Continuous-Time Analysis of Federated Averaging

Tom Overman [1] **Diego Klabjan** [2]

Abstract

Federated averaging (FedAvg) is a popular algorithm for horizontal federated learning (FL), where samples are gathered across different clients and are not shared with each other or a central server. Extensive convergence analysis of FedAvg exists for the discrete iteration setting, guaranteeing convergence for a range of loss functions and varying levels of data heterogeneity. We extend this analysis to the continuous-time setting where the global weights evolve according to a multivariate stochastic differential equation (SDE), which is the first time FedAvg has been studied from the continuous-time perspective. We use techniques from stochastic processes to establish convergence guarantees under different loss functions, some of which are more general than existing work in the discrete setting. We also provide conditions for which FedAvg updates to the server weights can be approximated as normal random variables. Finally, we use the continuous-time formulation to reveal generalization properties of FedAvg.

1. Introduction

Federated learning (FL) is a popular privacy-preserving machine learning framework allowing a server to train a central model without accessing data locked on clients. In FL, each client holds a subset of the overall data and is not willing to share this data with the server; however, the clients can send model weights to the server. In horizontal FL studied herein, each client holds a subset of samples but each client has access to all features. A popular algorithm for horizontal FL is FedAvg (McMahan et al., 2017), which is the focus in this work.

FedAvg works by averaging the model weights of all clients periodically. Specifically, each client trains a local model for a specific number of iterations using only the data on the client. Typically, this local training is done using batch stochastic gradient descent. Then, the clients send their local model weights to the server, where the server averages the model weights and sends them back to the clients. Understanding the conditions for FedAvg to converge and what factors influence convergence and generalization is of great interest to federated learning practitioners and has been the focus of significant work (Li et al., 2019; Karimireddy et al., 2020).

A major challenge in FL is dealing with non-IID data across clients (heterogeneous setting). The local updates on each client significantly diverge due to the different data distributions they are trained on, resulting in slower convergence. Many improvements to FedAvg exist, such as SCAFFOLD (Karimireddy et al., 2020), which reduces the impact of the data drift during local training and speeds up training in heterogeneous settings, and RADFed (Xue et al., 2022), which handles non-IID data by delaying aggregation with specialized redistribution rounds. Despite the abundance of new algorithms being introduced to tackle new challenges in FL, FedAvg remains an important algorithm that is the foundation of most new approaches.

The continuous-time perspective of stochastic gradient descent (SGD) has provided a framework for developing more compact proofs of convergence than the discrete process (Orvieto & Lucchi, 2019) and has allowed the study of generalization properties. Specifically, it has been shown through the continuous-time representation of SGD how the learning rate and batch size influence the width of local minima converged to during SGD (Jastrzebski et al., 2018). This local minima width impacts the model's generalization ability to unseen data (Keskar et al., 2017). The continuous-time representation also allows for studying first exit times of SGD from local minima and how this depends on the width of the local minima (Nguyen et al., 2019). Our work is focused on developing a continuous-time representation of FedAvg. It is our hope that this formulation provides a framework for new interesting analyses, just as the continuous-time representation did for SGD.

We focus on a theoretical analysis of FedAvg. We formulate a continuous-time representation of FedAvg in the form of

[1]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, USA [2]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA. Correspondence to: Tom Overman <tomoverman2025@u.northwestern.edu>.

a stochastic differential equation (SDE) and use this formulation to prove convergence properties. The convergence proofs are relatively compact and the proof framework may be extended to other FL algorithms. We show convergence of FedAvg to a stationary point for general, non-convex loss functions and demonstrate that it is likely not sufficient for only the server learning rate to decay; it is necessary that the client-side learning rate must decay at certain rates. We show convergence of FedAvg to the global minimum for weakly quasi-convex loss functions. To the best of our knowledge, weak quasi-convexity has not been studied for FedAvg up to this point. Next, we show that the server weight updates in FedAvg can be approximated as normal random variables under certain assumptions, even for heterogeneous data with extensive local updates before averaging. This is a surprising result and can assist in further analyses of FL algorithms. Finally, we use our continuous-time approach with a quadratic single variate form of each clients' loss landscape to determine how different FedAvg hyperparameters affect the trade-off between minimizing expected loss and ability to escape poorly-generalizing local minima.

Our contributions are as follows.

1. We are the first to formulate an SDE that models the evolution of *server* weights in continuous time during FedAvg. Existing work (Zhang et al., 2023) for modeling distributed learning algorithms uses ordinary differential equations without stochasticity.
2. Using the continuous-time formulation, we devise convergence proofs of FedAvg in deterministic and stochastic cases. We show convergence under a certain normality assumption to a stationary point for non-convex loss functions and convergence to a global minimum for weakly quasi-convex functions. To the best of our knowledge, no other works in either deterministic or stochastic regimes have shown global convergence of FedAvg for weakly quasi-convex loss functions, which are more general than convex functions and allow for locally concave regions. The new proof framework we provide allows for relatively compact proofs of FedAvg and can inspire compact proofs of other FL algorithms.
3. We show that the server weight updates converge in distribution to a normal distribution as the number of clients grow, even in non-IID data settings with many local client iterations without server averaging. This justifies the normality assumption in the convergence results. We demonstrate this through the Lyapunov central limit theorem.
4. Using a quadratic single variate loss function for each client's loss landscape, we uncover dynamics of how various hyperparameters in FedAvg affect the trade-off between generalization and the optimality gap of the expected loss in the continuous-time setting.

In Section 2, we discuss related work on the analysis of FedAvg in the discrete case and continuous-time analysis of SGD. In Section 3, we formulate the SDE that models the evolution of server weights during FedAvg. In Section 4, we provide convergence guarantees of FedAvg for non-convex and weakly quasi-convex loss functions using the SDE formulation. In Section 5, we provide conditions for which the server weight updates can be approximated as normally distributed; this is a key assumption for the SDE formulation to be an Itô process. In Section 6, we analyze the case where each client's local loss function is a quadratic; this allows us to examine how various parameters, such as the client learning rate and the number of local iterations, affect the trade-off between minimizing the loss function and generalization.

