{t-1}}[||x_t - \tilde{x}_t||^2]$ for any $0 \le t \le T$.

**Lemma 12.** *Assume Assumption 5 and Assumption 6, for any $t \ge 0$,*

$$a_{t+1} \le (1 - 2C_3\eta + 2C_1\eta^2)a_t + (2C_2\eta)\sqrt{a_t} + \eta^2(\sigma^2 d + 2C_2^2).$$

*Proof.* First, we observe that

$$\begin{aligned}
\mathbb{E}_{\pi_t}[||x_{t+1} - \tilde{x}_{t+1}||^2] &= \mathbb{E}_{\pi_t}[||(x_t - \eta\nabla g_t(x_t)) - (\tilde{x}_t - \eta(\frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||} + z_t))||^2] \\
&= \mathbb{E}_{\pi_t}[||(x_t - \tilde{x}_t) - \eta(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}) + \eta z_t||^2] \\
&= \mathbb{E}_{\pi_{t-1}}[||x - \tilde{x}_t||^2] + \eta^2\mathbb{E}_{\pi_t}[||\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}||^2] + \eta^2\mathbb{E}[||z_t||^2] \\
&\quad - 2\eta\mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||})] \\
&\quad - 2\eta^2\mathbb{E}_{\pi_t}[(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||})^T z_t] + 2\eta\mathbb{E}_{\pi_{t-1}}[(x_t - \tilde{x}_t)^T z_t] \\
&= \mathbb{E}_{\pi_{t-1}}[||x - \tilde{x}_t||^2] + \eta^2\mathbb{E}_{\pi_t}[||\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}||^2] + \eta^2\mathbb{E}[||z_t||^2] \\
&\quad - 2\eta\mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||})] \\
&\quad - 2\eta^2\mathbb{E}_{\pi_t}[\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}] \cdot \cancelto{0}{\mathbb{E}[z_t]} + 2\eta\mathbb{E}_{\pi_{t-1}}[x_t - \tilde{x}_t] \cdot \cancelto{0}{\mathbb{E}[z_t]} \\
&\le a_t + \underbrace{\eta^2\mathbb{E}_{\pi_t}[||\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}||^2]}_{\text{I}} + \eta^2\sigma^2 d \\
&\quad \underbrace{-2\eta\mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||})]}_{\text{II}}.
\end{aligned} \tag{6}$$

where the first equality is by the fact that $z_t$ is independent of any other random variables. We let $\text{I} := \eta^2\mathbb{E}_{\pi_t}[||\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r+||\nabla \tilde{g}_t(\tilde{x}_t)||}||^2]$ and $\text{II} := -2\eta\mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r+||\nabla \tilde{g}_t(\tilde{x}_t)||})]$.

Under the coupling condition $\pi_t$, $g_l$ and $\tilde{g}_l$ are identical for all $0 \leq l \leq t$. To ease the notation, when conditioning on $\pi_t$, we denote $\nabla g_t(x_t)$ by $h_t$ and $\nabla \tilde{g}_t(\tilde{x}_t) = \nabla g_t(\tilde{x}_t)$ by $\tilde{h}_t$. For I,

$$\begin{aligned}
\mathrm{I} &= \eta^2 \mathbb{E}_{\pi_t}[||h_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^2] = \eta^2 \mathbb{E}_{\pi_t}[||h_t - \tilde{h}_t + \tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^2] \\
&\leq \underbrace{2\eta^2 \mathbb{E}_{\pi_t}[||h_t - \tilde{h}_t||^2]}_{\mathrm{III}} + \underbrace{2\eta^2 \mathbb{E}_{\pi_t}[||\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^2]}_{\mathrm{IV}},
\end{aligned}$$

where the first inequality is due to the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for any two reals $a, b$. By Assumption 4, we have that

$$\mathrm{III} = 2\eta^2 \mathbb{E}_{\pi_t}[||h_t - \tilde{h}_t||^2] = 2\eta^2 \mathbb{E}[||\nabla f_i(x_t) - \nabla f_i(\tilde{x}_t)||^2] \leq 2\eta^2 C_1 \mathbb{E}_{\pi_{t-1}}[||x_t - \tilde{x}_t||^2] = 2\eta^2 C_1 a_t. \tag{7}$$

Furthermore,

$$\mathrm{IV} = 2\eta^2 \mathbb{E}_{\pi_t}[||\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^2] = 2\eta^2 \mathbb{E}[(\frac{r + ||\tilde{h}_t|| - 1}{r + ||\tilde{h}_t||}||\tilde{h}_t||)^2] \leq 2\eta^2 \mathbb{E}[||\tilde{h}_t||^2] \leq 2\eta^2 C_2^2 \tag{8}$$

where the second inequality is by Assumption 5. Combining Eq (7) and Eq (8), we obtain that

$$\mathrm{I} \leq 2\eta^2 C_1 a_t + 2\eta^2 C_2^2. \tag{9}$$

Then, we will upper bound II,

$$\begin{aligned}
\mathrm{II} &= -2\eta \mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T (h_t - \tilde{h}_t + \tilde{h}_t - \frac{\tilde{h}_t}{r + |\tilde{h}_t|})] \\
&= \underbrace{-2\eta \mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T (h_t - \tilde{h}_t)]}_{\mathrm{V}} \underbrace{-2\eta \mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T (\tilde{h}_t - \frac{\tilde{h}_t}{r + |\tilde{h}_t|})]}_{\mathrm{VI}}.
\end{aligned} \tag{10}$$

For the first term on the right hand side, by Assumption 6,

$$\mathrm{V} = -2\eta \mathbb{E}_{\pi_t}[(x_t - \tilde{x}_t)^T (h_t - \tilde{h}_t)] \leq -2\eta C_3 \mathbb{E}_{\pi_{t-1}}[||x_t - \tilde{x}_t||^2] = -2\eta C_3 a_t. \tag{11}$$

For the second term on the right hand side,

$$\begin{aligned}
\mathrm{VI} \leq |\mathrm{VI}| &\leq 2\eta \mathbb{E}_{\pi_t}[|(x_t - \tilde{x}_t)^T (\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||})|] \leq 2\eta \mathbb{E}_{\pi_t}[||(x_t - \tilde{x}_t)|| ||(\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||})||] \\
&\leq 2\eta \sqrt{\mathbb{E}_{\pi_t}[||(x_t - \tilde{x}_t)||^2]} \sqrt{\mathbb{E}[||(\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||})||^2]} \leq 2\eta \sqrt{a_t} C_2
\end{aligned} \tag{12}$$

where the second inequality is by Jensen inequality. The third inequality is by Cauchy-Schwarz over $\mathbb{R}^n$. The fourth inequality is by Cauchy-Schwarz over the probability space. The fifth inequality is by Eq (8). We remark that in order to use Cauchy-Schartz, we require that $r \geq 1$ so that $\frac{r + ||\tilde{h}_t|| - 1}{r + ||\tilde{h}_t||} > 0$. Combining Eq (11) and Eq (12), we obtain that

$$\mathrm{II} \leq -2\eta C_3 a_t + 2\eta C_2 \sqrt{a_t}. \tag{13}$$

Finally, combining Eq (6), Eq (9), and Eq (13), we obtain that

$$\begin{aligned}
a_{t+1} &\leq a_t + \mathrm{I} + \eta^2 \sigma^2 d + \mathrm{II} \\
&\leq a_t + \eta^2 (2C_1 a_t + 2C_2^2) + \eta^2 \sigma^2 d + (-2\eta C_3 a_t + 2\eta C_2 \sqrt{a_t}) \\
&\leq (1 - 2C_3 \eta + 2C_1 \eta^2) a_t + (2C_2 \eta) \sqrt{a_t} + \eta^2 (\sigma^2 d + 2C_2^2).
\end{aligned} \tag{14}$$