2. Related Work

Extensive work has been done in analyzing stochastic gradient descent (SGD) from the continuous-time perspective. Although theory of the discrete process of stochastic gradient descent has existed for a long time, continuous-time perspectives have added value in understanding the behavior of SGD particularly in the area of generalization (Mandt et al., 2015). Using the continuous-time perspective, it has been shown that learning rate magnitude and batch size affect the sharpness of the local minima that SGD converges to (Jastrzebski et al., 2018), and the sharpness of local minima has been linked to generalization ability (Keskar et al., 2017). Furthermore, there has been work in developing concise convergence proofs for the continuous SDE form of SGD (Orvieto & Lucchi, 2019). While the previously mentioned works form a Brownian motion representation with normally-distributed noise, there is also work on forming Lévy processes where the noise is the more general α-stable distribution that can have heavy tails (Nguyen et al., 2019). This Lévy process is used to describe first-exit times of SGD under heavy-tailed noise and demonstrates why SGD prefers converging to wide local minima which generalize better than narrow minima. All of these works take the actual discrete stochastic gradient descent process and form the continuous-time SDE that models the evolution of the weights over time.

Furthermore, the discrete process of FedAvg is well-studied. Convergence has been proven for the convex case, even when the data distribution across different clients is non-IID (Li et al., 2019). However, the convergence rate is slower when data is highly non-IID and the number of local iterations before averaging is required to be large due to "client drift" (Karimireddy et al., 2020).

The case of analyzing federated learning from the continuous-time perspective is less-studied. There has been work on analyzing the continuous-time system of a broad class of distributed optimization algorithms using a control-theory perspective, however this work assumes full gradients and thus has no stochastic component (Zhang et al., 2023). Furthermore, their framework does not work for the FedAvg algorithm specifically because FedAvg cannot be mapped to the double-feedback system they developed. Since most implementations of FedAvg use stochastic gradients on the clients, we focus on developing and analyzing an SDE that models the evolution of the global weights while assuming that the client updates use noisy, stochastic gradients. Recent work (Mehrjou, 2021) analyzes the continuous-time limit of local client updates to make connections to game theory, but does not form a continuous-time representation of the server weights and does not study convergence of the global weights. Existing work (Glasgow et al., 2022) uses the continuous-time limit of each client's local SGD iterates to uncover iterate bias, but the authors do not form the continuous-time SDE for the evolution of weights on the server, and thus their approach is very different from our approach.

3. Continuous-Time Formulation

We are given N samples with the loss of sample i being $F_i(w)$ and $w \in \mathbb{R}^d$ are the model weights. Furthermore, we assume the samples are gathered across Q clients in the horizontal federated learning fashion. We refer to each local client component of the overall objective function as $F^k(w) = \frac{1}{N_k}\sum_{i\in I_k} F_i(w)$ where I_k is the set of samples that belong to client k and $N_k = |I_k|$. We state the loss function as $F(w) = \sum_{k=1}^{Q} p_k F^k(w)$ where p_k is the weight of client k, usually set to $p_k = \frac{N_k}{N}$. We denote the copy of model weights on client k as w^k. The goal is to solve $\min_w F(w)$.

Client weights are updated using standard SGD, except for iterations where averaging occurs. We can write the evolution of client weights as $w_T^k = w_{T-1}^k - \eta_{k,T-1}\frac{1}{S}\sum_{i\in S_{k,T}} \nabla F_i(w_{T-1}^k)$, when $T \neq T_0 + E$, and $w_T^k = \sum_{\hat{k}=1}^{Q} p_{\hat{k}}\Big(w_{T-E}^{\hat{k}} + \hat{\eta}_{0,T-1}\big(w_{T-1}^{\hat{k}} - \eta_{\hat{k},T-1}\frac{1}{S}\sum_{i\in S_{\hat{k},T}} \nabla F_i(w_{T-1}^{\hat{k}}) - w_{T-E}^{\hat{k}}\big)\Big)$, when $T = T_0 + E$, where S is the batch size, $S_{\hat{k},T}$ is a random batch drawn from the available samples in $I_{\hat{k}}$ at iteration T, T_0 is the most recent iteration where averaging occurred, $\eta_{\hat{k},T-1}$ is the learning rate on client $\hat{k}$ and may vary over iterations, $\hat{\eta}_{0,T-1}$ is the global server learning rate which may vary over iterations, and E is the number of local SGD updates before averaging. When $T = T_0 + E$ and averaging occurs, the most recent iteration of average is incremeted as $T_0 \leftarrow T_0 + E$. It is important to note that the averaging step does not imply that clients have access to the updates of the other clients; this average is actually computed on the server and sent back to each client, and on this time step every client holds the same-valued weights. A common choice of server learning rate is $\hat{\eta}_{0,T-1} = 1$, which simplifies this update schedule to $w_T^k = w_{T-1}^k - \eta_{k,T-1}\frac{1}{S}\sum_{i\in S_{k,T}} \nabla F_i(w_{T-1}^k)$, when $T \neq T_0 + E$, and $w_T^k = \sum_{\hat{k}=1}^{Q} p_{\hat{k}}\Big(w_{T-1}^{\hat{k}} - \eta_{\hat{k},T-1}\frac{1}{S}\sum_{i\in S_{\hat{k},T}} \nabla F_i(w_{T-1}^{\hat{k}})\Big)$, when $T = T_0 + E$.

At iteration $T = T_0 + E$, each client performs one more step of SGD, then sends updated weights to the server, where the server averages the updates along with an optional server-side learning rate $\hat{\eta}_{0,T-1}$ that controls how much to update the server weights. While the mathematical expression for the case when $T = T_0 + E$ shows gradient passing, this is not the case in an actual implementation where it is equivalent to sending the updated weights after each client's local updates. The server then sends this average back to every client. We study the convergence of global server weights in this work, which we denote as $w_{0,T}$ for iteration T. Updates to the server weights only occur every E iterations during the averaging step, otherwise they remain the same as the most recent averaging iteration.