□

Define functions in $\eta$,

$$A_1(\eta) = 2C_1\eta^2 - 2C_3\eta + 1,$$
$$A_2(\eta) = 2C_2\eta,$$
$$A_3(\eta) = (\sigma^2 d + 2C_2^2)\eta^2.$$

Define functions in a free variable $y$,

$$p(y) = A_1 y + A_2\sqrt{y} + A_3,$$
$$q(y) = 1 - p(y) = (1 - A_1)y - A_2\sqrt{y} - A_3.$$

By Lemma 12, we have the following corollary,

**Corollary 13.** $a_{t+1} \leq p(a_t)$.

We want to show $p(y)$ is increasing in $y$ and $q(y)$ is increasing in $y$ when $y$ is large enough. For that, we need the following result,

**Lemma 14.** *When* $0 < \eta < \dfrac{C_3 - \sqrt{C_3^2 - 2C_1}}{2C_1}$, $0 < A_1(\eta) < 1$.

*Proof.* We first observe that $A_1(\eta)$ is a quadratic function in $\eta$. If $C_3^2 < 2C_1$ then there are no zeros and $A_1$ is always positive. If $C_3^2 \geq 2C_1$, the function has two zeros which are $\eta_1 = \dfrac{C_3 - \sqrt{C_3^2 - 2C_1}}{2C_1}$ and $\eta_2 = \dfrac{C_3 + \sqrt{C_3^2 - 2C_1}}{2C_1}$. Since we required the step size $\eta < \eta_1$, $A_1$ will always be positive. Furthermore, $A_1(\eta) < 1$ if and only if $0 < \eta < \dfrac{C_3}{C_1}$. Since we required that $0 < \eta < \dfrac{C_3 - \sqrt{C_3^2 - 2C_1}}{2C_1} < \dfrac{C_3}{C_1}$, we know that $A_1(\eta) < 1$.

□

We show the following property of $p(y)$ and $q(y)$, which will be helpful in our later proof.

**Lemma 15.** $q(y)$ *has a unique zero at*

$$y_* = (\frac{A_2 + \sqrt{A_2^2 + 4(1 - A_1)A_3}}{2(1 - A_1)})^2.$$

*When* $0 \leq y < y_*$, $q(y) < 0$. *When* $y > y_*$, $q(y) > 0$.

*Proof.* Let $z = \sqrt{y}$ so $q(z) = (1 - A_1)z^2 - A_2 z - A_3$ with domain $z > 0$. $q(z)$ is a quadratic with two roots

$$z_1 = \frac{A_2 - \sqrt{A_2^2 + 4(1 - A_1)A_3}}{2(1 - A_1)} < 0, \quad z_2 = \frac{A_2 + \sqrt{A_2^2 + 4(1 - A_1)A_3}}{2(1 - A_1)} > 0$$

and we find out $z_2$ is the unique zero in the domain so $y_* := z_2^2 = (\dfrac{A_2 + \sqrt{A_2^2 + 4(1 - A_1)A_3}}{2(1 - A_1)})^2$ is the unique zero of $q(y)$. Since $q(0) = -A_3 = -\eta^2(\sigma^2 d + 2C_2^2) < 0$ and $\lim_{y\to\infty} q(y) = \lim_{y\to\infty}(1 - A_1)y = \infty$, by continuity of $q(y)$, we conclude that when $0 \leq y < y_*$, $q(y) < 0$. When $y > y_*$, $q(y) > 0$. □

We define a new sequence $(y_t)_{t\geq 0}$ such that $y_0 = a_0$ and $y_{t+1} = p(y_t)$.

**Lemma 16.** *The following statements are true for* $(y_t)$*:*

1. $y_t \geq a_t$ *for any* $t$*;*
2. $\lim_{t\to\infty} y_t = y_*$.

*Proof.* We show all statements by induction on $(y_t)$. To show statement 1, we first see $y_0 \geq a_0$. Assume $y_t \geq a_t$, then $y_{t+1} = p(y_t) \geq p(a_t) \geq a_{t+1}$. The first inequality is by the fact that $p(y)$ is monotonically increasing. The second inequality is by Corollary 13.

We now show the second statement. If $y_* = y_0$, then by induction, since $y_*$ is the unique zero of $q(y)$, $y_{t+1} = p(y_t) = p(y_*) = y_*$ as desired. If $y_0 > y_*$, we claim that $(y_t)$ is bounded below by $y_*$ and decreasing so converges. We observe that $y_{t+1} - y_* = p(y_t) - p(y_*) > 0$ since $p$ is increasing and $y_t > y_*$ by assumption. Then, $y_t - y_{t+1} = y_t - p(y_t) = q(y_t) > 0$ where the inequality is by Lemma 15 and the fact that $y_t > y_*$. Thus, the limit of $(y_t)$ exists and we denote it $y_\infty := \lim_{t\to\infty} y_t$. We observe that $y_\infty = \lim_{t\to\infty} y_t = \lim_{t\to\infty} p(y_{t-1}) = \lim_{t\to\infty} p(y_t)$. Since $p$ is continuous, $p(y_\infty) = \lim_{t\to\infty} p(y_t) = y_\infty$ so $y_\infty$ is a zero of $q(y)$. Since $q$ has the unique zero $y_*$, we obtain that $y_\infty = y_*$.

If $y_0 < y_*$, we claim that $(y_t)$ is bounded above by $y_*$ and increasing so converges. We observe that $y_{t+1} - y_* = p(y_t) - p(y_*) < 0$ since $p$ is increasing and $y_t < y_*$ by assumption. Then, $y_t - y_{t+1} = y_t - p(y_t) = q(y_t) < 0$ where the inequality is by Lemma 15 and the fact that $y_t < y_*$. Thus, the limit of $(y_t)$ exists and we denote it $y_\infty := \lim_{t\to\infty} y_t$. We observe that $y_\infty = \lim_{t\to\infty} y_t = \lim_{t\to\infty} p(y_{t-1}) = \lim_{t\to\infty} p(y_t)$. Since $p$ is continuous, $p(y_\infty) = \lim_{t\to\infty} p(y_t) = y_\infty$ so $y_\infty$ is a zero of $q(y)$. Since $q$ has the unique zero $y_*$, we conclude that $y_\infty = y_*$.

□

From Lemma 16,

$$\begin{aligned}
\limsup_t a_t \leq \limsup_t y_t = \lim_t y_t = y_* &= \left(\frac{A_2 + \sqrt{A_2^2 + 4(1-A_1)A_3}}{2(1-A_1)}\right)^2 \\
&\leq \left(\frac{A_2 + \sqrt{4(1-A_1)A_3}}{2(1-A_1)}\right)^2 \\
&= \left(\frac{(2C_2\eta)^2 + \sqrt{4(2C_3\eta - 2C_1\eta^2)(\sigma^2 d + 2C_2^2)\eta^2}}{2(2C_3\eta - 2C_1\eta^2)}\right)^2 \\
&= \left(\frac{(2C_2)^2\eta + \sqrt{4(2C_3\eta - 2C_1\eta^2)(\sigma^2 d + 2C_2^2)}}{2(2C_3 - 2C_1\eta)}\right)^2 \\
&= 2\left(\frac{(2C_2)^2\eta}{2(2C_3 - 2C_1\eta)}\right)^2 + 2\left(\frac{\sqrt{4(2C_3\eta - 2C_1\eta^2)(\sigma^2 d + 2C_2^2)}}{2(2C_3 - 2C_1\eta)}\right)^2.
\end{aligned} \tag{15}$$

When $\eta$ approaches to 0 and $d$ approaches to $\infty$, we have that $\limsup_t a_t = O(\eta^2) + O(\eta d)$. We use the same contraction idea to give an upper bound on $b_{t+1} := \mathbb{E}_{\pi_t}[||x_{t+1} - \tilde{x}_{t+1}||^4]$.

**Lemma 17.**

$$b_{t+1} \leq Hb_t + Ib_t^{\frac{3}{4}} + Jb_t^{\frac{1}{2}} + Kb_t^{\frac{1}{4}} + L \tag{16}$$

*where*

$$\begin{aligned}
H &= 1 - \eta 2^2 C_3 + \eta^4 2^3 C_1^4 + \eta^2 2^{\frac{9}{2}} C_1^2 + \eta^2 2^{\frac{5}{2}} C_1^2 + \eta^4 2^{\frac{17}{4}} C_1^3 \\
I &= \eta 2^2 C_6^{\frac{1}{4}} + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1^2 C_6^{\frac{1}{4}} \\
J &= \eta^2 2^2 d\sigma^2 + \eta^2 2 d\sigma^2 + \eta^4 2^3 d\sigma^2 C_1 + \eta^4 2^2 d\sigma^2 C_1 + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_1^{\frac{1}{2}} + \eta^2 2^4 C_6^{\frac{1}{2}} + \eta^2 32^{\frac{1}{2}} C_6^{\frac{1}{2}} \\
&\quad + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1 C_6^{\frac{1}{2}} + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_1^{\frac{1}{2}} \\
K &= \eta^3 2^2 \sigma^3 d^{\frac{3}{2}} + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_1^{\frac{1}{2}} + \eta^4 2^{\frac{17}{4}} C_6^{\frac{3}{4}} + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_2 + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_2 \\
L &= \eta^4 8 C_6 + \eta^4 \sigma^4 (d^2 + 2d) + \eta^4 8 C_2^2 d\sigma^2 + \eta^4 4 C_2^2 d\sigma^2 + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_2).
\end{aligned}$$

*Proof.* First, we denote $E := x_t - \tilde{x}_t,\ F := -(\nabla g_t(x_t) - \frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||})$.

$$
\begin{aligned}
b_{t+1} &= \mathbb{E}_{\pi_t}[||x_{t+1} - \tilde{x}_{t+1}||^4] \\
&= \mathbb{E}_{\pi_t}[||(x_t - \eta\nabla g_t(x_t)) - (\tilde{x}_t - \eta(\frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||} + z_t))||^4] \\
&= \mathbb{E}_{\pi_t}[||\underbrace{(x_t - \tilde{x}_t)}_{E} + \underbrace{\eta(-\nabla g_t(x_t) + (\frac{\nabla \tilde{g}_t(\tilde{x}_t)}{r + ||\nabla \tilde{g}_t(\tilde{x}_t)||}))}_{F} + \eta z_t||^4] \\
&= \mathbb{E}_{\pi_t}[||E + \eta F + \eta z_t||^4] \\
&= \mathbb{E}_{\pi_t}[(||E||^2 + \eta^2||F||^2 + \eta^2||z_t||^2 + 2\eta E^T F + 2\eta E^T z_t + 2\eta^2 F^T z_t)^2] \\
&= \mathbb{E}_{\pi_t}[||E||^4 + \eta^4||F||^4 + \eta^4||z_t||^4 + 4\eta^2(E^T F)^2 + 4\eta^2(E^T z_t)^2 + 4\eta^4(F^T z_t)^2 \\
&\quad + 2\cdot||E||^2\cdot\eta^2||F||^2 + 2\cdot||E||^2\cdot\eta^2||z_t||^2 + 2\cdot||E||^2\cdot 2\eta(E^T F) + 2\cdot||E||^2\cdot 2\eta(E^T z_t) + 2\cdot||E||^2\cdot 2\eta^2(F^T z_t) \\
&\quad + 2\cdot\eta^2||F||^2\cdot\eta^2||z_t||^2 + 2\cdot\eta^2||F||^2\cdot 2\eta(E^T F) + 2\cdot\eta^2||F||^2\cdot 2\eta(E^T z_t) + 2\cdot\eta^2||F||^2\cdot 2\eta^2(F^T z_t) \\
&\quad + 2\cdot\eta^2||z_t||^2\cdot 2\eta(E^T F) + 2\cdot\eta^2||z_t||^2\cdot 2\eta(E^T z_t) + 2\cdot\eta^2||z_t||^2\cdot 2\eta^2(F^T z_t) \\
&\quad + 2\cdot 2\eta(E^T F)\cdot 2\eta(E^T z_t) + 2\cdot 2\eta(E^T F)\cdot 2\eta^2(F^T z_t) \\
&\quad + 2\cdot 2\eta(E^T z_t)\cdot 2\eta^2(F^T z_t)],
\end{aligned} \tag{17}
$$

where the third and the fourth equality is the expansion of the polynomials. We give an upper bound for each term separately in terms of $b_t$ and other constants.

1. $\mathbb{E}_{\pi_t}[||E||^4] = \mathbb{E}_{\pi_t}[||x_t - \tilde{x}_t||^4] = b_t$
2. Recall that $h_t := \nabla g_t(x_t)$, $\tilde{h}_t := \nabla \tilde{g}_t(\tilde{x}_t)$. Then,

$$
\mathbb{E}_{\pi_t}[\eta^4||F||^4] = \eta^4\mathbb{E}_{\pi_t}[||h_t - \tilde{h}_t + \tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^4] \le \underbrace{8\eta^4\mathbb{E}_{\pi_t}[||h_t - \tilde{h}_t||^4]}_{\text{VII}} + \underbrace{8\eta^4\mathbb{E}_{\pi_t}[||\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^4]}_{\text{VIII}}, \tag{18}
$$

where the inequality is by the fact that $(a+b)^4 \le 8a^4 + 8b^4$ for any two real numbers $a, b$. By Assumption 4, VII $\le 8\eta^4 C_1^4 b_t$. By Assumption 7,

$$
\text{VIII} = 8\eta^4\mathbb{E}_{\pi_t}[||\tilde{h}_t - \frac{\tilde{h}_t}{r + ||\tilde{h}_t||}||^4] \le 8\eta^4\mathbb{E}_{\pi_t}[(\frac{r + ||\tilde{h}_t|| - 1}{r + ||\tilde{h}_t||})^4||\tilde{h}_t||^4] \le 8\eta^4 C_6, \tag{19}
$$

Thus,

$$
\mathbb{E}_{\pi_t}[\eta^4||F||^4] \le \eta^4(8C_1^4 b_t + 8C_6). \tag{20}
$$

3. We calculate the fourth moment of the Gaussian distribution.

$$
\begin{aligned}
\mathbb{E}_{\pi_t}[\eta^4||z_t||^4] &= \eta^4\mathbb{E}_{\pi_t}[(\sum_{j=1}^{d}(z_t^j)^2)^2] = \eta^4(\sum_j \mathbb{E}_{\pi_t}[z_t]^4 + 2\sum_{k<l}\mathbb{E}_{\pi_t}[(z_t^k)^2]\mathbb{E}_{\pi_t}[(z_t^l)^2]) = \eta^4(3d\sigma^4 + d(d-1)\sigma^4) \\
&= \eta^4\sigma^4(d^2 + 2d).
\end{aligned}
$$