As shown in (Jastrzebski et al., 2018), the stochastic gradients, $\mathcal{G}_T^k$, on the local client updates can be assumed to be normally distributed as follows

$$\mathcal{G}_T^k = \frac{1}{S}\sum_{i\in S_{k,T+1}} \nabla F_i(w_T^k) \sim \mathcal{N}(\nabla F^k(w_T^k), \Sigma_k(w_T^k))$$

where $\Sigma_k(w_T^k) = (\frac{1}{S} - \frac{1}{N_k})\frac{1}{N_k - 1}\sum_{i=1}^{N_k}(\nabla F_i(w_T^k) - \nabla F^k(w_T^k))(\nabla F_i(w_T^k) - \nabla F^k(w_T^k))^T$. We can think of this as a normal random variable centered around the full gradient of the local loss function.

Therefore, we can write our discrete updates in the time region where no averaging occurs as

$$\begin{aligned} w_{T+1}^k &= w_T^k - \eta_{k,T}\mathcal{G}_T^k \\ &= w_T^k - \eta_{k,T}\mathcal{N}(\nabla F^k(w_T^k), \Sigma_k(w_T^k)) \\ &= w_T^k - \eta_{k,T}\nabla F^k(w_T^k) + \eta_{k,T}\mathcal{N}(0, \Sigma_k(w_T^k)). \end{aligned}$$

Expanding this for E time steps, we get

$$w_{T+E}^k = w_T^k + \sum_{i=0}^{E-1} \eta_{k,T+i}(N_{T+i}^k - G_{T+i}^k)$$

where $N_{T+i}^k \sim \mathcal{N}(0, \Sigma_k(w_{T+i}^k))$ and $G_{T+i}^k = \nabla F^k(w_{T+i}^k)$. Notice that the expressions for G_{T+i}^k and

N^k_{T+i} depend on w^k_{T+i} and not w^k_T. This results in the full expression being a very complicated recurrence relation.

We can re-index as

$$w^k_T = w^k_{T-E} + \sum_{i=0}^{E-1} \eta_{k,T-E+i}(N^k_{T-E+i} - G^k_{T-E+i}).$$

We assume that at iteration T (and thus also $T - nE$ for all integers n such that $T - nE \geq 0$), the server aggregates across clients, and we write the aggregated server weights at this time as $w_{0,T}$. Using the averaging technique specified in the FedAvg update schedule we get

$$w_{0,T} = w_{0,T-E} + \hat{\eta}_{0,T} \sum_{k=1}^{Q} p_k [\sum_{i=0}^{E-1} \eta_{k,T-E+i}(N^k_{T-E+i} - G^k_{T-E+i})].$$

We split the global learning rate, $\hat{\eta}_{0,T}$, into the product of a constant term, h, used for lifting the difference equation to continuous-time and the learning rate, $\eta_{0,T}$, which depends on iteration T. We write this as $\hat{\eta}_{0,T} = h\eta_{0,T}$. Thus, we rewrite the difference equation as

$$w_{0,T} = w_{0,T-E} + h\eta_{0,T} \underbrace{\sum_{k=1}^{Q} p_k [\sum_{i=0}^{E-1} \eta_{k,T-E+i}(N^k_{T-E+i} - G^k_{T-E+i})]}_{A_T}.$$

Assumption 3.1. A_T is a normally distributed random variable such that $A_T \sim \mathcal{N}(M_T, V_T)$. Both M_T and V_T are functions of w^k_{T-E}.

We further discuss Assumption 3.1 in Section 5. We can then write the evolution of iterates of the global server weights as

$$w_{0,T} - w_{0,T-E} = h\eta_{0,T} M_T + h\eta_{0,T}\mathcal{N}(0, V_T) \quad (1)$$

where $M_T = \mathbb{E}[A_T]$ and $V_T = \mathbb{E}[A_T A_T^T] - M_T M_T^T$.

According to the FedAvg server update schedule, the global server weights only change every E iterations as indexed by T, thus $w_{0,T} - w_{0,T-E}$ in (1) represents the difference in a single update of the server weights. So, we can view this as the discretization of an SDE using the Euler-Maruyama method with a step size of h. This discretization is more accurate when h is small. We now form the SDE as

$$dw_0(t) = \eta_0(t)\hat{M}(w_0(t))dt + \eta_0(t)\sqrt{h}\hat{V}^{1/2}(w_0(t))dB(t) \quad (2)$$

where $B(t)$ is a standard Brownian motion, $\hat{M}(w_0(t)) = \mathbb{E}[\hat{A}(w_0(t))]$, $\hat{V}^{1/2}(w_0(t))(\hat{V}^{1/2}(w_0(t)))^T = \hat{V}(w_0(t))$, $\hat{V}(w_0(t)) = \mathbb{E}[\hat{A}(w_0(t))\hat{A}(w_0(t))^T] - \hat{M}(w_0(t))\hat{M}(w_0(t))^T$, and rewriting A_T in continuous-time as

$$\hat{A}(w_0(t)) = \sum_{k=1}^{Q} p_k [\sum_{i=0}^{E-1} \eta_k(t)(N^k(t,i) - G^k(t,i)))]$$

where $N^k(t,i) = \mathcal{N}(0, \Sigma_k(w^k(t,i)))$, $G^k(t,i) = \nabla F^k(w^k(t,i))$, $w^k(t,i) = w^k(t,i-1) - \eta_k(t)[G^k(t,i-1) + N^k(t,i-1)]$, and $w^k(t,i=0) = w_0(t)$ for all clients $k = 1, ..., Q$.

We show in Lemma A.4 that the drift term, $\eta_0(t)\hat{M}(w_0(t))$, in (2) is Lipschitz continuous which is an important property for our convergence proofs.

4. Convergence Proofs

We use the formulation of (2) in Section 3, and similar tools used in (Orvieto & Lucchi, 2019) to form convergence proofs for FedAvg in various settings. While the techniques used in convergence proofs for continuous-time SGD are similar to our approaches for FedAvg, the case for FedAvg that we show is much more complicated.

Assumption 4.1. $F_i(w)$ each are μ-smooth functions. We require each $F_i(w)$ to be twice differentiable across their entire domain and $||\nabla F_i(w)||_\infty \leq L$ and $||\text{diag}(H)||_\infty \leq L$ over the entire domain, where H is the Hessian of $F_i(w)$.

Assumption 4.2. The learning rates on each client are the same, may depend on t, and can be written as $\eta_k(t) = \eta(t)$ for all $k = 1, 2, .., Q$.