4.

$$
\begin{aligned}
\mathbb{E}_{\pi_t}[4\eta^2(E^T F)^2] &\le \mathbb{E}_{\pi_t}[4\eta^2||E||^2||F||^2] \le 4\eta^2\sqrt{\mathbb{E}_{\pi_t}[||E||^4]\mathbb{E}_{\pi_t}[||F||^4]} \le 4\eta^2\cdot\sqrt{b_t}\cdot\sqrt{8C_1^4 b_t + 8C_6} \\
&\le 4\eta^2\sqrt{b_t}(\sqrt{8C_1^4 b_t} + \sqrt{8C_6}) = 8\sqrt{2}\eta^2 C_1^2 b_t + 8\sqrt{2}\eta^2 C_6^{\frac{1}{2}} b_t^{\frac{1}{2}}
\end{aligned} \tag{21}
$$

where the first inequality is by Cauchy-Schwarz in $\mathbb{R}^d$. The second inequality is by Cauchy-Schwarz over the probability space. The third inequality is by Eq (20). The fourth inequality is by subadditivity of the square root function.

5.

$$\mathbb{E}_{\pi_t}[4\eta^2(E^T z_t)^2] \le \mathbb{E}_{\pi_t}[4\eta^2||E||^2||z_t||^2] = 4\eta^2\mathbb{E}_{\pi_t}[||E||^2]\mathbb{E}_{\pi_t}[||z_t||^2] \le 4\eta^2 \cdot \sqrt{b_t} \cdot d\sigma^2$$

where the first equality is by independence and the second inequality is by Jensen's inequality.

6.

$$\mathbb{E}_{\pi_t}[4\eta^4(F^T z_t)^2] \le 4\eta^4\mathbb{E}_{\pi_t}[||F||^2] \cdot \mathbb{E}_{\pi_t}[||z_t||^2] = 4\eta^4 \cdot (2C_1\sqrt{b_t} + 2C_2^2) \cdot d\sigma^2$$

where the first equality is by Eq (9).

7.

$$\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot \eta^2||F||^2] \le 2\eta^2\sqrt{\mathbb{E}_{\pi_t}[||E||^4]\mathbb{E}_{\pi_t}[||F||^4]} \le 4\sqrt{2}\eta^2 C_1^2 b_t + 4\sqrt{2}\eta^2 C_6^{\frac{1}{2}} b_t^{\frac{1}{2}},$$

followed by Eq(21) and subadditivity of square root fuction.

8.

$$\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot \eta^2||z_t||^2] = 2\eta^2 \cdot \mathbb{E}_{\pi_t}[||E||^2] \cdot \mathbb{E}_{\pi_t}[||z_t||^2] \le 2\eta^2 \cdot \sqrt{b_t} \cdot d\sigma^2$$

by independence of Gaussian distribution.

9.

$$\begin{aligned}\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot 2\eta(E^T F)] &= 4\eta\mathbb{E}_{\pi_t}[-||E||^2 E^T(h - \tilde{h} + \tilde{h} - \frac{\tilde{h}}{r + ||\tilde{h}||})] \\ &= 4\eta \cdot (\underbrace{-\mathbb{E}_{\pi_t}[||E||^2(x_t - \tilde{x}_t)^T(h - \tilde{h})]}_{\text{IX}} \underbrace{-\mathbb{E}_{\pi_t}[||E||^2 E^T(\tilde{h} - \frac{\tilde{h}}{r + ||\tilde{h}||})]}_{\text{X}}).\end{aligned} \tag{22}$$

We observe that

$$\text{IX} \le -\mathbb{E}_{\pi_t}[||E||^2 \cdot C_3||E||^2] = -C_3 b_t.$$

where the inequality is by Assumption 6. Furthermore,

$$\text{X} \le \mathbb{E}_{\pi_t}[||E||^2 \cdot ||E|| \cdot ||\tilde{h} - \frac{\tilde{h}}{r + \tilde{h}}||] \le (\mathbb{E}_{\pi_t}[(||E||^3)^{\frac{4}{3}}])^{\frac{3}{4}}(\mathbb{E}_{\pi_t}[||\tilde{h} - \frac{\tilde{h}}{r + \tilde{h}}||^4])^{\frac{1}{4}} \le b_t^{\frac{3}{4}} \cdot C_6^{\frac{1}{4}},$$

where the first inequality is by Cauchy-Schwarz over $\mathbb{R}^d$. The second inequality is by Hölder's inequality. The third inequality is by Eq (19). Therefore,

$$\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot 2\eta(E^T F)] \le 4\eta(-C_3 b_t + b_t^{\frac{3}{4}} \cdot C_6^{\frac{1}{4}}).$$

10. $\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot 2\eta(E^T z_t)] = \mathbb{E}_{\pi_t}[(4\eta||E||^2 E)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[4\eta||E||^2 E^j]\mathbb{E}_{\pi_t}[z_t^j] = 0.$

11. $\mathbb{E}_{\pi_t}[2 \cdot ||E||^2 \cdot 2\eta^2(F^T z_t)] = \mathbb{E}_{\pi_t}[(4\eta^2||E||^2 F)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[4\eta^2||E||^2 F^j]\mathbb{E}_{\pi_t}[z_t^j] = 0.$

12.

$$\mathbb{E}_{\pi_t}[2 \cdot \eta^2||F||^2 \cdot \eta^2||z_t||^2] = 2\eta^4\mathbb{E}_{\pi_t}[||F||^2]\mathbb{E}_{\pi_t}[||z_t||^2] \le 2\eta^4 d\sigma^2(2C_1\sqrt{b_t} + 2C_2^2),$$

where the inequality is by Eq (20).

13.

$$\begin{aligned}\mathbb{E}_{\pi_t}[2 \cdot \eta^2||F||^2 \cdot 2\eta(E^T F)] &\le 4\eta^3\mathbb{E}_{\pi_t}[||F||^3||E||] \le 4\eta^3(\mathbb{E}_{\pi_t}[||F||^4])^{\frac{3}{4}}(\mathbb{E}_{\pi_t}[||E||^4])^{\frac{1}{4}} \\ &\le 4\eta^3(8C_1^4 b_t + 8C_6)^{\frac{3}{4}} b_t^{\frac{1}{4}} \le 16\sqrt[4]{2}\eta^3 C_1^3 b_t + 16\sqrt[4]{2}\eta^3 C_6^{\frac{3}{4}} b_t^{\frac{1}{4}},\end{aligned}$$

where the first inequality is by Cauchy-Schwarz over $\mathbb{R}^d$. The second inequality is by Hölder's inequality. The third inequality is by Eq (20). The fourth inequality is by subadditivity.

14. $\mathbb{E}_{\pi_t}[2\cdot\eta^2||F||^2\cdot 2\eta(E^T z_t)] = \mathbb{E}_{\pi_t}[(4\eta^3||F||^2E)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[4\eta^3||F||^2E^j]\mathbb{E}_{\pi_t}[z_t^j] = 0.$

15. $\mathbb{E}_{\pi_t}[2\cdot\eta^2||F||^2\cdot 2\eta^2(F^T z_t)] = \mathbb{E}_{\pi_t}[(2\cdot\eta^2||F||^2\cdot 2\eta^2F)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[2\cdot\eta^2||F||^2\cdot 2\eta^2F^j]\mathbb{E}_{\pi_t}[z_t^j] = 0.$

16. $$\mathbb{E}_{\pi_t}[2\cdot\eta^2||z_t||^2\cdot 2\eta(E^TF)] = 4\eta^3\mathbb{E}_{\pi_t}[||z_t||^2]\mathbb{E}_{\pi_t}[(E^TF)] \le 4\eta^3 d\sigma^2\sqrt{\mathbb{E}_{\pi_t}[||E||^2]\mathbb{E}_{\pi_t}[||F||^2]}$$
$$\le 4\eta^3 d\sigma^2 b_t^{\frac{1}{4}}(2C_1\sqrt{b_t}+2C_2^2)^{\frac{1}{2}} \le 4\sqrt{2}\eta^3 d\sigma^2 C_1^{\frac{1}{2}}b_t^{\frac{1}{2}} + 4\sqrt{2}\eta^3 d\sigma^2 C_2 b_t^{\frac{1}{4}}$$
where the first inequality is by Cauchy-Schwarz over the probability space. The second inequality is by Eq (20) and Jensen's inequality. The third inequality is by subadditivity.

17. $$\mathbb{E}_{\pi_t}[2\cdot\eta^2||z_t||^2\cdot 2\eta(E^Tz_t)] \le \mathbb{E}_{\pi_t}[4\eta^3||z_t||^3||E||] = 4\cdot\eta^3\mathbb{E}_{\pi_t}[||z_t||^3]\mathbb{E}_{\pi_t}[||E||] \le 4\eta^3 d^{\frac{3}{2}}\sigma^3 b_t^{\frac{1}{4}},$$
where the first inequality is by Cauchy-Schwarz over $\mathbb{R}^d$. The first equality is by the independence of random variables. The second inequality is by Jensen's inequality.

18. $$\mathbb{E}_{\pi_t}[2\cdot\eta^2||z_t||^2\cdot 2\eta^2(F^Tz_t)] \le \mathbb{E}_{\pi_t}[4\eta^3||z_t||^3||F||] = 4\cdot\eta^3\mathbb{E}_{\pi_t}[||z_t||^3]\mathbb{E}_{\pi_t}[||F||]$$
$$\le 4\eta^3 d^{\frac{3}{2}}\sigma^3\cdot\sqrt{2C_1\sqrt{b_t}+2C_2^2} \le 4\sqrt{2}\eta^3 d^{\frac{3}{2}}\sigma^3 C_1^{\frac{1}{2}}b_1^{\frac{1}{4}} + 4\sqrt{2}\eta^3 d^{\frac{3}{2}}\sigma^3 C_2, \tag{23}$$
where the first inequality is by Cauchy-Schwarz over $\mathbb{R}^d$. The first equality is by the independence of random variables. The second inequality is by the Jensen's inequality and Eq (20). The third inequality is by subadditivity of square root function.

19. $\mathbb{E}_{\pi_t}[2\cdot 2\eta(E^TF)\cdot 2\eta(E^Tz_t)] = \mathbb{E}_{\pi_t}[(2\cdot 2\eta(E^TF)\cdot 2\eta\cdot E)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[(2\cdot 2\eta(E^TF)\cdot 2\eta\cdot E^j\cdot z_t^j] = 0.$

20. $\mathbb{E}_{\pi_t}[2\cdot 2\eta(E^TF)\cdot 2\eta^2(F^Tz_t)] = \mathbb{E}_{\pi_t}[(2\cdot 2\eta(E^TF)\cdot 2\eta^2\cdot F)^T z_t] = \sum_{j=1}^d \mathbb{E}_{\pi_t}[2\cdot 2\eta(E^TF)\cdot 2\eta^2\cdot F^j\cdot z_t^j] = 0$

21. $$\mathbb{E}_{\pi_t}[2\cdot 2\eta(E^Tz_t)\cdot 2\eta^2(F^Tz_t)]] \le \mathbb{E}_{\pi_t}[8\eta^3||E||||F||||z_t||^2]$$
$$\le 8\eta^3\sqrt{\mathbb{E}_{\pi_t}[||E||^2]}\sqrt{\mathbb{E}_{\pi_t}[||F||^2]}\cdot d\sigma^2$$
$$\le 8\eta^3\cdot d\sigma^2\cdot b_t^{\frac{1}{4}}\sqrt{2C_1\sqrt{b_t}+2C_2^2}$$
$$\le 8\sqrt{2}\eta^3 d\sigma^2 C_1^{\frac{1}{2}}b_t^{\frac{1}{2}} + 8\sqrt{2}\eta^3 d\sigma^2 C_2 b_t^{\frac{1}{4}}$$
where the first inequality is by Cauchy-Schwarz over $\mathbb{R}^d$. The second inequality is by Cauchy-Schwarz over the probability space. The third inequality is using Jensen's inequality and Eq (9). The fourth inequality is by subadditivity of square root function.

By combining Eq (17) and the above term analysis, we obtain that

$$\begin{aligned}
b_{t+1} \le\ & b_t + \eta^4(8C_1^4b_t + 8C_6) + \eta^4\sigma^4(d^2+2d) + (8\sqrt{2}\eta^2C_1^2b_t + 8\sqrt{2}\eta^2C_6^{\frac{1}{2}}b_t^{\frac{1}{2}}) \\
& + 4\eta^2\cdot\sqrt{b_t}\cdot d\sigma^2 + 4\eta^4\cdot(2C_1\sqrt{b_t}+2C_2^2)\cdot d\sigma^2 \\
& + (4\sqrt{2}\eta^2C_1^2b_t + 4\sqrt{2}\eta^2C_6^{\frac{1}{2}}b_t^{\frac{1}{2}}) + 2\eta^2\cdot\sqrt{b_t}\cdot d\sigma^2 + 4\eta(-C_3b_t + b_t^{\frac{3}{4}}\cdot C_6^{\frac{1}{4}}) + 0 + 0 \\
& + 2\eta^4 d\sigma^2(2C_1\sqrt{b_t}+2C_2^2) + (16\sqrt[4]{2}\eta^3C_1^3b_t + 16\sqrt[4]{2}\eta^3C_6^{\frac{3}{4}}b_t^{\frac{1}{4}}) + 0 + 0 \\
& + (4\sqrt{2}\eta^3 d\sigma^2C_1^{\frac{1}{2}}b_t^{\frac{1}{2}} + 4\sqrt{2}\eta^3 d\sigma^2C_2b_t^{\frac{1}{4}}) + 4\eta^3 d^{\frac{3}{2}}\sigma^3 b_t^{\frac{1}{4}} + (4\sqrt{2}\eta^3 d^{\frac{3}{2}}\sigma^3C_1^{\frac{1}{2}}b_1^{\frac{1}{4}} + 4\sqrt{2}\eta^3 d^{\frac{3}{2}}\sigma^3C_2) \\
& + 0 + 0 \\
& + (8\sqrt{2}\eta^3 d\sigma^2C_1^{\frac{1}{2}}b_t^{\frac{1}{2}} + 8\sqrt{2}\eta^3 d\sigma^2C_2b_t^{\frac{1}{4}})
\end{aligned} \tag{24}$$