As a consequence of Assumption 4.2, we rewrite

$$\hat{A}(w_0(t)) = \eta(t) \underbrace{\sum_{k=1}^{Q} p_k [\sum_{i=0}^{E-1} (N^k(t,i) - G^k(t,i)))]}_{\hat{A}_1(w_0(t))} = \eta(t)\hat{A}_1(w_0(t))$$

and

$$\hat{V}(w_0(t)) = \mathbb{E}(\hat{A}(w_0(t))\hat{A}(w_0(t))^T) - \hat{M}(w_0(t))\hat{M}(w_0(t))^T = \eta(t)^2 \Big(\underbrace{\mathbb{E}(\hat{A}_1\hat{A}_1^T) - \mathbb{E}[\hat{A}_1]\mathbb{E}[\hat{A}_1]^T}_{\hat{V}_1(w_0(t))} \Big) = \eta(t)^2 \hat{V}_1(w_0(t)).$$

Assumption 4.3. We assume that the variances of the server updates are bounded for all t. More precisely, $V^* = \max_t ||\hat{V}_1(w_0(t))||_S < \infty$, where $||\cdot||_s$ is the spectral norm.

Assumption 4.4. We assume $\Sigma_k(w^k_T) = \Sigma_k$, where $\Sigma_k \in \mathbb{R}^{d\times d}$ does not vary over iterations and may be different for each client k.

Assumption 4.4 is similar to the assumptions made in the continuous-time SGD literature (Mandt et al., 2015).

4.1. Non-convex loss functions

We prove convergence to a stationary point for general, non-convex loss functions.

Theorem 4.5. *We assume Assumptions 3.1, 4.1, 4.2, 4.3, and 4.4 are met, and the server learning rate $\eta_0(t) = 1$. For a random time point $\tilde{t} \in [0, t]$ that follows the distribution $\frac{\eta(\tilde{t})}{\int_0^t \eta(s)ds}$, we have*

$$\mathbb{E}_{\tilde{t},\mathcal{G}}||\nabla F(w_0(\tilde{t}))||^2 \leq \frac{F(w_0(0)) - F(w_0^*)}{E\varphi(t)} + \frac{1}{E\varphi(t)} \int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2 V^* L}{2}]dt' \quad (3)$$

where $C_1 = \frac{E^2 L\mu \sum_{k=1}^Q p_k[L+\sqrt{Tr(\Sigma_k)}]}{2}$, $\varphi(t) = \int_0^t \eta(t')dt'$, $w_0^* = argmin_{w_0} F(w_0)$, and the expectation $\mathbb{E}_{\tilde{t},\mathcal{G}}$ is over the random time point $\tilde{t}$ and stochasticity in gradients $\mathcal{G}$.

Proof. Proof provided in Section A.3. □

Theorem 4.5 provides a general expression that can be used to easily derive convergence rates for a variety of learning rates, $\eta(t)$. We obtain more concrete convergence rates for specific choices of $\eta(t)$ and asymptotic rates for intervals of $\eta(t)$ in Corollary 4.6.

Corollary 4.6. *For a random time point $\tilde{t} \in [0, t]$ that follows the distribution $\frac{\eta(\tilde{t})}{\int_0^t \eta(s)ds}$, and with $\eta(t) = \frac{1}{t+1}$, we have*

$$\mathbb{E}_{\tilde{t},\mathcal{G}}||\nabla F(w_0(\tilde{t}))||^2 \leq \frac{F(w_0(0)) - F(w_0^*)}{E} \cdot \frac{1}{\log(t+1)} + \frac{C_1 + \frac{hV^*L}{2}}{E} \cdot \frac{1}{\log(t+1)},$$

for $\eta(t) = \frac{1}{\sqrt{t+1}}$, we have

$$\mathbb{E}_{\tilde{t},\mathcal{G}}||\nabla F(w_0(\tilde{t}))||^2 \leq \frac{F(w_0(0)) - F(w_0^*)}{E(2\sqrt{t+1} - 2)} + \frac{[C_1 + \frac{hV^*L}{2}]\log(t+1)}{E(2\sqrt{t+1} - 2)},$$

and with $\eta(t) = 1/(t+1)^b$, we have

$$\mathbb{E}_{\tilde{t},\mathcal{G}}||\nabla F(w_0(\tilde{t}))||^2 \leq \begin{cases} \mathcal{O}(\frac{1}{t^b}) & 0 < b < \frac{1}{2} \\ \mathcal{O}(\frac{\log(t)}{\sqrt{t}}) & b = \frac{1}{2} \\ \mathcal{O}(\frac{1}{t^{1-b}}) & \frac{1}{2} < b < 1 \\ \mathcal{O}(\frac{1}{\log(t)}) & b = 1 \end{cases}.$$

Proof. Proof provided in Section A.4. □

Next, we show in Corollary 4.7 that convergence to a stationary point requires a decreasing learning rate on the clients, $\eta(t)$. Interestingly, a decreasing global server learning rate, $n_0(t)$, by itself is likely not sufficient for convergence. This is because the bound on the expected client drift from Lemma A.1 in the appendix is dependent on the client-side learning rate $\eta(t)$.

Corollary 4.7. *For a random time point $\tilde{t} \in [0, t]$ that follows the distribution $\frac{\eta(\tilde{t})}{\int_0^t \eta(s)ds}$ and with constant client learning rate $\eta(t) = \eta_c$ and decreasing global server-side learning rate $\eta_0(t) = 1/(t+1)$, we have*

$$\mathbb{E}_{\tilde{t},\mathcal{G}}||\nabla F(w_0(\hat{t}))||^2 \leq \frac{F(w_0(0)) - F(w_0^*) + \eta_c^2 hV^* L/2}{E\eta_c \log(t+1)} + \eta_c C_1.$$

Proof. Proof provided in Section A.5. □

Corollary 4.7 shows that despite a decreasing server-side learning rate, FedAvg is only guaranteed to converge to a neighborhood of a stationary point that depends on the size of the constant client-side learning rate $\eta(t) = \eta_c$. The algorithm might still converge to a stationary point since we provide an upper bound.

4.2. Weakly quasi-convex loss functions

We prove convergence to the global minimum for weakly quasi-convex loss functions as defined in Assumption 4.8.

Assumption 4.8. Function F_i is weakly quasi-convex if for some w^* and $\tau > 0$

$$\langle \nabla F_i(w), w - w^* \rangle \geq \tau(F_i(w) - F_i(w^*))$$

holds for every w.

This class of functions is more general than convex functions that are studied in previous discrete analyses of FedAvg. In fact, weakly quasi-convex functions can have locally concave regions, and it has been shown that some non-convex LSTMs induce weakly-quasi convex loss functions (Hardt et al., 2018). To the best of our knowledge, this is the first work to provide convergence results of FedAvg for weakly quasi-convex loss functions.

Theorem 4.9. *We assume Assumptions 3.1, 4.1, 4.2, 4.3, 4.4, and 4.8 are met, and the serving learning rate is a constant $\eta_0(t) = \eta_0$. For a random time point $\tilde{t} \in [0, t]$ that*

follows the distribution $\frac{\eta(\tilde{t})}{\int_0^t \eta(s)ds}$*, we have*

$$\mathbb{E}_{\tilde{t},\mathcal{G}}[(F(w_0(\hat{t})) - F(w_0^*))] \leq \frac{||w_0(0) - w_0^*||}{\tau\varphi(t)} + \frac{C_2}{\tau\varphi(t)} \int_0^t \left[\eta(s)^2 \left(LE \int_0^s \eta(t')dt' + \sqrt{h}V^* \sqrt{\int_0^s \eta(t')^2 dt'}\right)\right] ds + \frac{C_3}{\tau\varphi(t)} \int_0^t \eta(s)^2 ds$$

where $C_2 = \mu E^2 \sum_{k=1}^Q p_k[L + \sqrt{Tr(\Sigma_k)}]$, $C_3 = \left[\frac{dh\eta_0^2 V^*}{2} + \eta_0 C_2 ||w_0(0) - w_0^*||\right]$*, and* w_0^* *is the global minimum point.*

Proof. Proof provided in Section A.6. □

Theorem 4.9 provides a general expression that can be used to easily derive convergence rates for a variety of learning rates, $\eta(t)$. We obtain more concrete convergence rates for a specific choice of $\eta(t)$ in Corollary 4.10.

Corollary 4.10. *If we choose a random time point* $\tilde{t} \in [0, t]$ *that follows the distribution* $\frac{\eta(\tilde{t})}{\int_0^t \eta(s)ds}$*, the serving learning rate is a constant* $\eta_0(t) = \eta_0$*, and for* $\eta(t) = \frac{1}{t+1}$ *we have*

$$\mathbb{E}_{\tilde{t},\mathcal{G}}[(F(w_0(\hat{t})) - F(w_0^*))] \leq \frac{\eta_0^2 C_2 LE}{\tau\eta_0} \cdot \frac{t - \log(t+1)}{t\log(t+1)} + \frac{||w_0(0) - w_0^*|| + C_3 + \eta_0^2 C_2 \sqrt{h} V^*}{\tau\eta_0 \log(t+1)} = \mathcal{O}\left(\frac{1}{\log(t)}\right).$$

Proof. Proof provided in Section A.7. □

5. On the Assumption of Normally Distributed Server Updates

In order for the SDE specified in (2) to be an Itô process, we require A_T to be normally distributed. We show in this section that this is a reasonable assumption in many cases, such as in the regime of a very large number of clients.

Large IID Client Setting Assuming data is independent and identically distributed across clients, the weights p_k should be approximately equal, and assuming bounded variance (Assumption 4.3), we have the conditions met for the traditional central limit theorem. Thus, as the number of clients Q goes to ∞, we have $A_T \xrightarrow{\mathcal{D}} \mathcal{N}(M_T, V_T)$.

Large non-IID Client Setting We now show that we can approximate A_T as a normal random variable under certain conditions even if the data is not identically distributed across clients. The proof hinges on the Lyapunov central limit theorem which generalizes the traditional central limit theorem to cases where random variables are not identically distributed (Billingsley, 2012).

Assumption 5.1. The covariance matrix $\Sigma_k(w_t^k)$ is a diagonal matrix for each client $k = 1, ..., Q$.

Assumption 5.2. We require

$$0 < C \leq \left(p_k^2 \eta_{k,t}^2 \mathbb{E}[(N_{k,i} + R_{k,i})^2]\right)^2 - \frac{1}{2Q} \sum_{j=1}^Q \left(p_k^2 \eta_{k,t}^2 \mathbb{E}[(N_{k,i} + R_{k,i})^2])^2 - p_j^2 \eta_{j,t}^2 \mathbb{E}[(N_{j,i} + R_{j,i})^2])^2\right)^2,$$

holds for all clients $k = 1, ..., Q$ and coordinates $i = 1, .., d$, where $\eta_{k,t}$ is the client-side learning rate on client k. Values $N_{k,i}$ and $R_{k,i}$ are the i-th coordinates of vectors N_k and R_k, respectively, where $N_k = \sum_{i=0}^{E-1} N_{t-E+i}^k$ and $R_k = \sum_{i=0}^{E-1} (\mathbb{E}[G_{t-E+i}^k] - G_{t-E+i}^k)$.

Assumption 5.2 requires that for all clients indexed by k, the square of the second moment of $N_k + R_k$ to be greater than the square of the difference of the second moments of $N_k + R_k$ and $N_j + R_j$ averaged over all clients indexed by j. This essentially requires the second moments to not be too different across different clients.

Assumption 5.3. The fourth-order mixed moments of N_k and R_k are bounded for all clients $k = 1, ..., Q$. More formally, we require $\left|\mathbb{E}[N_{k,i}^u R_{k,i}^v]\right| \leq D$ for all clients $k = 1, ..., Q$, coordinates $i = 1, .., d$, and $0 \leq u, v \leq 4$ with $u + v = 4$.

We note that Assumption 5.3 is guaranteed to hold if N_k and R_k have bounded support, which is usually the case in practice because gradients are typically clipped.

Theorem 5.4. *With assumptions 5.1, 5.2 and 5.3, as the number of clients* Q *goes to* ∞*, we have* $A_T \xrightarrow{\mathcal{D}} \mathcal{N}(M_T, V_T)$.