Since for any two positive real numbers $a, b$, $(a+b)^{\frac{1}{2}} \leq a^{\frac{1}{2}} + b^{\frac{1}{2}}$ and $(a+b)^{\frac{1}{4}} \leq a^{\frac{1}{4}} + b^{\frac{1}{4}}$, we have that

$$b_{t+1} \leq Hb_t + Ib_t^{\frac{3}{4}} + Jb_t^{\frac{1}{2}} + Kb_t^{\frac{1}{4}} + L \tag{25}$$

where

$$\begin{aligned}
H &= 1 - \eta 2^2 C_3 + \eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2) + \eta^4(2^{\frac{17}{4}}C_1^3 + 2^3 C_1^4) \\
I &= \eta 2^2 C_6^{\frac{1}{4}} + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1^2 C_6^{\frac{1}{4}} \\
J &= \eta^2 2^2 d\sigma^2 + \eta^2 2 d\sigma^2 + \eta^4 2^3 d\sigma^2 C_1 + \eta^4 2^2 d\sigma^2 C_1 + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_1^{\frac{1}{2}} + \eta^2 2^4 C_6^{\frac{1}{2}} + \eta^2 32^{\frac{1}{2}} C_6^{\frac{1}{2}} \\
&\quad + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1 C_6^{\frac{1}{2}} + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_1^{\frac{1}{2}} \\
K &= \eta^3 2^2 \sigma^3 d^{\frac{3}{2}} + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_1^{\frac{1}{2}} + \eta^4 2^{\frac{17}{4}} C_6^{\frac{3}{4}} + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_2 + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_2 \\
L &= \eta^4 8 C_6 + \eta^4 \sigma^4 (d^2 + 2d) + \eta^4 8 C_2^2 d\sigma^2 + \eta^4 4 C_2^2 d\sigma^2 + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_2
\end{aligned}$$

□

**Lemma 18.** *When* $0 < \eta < \min\{\frac{1}{4C_3}, \frac{C_3}{16(C_1^2+C_1^3+C_1^4)}, 1\}$*, it is true that* $0 < H < 1$ *and* $I, J, K, L > 0$.

*Proof.* First, since $\eta < \frac{1}{4C_3}$,

$$H > 1 - \frac{2^2 C_3}{4C_3} + \eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2) + \eta^4(2^{\frac{17}{4}}C_1^3 + 2^3 C_1^4) = \eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2) + \eta^4(2^3 C_1^4 + 2^{\frac{17}{4}}C_1^3) > 0.$$

Furthermore, since $\eta < \frac{C_3}{16(C_1^2+C_1^3+C_1^4)}$ and $\eta < 1$,

$$\begin{aligned}
\eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2) + \eta^4(2^{\frac{17}{4}}C_1^3 + 2^3 C_1^4) &< \eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2 + 2^{\frac{17}{4}}C_1^3 + 2^3 C_1^4) \\
&< \eta^2(2^6 C_1^2 + 2^6 C_1^3 + 2^6 C_1^4) \\
&< \frac{\eta C_3}{16(C_1^2 + C_1^3 + C_1^4)}(2^6 C_1^2 + 2^6 C_1^3 + 2^6 C_1^4) \\
&< \eta 4 C_3
\end{aligned}$$

and so

$$H = 1 - \eta 4 C_3 + \eta^2(2^{\frac{9}{2}}C_1^2 + 2^{\frac{5}{2}}C_1^2) + \eta^4(2^{\frac{17}{4}}C_1^3 + 2^3 C_1^4) < 1.$$

Since $\eta > 0$, $I, J, K, L > 0$. □

We define three functions

$$l_1(y) = y, \tag{26}$$

$$l_2(y) = Hy + Iy^{\frac{3}{4}} + Jy^{\frac{1}{2}} + Ky^{\frac{1}{4}} + L, \tag{27}$$

$$l(y) = l_1(y) - l_2(y) = (1-H)y - Iy^{\frac{3}{4}} - Jy^{\frac{1}{2}} - Ky^{\frac{1}{4}} - L. \tag{28}$$

We define a new sequence $(y_t)$ where $y_0 = b_0$ and

$$y_{t+1} = l_2(y_t).$$

By Lemma 17 and induction, we have the following result.

**Corollary 19.** *For each* $t \geq 1$*,* $y_t \geq b_t$.

**Lemma 20.** $l(y)$ *has a unique nonnegative zero. If we denote that zero by* $y_*$*, then* $\lim_{t\to\infty} y_t = y_*$.

*Proof.* It is sufficient to show that $l_1(y)$ and $l_2(y)$ have a unique intersection when $y$ is nonnegative. Since $l_2$ is concave, $l_2'(y)$ is decreasing. Since $l(0) < 0$ and $\lim_{y\to\infty} l(y) = \infty$, $l_1, l_2$ has at least one intersection $y_*$. It suffices to show that for any $y > y_*$, $l_1(y) > l_2(y)$ and for any $y < y_*$, $l_1(y) < l_2(y)$. First, we claim that $l_1'(y_*) > l_2'(y_*)$. Assume $l_2'(y_*) \geq l_1'(y_*)$. Since $l_2'(y)$ is strictly decreasing, for any $y < y_*$, $l_2'(y) > l_2'(y_*) > l_1'(y_*) = l_1'(y)$. Thus, $l_2(y_*) = \int_0^{y_*} l_2'(t)\ dt > \int_0^{y_*} l_1'(t)\ dt = l_1(y_*)$, contradicting that $y_*$ is an intersection. Since $l_2'(y)$ is strictly decreasing and $l_1'(y) = 1$, for any $y > y_*$, $l_1'(y) > l_2'(y)$.

Thus, for any $y > y_*$, $l_2(y) = y_* + \int_{y_*}^{y} l_2'(t)\ dt < y_* + \int_{y_*}^{y} l_1'(t)\ dt = l_1(y)$. Since $l_2'(0) > l_1'(0)$, there exists $0 < \tilde{y} < y_*$ such that $l_2'(\tilde{y}) = l_1'(\tilde{y})$. For any $0 \le y \le \tilde{y}$, $l_2(y) = l_2(0) + \int_{t=0}^{y} l_2'(t)dt > l_1(0) + \int_{t=0}^{y} l_1'(t)dt = l_1(y)$. For any $\tilde{y} < y < y_*$, $l_2(y) = l_2(y_*) - \int_{t=y_*}^{y} l_2'(t)dt > l_1(y) + \int_{t=0}^{y_*} l_1'(t)dt = l_1(y)$. Therefore, $y_*$ is the unique zero.

We now show that for any $y_0 \ge 0$, the sequence $(y_t)$ converges to $y_*$. If $y_0 = y_*$ then $y_t = y_*$ for all $t$ since $y_*$ is the stationary point. If $y_0 \ge y_*$, we show by induction that $y_{t+1} \ge y_*$. Assume $y_t \ge y_*$ then $y_{t+1} - y_* = l_2(y_t) - l_2(y_*) = H(y_t - y_*) + I(y_t^{\frac{3}{4}} - y_*^{\frac{3}{4}}) + J(y_t^{\frac{1}{2}} - y_*^{\frac{1}{2}}) + K(y_t^{\frac{1}{4}} - y_*^{\frac{1}{4}}) \ge 0$. Thus, $y_t \ge y_*, \forall t$. Also, since $y_t \ge y_*$, $l(y_t) > 0$. Thus, $y_{t+1} - y_t = l_2(y_t) - y_t = -l(y_t) < 0$. Therefore, $y_{t+1} < y_t, \forall t$. Therefore, the sequence $(y_t)$ is decreasing and bounded so must converge.