Proof. The proof uses the Lyapunov Central Limit Theorem, which is more flexible than the standard Central Limit Theorem and allows for non-identically distributed random variables as long as the Lyapunov condition is met. The full proof is provided in Section A.8. □

Section 6 further demonstrates the normality assumption being met when each client's loss landscape is a quadratic form.

6. Analysis in the Quadratic Case

Quadratic Client Loss Landscape We can show that the server updates in FedAvg are normally distributed if we

assume the loss landscape of each client follows a different quadratic form.

Assumption 6.1. Each client's loss landscape is $F^k(w) = \frac{1}{2}(w-a_k)^T U_k (w-a_k)$ for some $U_k \in \mathbb{R}^{d\times d}$ and $a_k \in \mathbb{R}^d$.

With Assumption 6.1, the loss landscape on the server is $F(w) = \frac{1}{2}\sum_{k=1}^{Q} p_k (w-a_k)^T U_k (w-a_k)$. This global loss function can be rewritten as

$$F(w) = \frac{1}{2}\Big(w-a\Big)^T \Big(\sum_{k=1}^{Q} p_k U_k\Big)\Big(w-a\Big),$$

where $a = (\sum_{k=1}^{Q} p_k U_k)^{-1} \sum_{k=1}^{Q} p_k U_k a_k$, which is clearly another quadratic form.

Theorem 6.2. *With Assumption 4.4 and 6.1, a server learning rate of $\eta_0(t) = 1$, and a constant client-side learning rate of $\eta_k(t) = \eta$, the local weight vector on client k after E local updates is*

$$\begin{aligned} w_{t+E}^k \sim w_{0,t} &- \eta \sum_{j=0}^{E-1} (I-\eta U_k)^j U_k (w_{0,t} - a_k) \\ &- \eta \sum_{j=0}^{E-1}\sum_{i=1}^{j} (I-\eta U_k)^{j-i} U_k \mathcal{N}(0,\Sigma_k) \\ &+ E\eta \mathcal{N}(0,\Sigma_k), \end{aligned}$$

where $w_{0,t}$ is the shared server weight vector from the most recent averaging step.

Proof. Proof provided in Section A.9. □

Theorem 6.2 shows that no matter how many local updates are performed on a client, the resulting final update to the weights is normally distributed. Therefore, after averaging, the updates to the global weights are also normally distributed.

Generalization Analysis of Single Variate Quadratic Loss Landscape

Using the evolution of local weights for quadratic loss functions provided in Theorem 6.2, assuming we have the single variate case, and using the same procedure for building the SDE as in Section 3, we obtain the SDE for the global weights as

$$\begin{aligned} dw_0(t) = &-\Big[\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k (w_0(t) - a_k)\Big] dt \\ &+ \sqrt{\eta}\Big[E + \sum_{k=1}^{Q}\sum_{j=0}^{E-1}\sum_{i=1}^{j} p_k \sqrt{\Sigma_k}(1-\eta U_k)^{j-i} U_k\Big] dB(t). \end{aligned} \tag{4}$$

The SDE in (4) is a linear SDE and allows us to find analytical solutions as shown in Theorem 6.3.

Theorem 6.3. *The solution to the SDE specified in (4) is normally distributed as*

$$w_0(t) \sim \mathcal{N}(m_0(t), v_0(t)),$$

where

$$m_0(t) = C_4 + (w_0(0) - C_4) e^{-(\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k)t},$$

$$C_4 = \frac{\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k a_k}{\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k},$$

and

$$\begin{aligned} v_0(t) = &\frac{-\eta\big[E + \sum_{k=1}^{Q}\sum_{j=0}^{E-1}\sum_{i=1}^{j} p_k \sqrt{\Sigma_k}(1-\eta U_k)^{j-i} U_k\big]^2}{2\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k} \\ &\cdot \Big(\exp\big(-(\sum_{k=1}^{Q} p_k \sum_{j=0}^{E-1} (1-\eta U_k)^j U_k)t\big) - 1\Big). \end{aligned}$$

Proof. Proof is provided in Section A.11. □

We are interested in how various hyperparameters influence the expected loss and the variance of the solution distribution shown in Theorem 6.3. In the limit as $\eta \to 0$ and $t \to \infty$, the solution approaches $\mathcal{N}(\frac{\sum_{k=1}^{Q} p_k U_k a_k}{\sum_{k=1}^{Q} p_k U_k}, 0)$ which is the true global minimum with zero variance. This matches the behavior of SGD, that as the learning rate is decreased, the expected loss approaches the true minimum but the variance of the solution decreases and thus may not be able to escape poor local minima (Bishop, 1995; Jastrzebski et al., 2018).

As the number of local iterations before communication, E, grows, the mean of the solution diverges from the true solution. Furthermore, as E grows, the variance of the solution increases as $\mathcal{O}(E^2)$. This causes the expected solution to have a high bias with regards to the true solution, but an increase in E may allow FedAvg to escape poorly-generalizing local minima. This demonstrates the trade-off between expected loss and ability to escape poor local minima.

This reasoning demonstrates how our proposed continuous-time limit of FedAvg can be used to explore behaviors of FL algorithms such as generalization ability. We leave further exploration of these ideas to future work, such as exploring the first exit time of the continuous-time formulation of FedAvg from a local minimum, similar to the existing work in SGD (Nguyen et al., 2019).

7. Conclusion

We have developed a continuous-time SDE that models the evolution of server weights during FedAvg. Using this

formulation, we prove convergence to stationary points for general non-convex loss functions and are the first to show global convergence for weakly quasi-convex loss functions. We discuss the various assumptions required for the SDE to be an Itô process and demonstrate that these assumptions are reasonable. We demonstrate that the updates to the server weights can be approximated as a normal random variable in many cases, even if data is not identically distributed across clients. Finally, we provide a generalization analysis using our continuous-time formulation and quadratic client losses. Our SDE formulation serves as a framework for future efforts in analyzing FedAvg and other FL algorithms from the continuous-time perspective.

Impact Statement

This paper is focused on mathematically analyzing the underlying behavior of federated learning from a new perspective.

References

Billingsley, P. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 2012.

Bishop, C. *Neural Networks for Pattern Recognition*. Advanced Texts in Econometrics. Clarendon Press, 1995.