Assume $\lim_{t\to\infty} y_t = y_\infty$ so $l_2(y_\infty) = y_\infty = l_1(y_\infty)$. Since $y_*$ is the unique zero, $y_\infty = y_*$. The proof is similar if $y_0 < y_*$. We first show by induction that $y_{t+1} \le y_*$. Assume $y_t \le y_*$, $y_{t+1} - y_* = l_2(y_t) - l_2(y_*) = H(y_t - y_*) + I(y_t^{\frac{3}{4}} - y_*^{\frac{3}{4}}) + J(y_t^{\frac{1}{2}} - y_*^{\frac{1}{2}}) + K(y_t^{\frac{1}{4}} - y_*^{\frac{1}{4}}) \le 0$. Thus, $y_{\le} y_*, \forall t$. Also since $y_t \le y_*$, $f(y_t) < 0$. Thus, $y_{t+1} - y_t = l_2(y_t) - y_t = -l(y_t) > 0$. Therefore, $y_{t+1} > y_t, \forall t$. Therefore, the sequence $(y_t)$ is increasing and bounded so must converges. Assume $\lim_{t\to\infty} y_t = y_\infty$ so $l_2(y_\infty) = y_\infty = l_1(y_\infty)$. Since $y_*$ is the unique zero, $y_\infty = y_*$.

□

**Lemma 21.** *Let $y_*$ be the unique nonnegative zero of $l(y)$, then*

$$y_* \le \max\{(\frac{I+J+K+L}{1-H})^4, (\frac{(I+J)+\sqrt{(1-H)(K+L)}}{(1-H)})^2\}. \tag{29}$$

*Proof.* We now give an upper bound for $y_*$. Let $l_3(y) = (1-H)y - (I+J+K+L)y^{\frac{3}{4}}$ and $l_4(y) = (1-H)y - (I+J)y^{\frac{1}{2}} - (K+L)$. For any $0 \le y \le 1$, $l(y) \ge l_4(y)$ and for any $y > 1$, $l(y) \ge l_3(y)$. Thus, for any $y > 0$, $l(y) \ge \min\{l_3(y), l_4(y)\}$. We observe that $l_3(y)$ has a unique zero $z_3 = (\frac{I+J+K+L}{1-H})^4$ and for any $y < z_3$, $l_3(y) < 0$ and for any $y > z_3$, $l_3(y) > 0$. Similarly, $l_4(y)$ has a unique zero $z_4 = (\frac{(I+J)+\sqrt{(I+J)^2+4(1-H)(K+L)}}{2(1-H)})^2 \le (\frac{2(I+J)+\sqrt{4(1-H)(K+L)}}{2(1-H)})^2$ such that for any $y < z_4$, $l_4(y) < 0$ and for any $y > z_4$, $l_4(y) > 0$. If $l(y_*) \ge l_3(y_*)$ then $y_* \le z_3$ and if $l(y_*) \ge l_4(y_*)$ then $y_* \le z_4$. Therefore, $y_* \le \max\{z_3, z_4\}$.

□

**Lemma 22.**

$$\limsup_{t\to\infty} b_t \le \max\{(\frac{I+J+K+L}{1-H})^4, (\frac{(I+J)+\sqrt{(1-H)(K+L)}}{(1-H)})^2\} \tag{30}$$

*Proof.* By Corollary 19, Lemma 20, and Lemma 21,

$$\limsup_{t\to\infty} b_t \le \lim_{t\to\infty} y_t \le y_* \le \max\{(\frac{I+J+K+L}{1-H})^4, (\frac{(I+J)+\sqrt{(1-H)(K+L)}}{(1-H)})^2\}.$$

□

When $\eta$ approaches to 0 and $d$ approaches to infinity, we observe that

$$\begin{aligned}
H &= 1 - \eta 2^2 C_3 + \eta^4 2^3 C_1^4 + \eta^2 2^{\frac{9}{2}} C_1^2 + \eta^2 2^{\frac{5}{2}} C_1^2 + \eta^4 2^{\frac{17}{4}} C_1^3 = O(\eta),\\
I &= \eta 2^2 C_6^{\frac{1}{4}} + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1^2 C_6^{\frac{1}{4}} = O(\eta),\\
J &= \eta^2 2^2 d\sigma^2 + \eta^2 2 d\sigma^2 + \eta^4 2^3 d\sigma^2 C_1 + \eta^4 2^2 d\sigma^2 C_1 + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_1^{\frac{1}{2}} + \eta^2 2^4 C_6^{\frac{1}{2}} + \eta^2 32^{\frac{1}{2}} C_6^{\frac{1}{2}}\\
&\quad + \eta^4 \cdot 3 \cdot 2^{\frac{17}{4}} C_1 C_6^{\frac{1}{2}} + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_1^{\frac{1}{2}} = O(\eta^2 d),\\
K &= \eta^3 2^2 \sigma^3 d^{\frac{3}{2}} + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_1^{\frac{1}{2}} + \eta^4 2^{\frac{17}{4}} C_6^{\frac{3}{4}} + \eta^3 2^{\frac{5}{2}} d\sigma^2 C_2 + \eta^3 2^{\frac{7}{2}} d\sigma^2 C_2 = O(\eta^3 d^{\frac{3}{2}}),\\
L &= \eta^4 8 C_6 + \eta^4 \sigma^4 (d^2 + 2d) + \eta^4 8 C_2^2 d\sigma^2 + \eta^4 4 C_2^2 d\sigma^2 + \eta^3 2^{\frac{5}{2}} d^{\frac{3}{2}} \sigma^3 C_2 = O(\eta^4 d^2 + \eta^3 d^{\frac{3}{2}}).
\end{aligned}$$

Thus,

$$\begin{aligned}(\frac{I+J+K+L}{1-H})^4 &= O((\frac{\eta+\eta^3 d^{\frac{3}{2}}+\eta^2 d+\eta^4 d^2+\eta^3 d^{\frac{3}{2}}}{\eta})^4)\\ &= O((1+\eta^2 d^{\frac{3}{2}}+\eta d+\eta^3 d^2+\eta^2 d^{\frac{3}{2}})^4)\\ &= O(1+\eta^8 d^6+\eta^4 d^4+\eta^{12} d^8+\eta^8 d^6)\\ &= O(1+\eta^8 d^6+\eta^4 d^4+\eta^{12} d^8),\end{aligned}$$

and

$$\begin{aligned}(\frac{(I+J)+\sqrt{(1-H)(K+L)}}{(1-H)})^2 &= O((\frac{\eta+\eta^2 d+\sqrt{\eta(\eta^3 d^{\frac{3}{2}}+\eta^4 d^2)}}{\eta})^2)\\ &= O((1+\eta d+\eta\sqrt{d^{\frac{3}{2}}+\eta d^2})^2)\\ &= O(1+\eta^2 d^2+\eta^2(d^{\frac{3}{2}}+\eta d^2))\\ &= O(1+\eta^2 d^2).\end{aligned}$$

Therefore,

$$\begin{aligned}\limsup_{t\to\infty} b_t &= \max\{O(1+\eta^8 d^6+\eta^4 d^4+\eta^{12} d^8), O(1+\eta^2 d^2)\}\\ &= O(1+\eta^2 d^2+\eta^4 d^4+\eta^8 d^6+\eta^{12} d^8)\end{aligned} \tag{31}$$

Before we begin our proof on an upper bound on $\mathbb{E}_{\pi_{t-1}}||f(x_t)-f(\tilde{x}_t)||^2$, we first state a result on the powers of gradient of $f$.