Eriksson, K., Johnson, C., and Estep, D. *Vector-Valued Functions of Several Real Variables*, pp. 789–814. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

Glasgow, M. R., Yuan, H., and Ma, T. Sharp bounds for federated averaging (local SGD) and continuous perspective. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9050–9090. PMLR, 28–30 Mar 2022.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD, 2018. URL https://arxiv.org/abs/1711.04623.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Krichene, W. and Bartlett, P. L. Acceleration and averaging in stochastic descent dynamics. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations*, 2019.

Mandt, S., Hoffman, M. D., Blei, D. M., et al. Continuous-time limit of stochastic gradient descent revisited. *Advances in Neural Information Processing Systems*, 2015.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Mehrjou, A. Federated learning as a mean-field game, 2021. URL https://arxiv.org/abs/2107.03770.

Mertikopoulos, P. and Staudigl, M. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

Movellan, J. R. Tutorial on stochastic differential equations. *MPLab Tutorials Version*, 6, 2011.

Nguyen, T. H., Simsekli, U., Gurbuzbalaban, M., and Richard, G. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Orvieto, A. and Lucchi, A. Continuous-time models for stochastic optimization algorithms. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Xue, Y., Klabjan, D., and Luo, Y. Aggregation delayed federated learning. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 85–94, 2022.

Zhang, X., Hong, M., and Elia, N. Understanding a class of decentralized and federated optimization algorithms: A multirate feedback control perspective. *SIAM Journal on Optimization*, 33(2):652–683, 2023.

A. Proofs

For a nice review of the mathematical preliminaries of SDEs, see the starting sections of the Supplementary Materials in (Orvieto & Lucchi, 2019).
We introduce a bound on the expected drift between client weights and the server weights. Lemma A.1 is used in the proofs of Theorems 4.5 and 4.9. The proof follows similar ideas as (Li et al., 2019) but has several key differences due to the continuous-time setting.

Lemma A.1. *With Assumptions 4.1, 4.2, 4.4, the expected drift between server global weights and the weights of client k is bounded as*

$$\mathbb{E}||w_0(t) - w^k(t,i)|| \leq i\eta(t)[L + \sqrt{Tr(\Sigma_k)}].$$

A.1. Proof of Lemma A.1

Proof. We first examine the form of $w^k(t,i)$ as

$$w^k(t,i) = w_0(t) + \sum_{j=0}^{i-1} \eta(t)[N^k(t,j) - G^k(t,j)]$$

Now we form the bound

$$\begin{aligned}
\mathbb{E}||w_0(t) - w^k(t,i)|| &= \mathbb{E}||w_0(t) - w_0(t) - \sum_{j=0}^{i-1} \eta[N^k(t,j) - G^k(t,j)]|| \\
&= \mathbb{E}||\sum_{j=0}^{i-1} \eta[G^k(t,j) - N^k(t,j)]|| \\
&\leq \sum_{j=0}^{i-1} \eta(t)[\mathbb{E}||\nabla F^k(w^k(t,j))|| + \mathbb{E}||N^k(t,j)||] \\
&\leq \sum_{j=0}^{i-1} \eta(t)[L + \sqrt{\mathrm{Tr}(\Sigma_k)}] \\
&= i\eta(t)[L + \sqrt{\mathrm{Tr}(\Sigma_k)}].
\end{aligned}$$

□

A.2. Other Lemmas

We first include some helpful lemmas that have been proved in other works or are straightforward to show.

Lemma A.2. *For two symmetric, $d \times d$ matrices A and B, we have the following hold*

$$Tr(AB) \leq d||A||_S||B||_S.$$

Proof. (Mertikopoulos & Staudigl, 2018), (Krichene & Bartlett, 2017), and (Orvieto & Lucchi, 2019) all provide proofs of this statement. □

Lemma A.3. *The probability density function $f(t') = \frac{\eta(t')}{\varphi(t)}$ defined over support $t \in [0,t]$ is a valid probability density function.*

Proof. We define the learning rate $\eta(t)$ as strictly positive, and thus the resulting probability density function is non-negative.

We also show that

$$\begin{aligned}\int_{-\infty}^{\infty} f(t')dt' &= \int_0^t \frac{\eta(t')}{\varphi(t)}dt' \\ &= \frac{1}{\varphi(t)} \underbrace{\int_0^t \eta(t')dt'}_{\varphi(t)} \\ &= 1.\end{aligned}$$

Therefore, the probability density function integrated over its entire support is 1, and the necessary conditions of a probability density function are met. □

Lemma A.4. *The drift term, $\eta_0(t)\hat{M}(w_0(t))$, in Equation 2 is Lipschitz continuous.*

Proof. We require Assumption 4.1. The drift term is a vector-valued function, and it suffices to prove Lipschitz continuity for each component of the function to prove Lipschitz continuity of the whole vector-valued function (Eriksson et al., 2004). Furthermore, it is well known that for everywhere differentiable functions, the function is Lipschitz continuous if and only if the absolute value of the derivative of the function is bounded by a finite value for the entire input domain. We prove this lemma by combining these two facts and proving the bounded derivative, component-wise for $\hat{M}(t)$.

$$\begin{aligned}\hat{M}(w_0(t)) &= \mathbb{E}\Big[\sum_{k=1}^{Q} p_k[\sum_{i=0}^{E-1} \eta_k(t)(N^k(t,i) - G^k(t,i)))]\Big] \\ &= -\sum_{k=1}^{Q} p_k[\sum_{i=0}^{E-1} \eta_k(t)\mathbb{E}[G^k(t,i)]] \\ &= -\sum_{k=1}^{Q} p_k[\sum_{i=0}^{E-1} \eta_k(t)\mathbb{E}[\nabla F^k(w^k(t,i))]]\end{aligned}$$

where $N^k(t,i) = \mathcal{N}(0, \Sigma_k(w^k(t,i)))$, $G^k(t,i) = \nabla F^k(w^k(t,i))$, $w^k(t,i) = w^k(t,i-1) - \eta_k(t)[G^k(t,i-1) + N^k(t,i-1)]$, and $w^k(t,i=0) = w_0(t)$ for all $k = 1, ..., Q$.