**Lemma 23.** *If each function in $\{f_i\}$ satisfies Assumption 5 and Assumption 7, then for each $x\in\mathbb{R}^d$, $||f(x)||^2<C_2$ and $||f(x)||^4<C_6$.*

*Proof.* First,

$$||f(x)||^2=||\frac{1}{n}\sum_{i=1}^n f_i(x)||^2=||\sum_{i=1}^n \frac{1}{n}f_i(x)||^2=||\mathbb{E}_{g\sim\{f_i\}_i}[g(x)]||^2\le\mathbb{E}_{g\sim\{f_i\}_i}[||g(x)||^2]<C_2$$

where the first inequality is by Jensen's inequality. Second,

$$||f(x)||^4=||\frac{1}{n}\sum_{i=1}^n f_i(x)||^4=||\sum_{i=1}^n \frac{1}{n}f_i(x)||^4=||\mathbb{E}_{g\sim\{f_i\}_i}[g(x)]||^4\le\mathbb{E}_{g\sim\{f_i\}_i}[||g(x)||^4]<C_6$$

where the first inequality is by Jensen's inequality. □

We now give a bound on $\mathbb{E}_{\pi_{t-1}}||f(x_t)-f(\tilde{x}_t)||^2$. By Assumption 4 and Taylor expansion, we observe that

$$f(\tilde{x}_t)\le f(x_t)+\nabla f(x_t)^T(\tilde{x}_t-x_t)+\frac{C_1}{2}||x_t-\tilde{x}_t||^2.$$

Combining

$$-f(\tilde{x}_t)\le -f(x_t)-\nabla f(x_t)^T(\tilde{x}_t-x_t)+\frac{C_1}{2}||x_t-\tilde{x}_t||^2,$$

we obtain that

$$\begin{aligned}|f(\tilde{x}_t)-f(x_t)|^2 &\le (|\nabla f(x_t)^T(\tilde{x}_t-x_t)|+\frac{C_1}{2}||x_t-\tilde{x}_t||^2)^2\\ &\le (||\nabla f(x_t)||||(\tilde{x}_t-x_t)||+\frac{C_1}{2}||x_t-\tilde{x}_t||^2)^2\\ &\le ||\nabla f(x_t)||^2||(\tilde{x}_t-x_t)||^2+\frac{C_1^2}{4}||x_t-\tilde{x}_t||^4+C_1||\nabla f(x_t)||||x_t-\tilde{x}_t||^3\end{aligned}$$

which implies

$$\begin{aligned}
\mathbb{E}_{\pi_{t-1}}||f(x_t)-f(\tilde{x}_t)||^2 &\le \mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||^2||(\tilde{x}_t-x_t)||^2]+\frac{C_1^2}{4}\mathbb{E}_{\pi_{t-1}}[||x_t-\tilde{x}_t||^4]+C_1\mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||||x_t-\tilde{x}_t||^3]\\
&\le \sqrt{\mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||^4]}\sqrt{\mathbb{E}_{\pi_{t-1}}[||\tilde{x}_t-x_t||^4]}+\frac{C_1^2}{4}\mathbb{E}_{\pi_{t-1}}[||x_t-\tilde{x}_t||^4]\\
&\quad+(\mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||^4])^{\frac{1}{4}}(\mathbb{E}_{\pi_{t-1}}[||x_t-\tilde{x}_t||^4])^{\frac{3}{4}}\\
&= \sqrt{\mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||^4]}b_t^{\frac{1}{2}}+\frac{C_1^2}{4}b_t+(\mathbb{E}_{\pi_{t-1}}[||\nabla f(x_t)||^4])^{\frac{1}{4}}b_t^{\frac{3}{4}}\\
&\le C_6^{\frac{1}{2}}b_t^{\frac{1}{2}}+\frac{C_1^2}{4}b_t+C_6^{\frac{1}{4}}b_t^{\frac{3}{4}}.
\end{aligned}$$

By Eq (31),

$$\begin{aligned}
\limsup_{t\to\infty}\mathbb{E}_{\pi_{t-1}}||f(x_t)-f(\tilde{x}_t)||^2 &= O(1+\eta d+\eta^2d^2+\eta^4d^3+\eta^6d^4\\
&\quad+1+\eta^2d^2+\eta^4d^4+\eta^8d^6+\eta^{12}d^8\\
&\quad+1+\eta^{\frac{3}{2}}d^{\frac{3}{2}}+\eta^3d^3+\eta^6d^{\frac{9}{2}}+\eta^9d^6)\\
&= O(1+\eta d+\eta^{\frac{3}{2}}d^{\frac{3}{2}}+\eta^2d^2+\eta^3d^3+\eta^4d^4+\eta^6d^{\frac{9}{2}}+\eta^8d^6+\eta^{12}d^8).
\end{aligned}\tag{32}$$

Finally,

$$\begin{aligned}
\limsup_{t\to\infty}W_2(f(\tilde{x}_t),f(x_t)) &= \limsup_{t\to\infty}\inf_{\tilde{\pi}_t\in\text{Coup}(x_t,\tilde{x}_t)}\{(\mathbb{E}_{(x_t,\tilde{x}_t)\sim\tilde{\pi}_t}[||f(x_t)-f(\tilde{x}_t)||^2])^{\frac{1}{2}}\}\\
&\le \limsup_{t\to\infty}(\mathbb{E}_{(x_t,\tilde{x}_t)\sim\pi_t}[||f(x_t)-f(\tilde{x}_t)||^2])^{\frac{1}{2}}\\
&= (\limsup_{t\to\infty}\mathbb{E}_{(x_t,\tilde{x}_t)\sim\pi_t}[||f(x_t)-f(\tilde{x}_t)||^2])^{\frac{1}{2}}\\
&= O(1+\eta^{\frac{1}{2}}d^{\frac{1}{2}}+\eta^{\frac{3}{4}}d^{\frac{3}{4}}+\eta d+\eta^{\frac{3}{2}}d^{\frac{3}{2}}+\eta^2d^2+\eta^3d^{\frac{9}{4}}+\eta^4d^3+\eta^6d^4)
\end{aligned}$$

where $\text{Coup}(x_{t^*},\tilde{x}_{t^*})$ is the set of couplings between $x_{t^*},\tilde{x}_{t^*}$.

### B.2 Proof of Theorem 11

There is an existing result on the convergence of $W_2(f(x_t),f^*)$:

**Lemma 24** ([7]). *Under Assumptions 4–9 and appropriate step size conditions:*

$$\lim_{t\to\infty}W_2(f(x_t),f^*)=O(\eta)$$

Combining Lemma 10 and Lemma 24,

$$\begin{aligned}
\limsup_{t\to\infty}W_2(f(\tilde{x}_t),f^*) &\le \limsup_{t\to\infty}W_2(f(\tilde{x}_t),f(x_t))+\lim_{t\to\infty}W_2(f(x_t),f^*)\\
&= O(1+\eta^{\frac{1}{2}}d^{\frac{1}{2}}+\eta^{\frac{3}{4}}d^{\frac{3}{4}}+\eta d+\eta^{\frac{3}{2}}d^{\frac{3}{2}}+\eta^2d^2+\eta^3d^{\frac{9}{4}}+\eta^4d^3+\eta^6d^4)+O(\eta)\\
&= O(1+\eta^{\frac{1}{2}}d^{\frac{1}{2}}+\eta^{\frac{3}{4}}d^{\frac{3}{4}}+\eta d+\eta^{\frac{3}{2}}d^{\frac{3}{2}}+\eta^2d^2+\eta^3d^{\frac{9}{4}}+\eta^4d^3+\eta^6d^4)
\end{aligned}$$