We examine at the component-level and index for the j-th component. We use the notation $F'(x) = \frac{d}{dx}F(x)$. We also change $F^k()$ to $F_k()$ for readability when using the $'$ derivative notation. We have

$$\begin{aligned}\hat{M}(t)_j &= -\sum_{k=1}^{Q} p_k[\sum_{i=0}^{E-1} \eta_k(t)\mathbb{E}[\frac{d}{dw_j^k(t,i)}F_k(w_j^k(t,i))]] \\ &= -\sum_{k=1}^{Q} p_k[\sum_{i=0}^{E-1} \eta_k(t)\mathbb{E}[F_k'(w_j^k(t,i))]].\end{aligned}$$

Now we need to show $|\frac{d}{dw_j^k(t,0)}\hat{M}(t)_j| \leq M$ for some finite M. Note, the derivative is with respect to $w_j^k(t,0)$, but the function arguments are $w_j^k(t,i)$, so chain rule needs to be applied consecutively. We have

$$\begin{aligned}\frac{d}{dw_j^k(t,0)}F_k'(w_j^k(t,i)) &= F_k''(w_j^k(t,i))\cdot\frac{dw_j^k(t,i)}{dw_j^k(t,0)}\\ &= F_k''(w_j^k(t,i))\cdot\frac{d}{dw_j^k(t,0)}\Big(w_j^k(t,i-1)-\eta_k(t)[F_k'(w_j^k(t,i-1))+N_j^k(t,i-1)]\Big)\\ &= F_k''(w_j^k(t,i))\cdot\left(\frac{dw_j^k(t,i-1)}{dw_j^k(t,0)}-\eta_k(t)F_k''(w_j^k(t,i-1))\cdot\frac{dw_j^k(t,i-1)}{dw_j^k(t,0)}\right)\\ &= F_k''(w_j^k(t,i))\cdot\left(\frac{dw_j^k(t,i-1)}{dw_j^k(t,0)}[1-\eta_k(t)F_k''(w_j^k(t,i-1))]\right).\end{aligned}$$

Using a new local iteration, we have that

$$\frac{dw_j^k(t,i-1)}{dw_j^k(t,0)} = \left(\frac{dw_j^k(t,i-2)}{dw_j^k(t,0)}[1-\eta_k(t)F_k''(w_j^k(t,i-2))]\right).$$

Combining we get

$$\frac{d}{dw_j^k(t,0)}F_k'(w_j^k(t,i)) = F_k''(w_j^k(t,i))\cdot\left[\left(\frac{dw_j^k(t,i-2)}{dw_j^k(t,0)}[1-\eta_k(t)F_k''(w_j^k(t,i-2))]\right)[1-\eta_k(t)F_k''(w_j^k(t,i-1))]\right].$$

Since $\frac{dw_j^k(t,0)}{dw_j^k(t,0)}=1$, we can expand as

$$\begin{aligned}\frac{d}{dw_j^k(t,0)}F_k'(w_j^k(t,i)) &= F_k''(w_j^k(t,i))\cdot\left[\left(\frac{dw_j^k(t,i-2)}{dw_j^k(t,0)}[1-\eta_k(t)F_k''(w_j^k(t,i-2))]\right)[1-\eta_k(t)F_k''(w_j^k(t,i-1))]\right]\\ &= F_k''(w_j^k(t,i))\cdot\left[\prod_{l=1}^{i-1}[1-\eta_k(t)F_k''(w_j^k(t,i-l))]\right].\end{aligned}$$

Returning to the original expression for $\hat{M}(t)_j$ we have

$$\frac{d}{dw_j^k(t,0)}\hat{M}(t)_j = -\sum_{k=1}^{Q}p_k[\sum_{i=0}^{E-1}\eta_k(t)F_k''(w_j^k(t,i))\cdot\left[\prod_{l=1}^{i-1}[1-\eta_k(t)F_k''(w_j^k(t,i-l))]\right]].$$

We now bound the absolute value of this derivative as

$$\begin{aligned}|\frac{d}{dw_j^k(t,0)}\hat{M}(t)_j| &\le \sum_{k=1}^{Q}p_k\sum_{i=0}^{E-1}\eta_k(t)|F_k''(w_j^k(t,i))|\cdot\left[\prod_{l=1}^{i-1}[1+\eta_k(t)|F_k''(w_j^k(t,i-l))|]\right]\\ &\le \sum_{k=1}^{Q}p_k\sum_{i=0}^{E-1}\eta_k(t)L\cdot\left[\prod_{l=1}^{i-1}[1+\eta_k(t)L]\right]\\ &\le \sum_{k=1}^{Q}p_k\sum_{i=0}^{E-1}[1+\eta_k(t)L]^i.\end{aligned}$$

This holds for all components, therefore, we can conclude the drift term, $\eta_0(t)\hat{M}(t)$, is Lipschitz continuous.

□

A.3. Proof of Theorem 4.5

Proof. The proof structure is inspired by the structure of the continuous-time proofs for standard SGD, but have several additional complexities due to the averaging of client weights to determine server weights after a certain amount of time. The main steps are finding a suitable energy function of the stochastic process, bounding the infinitesimal diffusion generator, and using Dynkin's formula to complete the proof.

We first write out $\hat{M}(t)$ as

$$\hat{M}(t) = -\sum_{k=1}^{Q} p_k (\sum_{i=0}^{E-1} \eta_k(t) \mathbb{E}(G(t-1,i))).$$

A good resource for reviewing the background mathematics used in this proof is in the Supplementary materials of (Orvieto & Lucchi, 2019).

Using a similar approach as in (Orvieto & Lucchi, 2019), we define an appropriate energy function and then use Dynkin's formula to complete the proof. We are able to use all of the machinery for Itô diffusions due to our SDE being an Itô diffusion. In particular, we show in Lemma A.4 that the drift term is Lipschitz continuous and the proof for Lemma 1 in the supplementary materials of (Orvieto & Lucchi, 2019) proves Lipschitz continuity of the volatility matrix $\eta_0(t)\sqrt{h}\hat{V}^{1/2}(t)$.

We define $\partial_x(\cdot)$ as the vector of first derivatives with respect to each component of x and $\partial_{xx}(\cdot)$ as the matrix of partial derivatives of $\partial_{xx}(\cdot)$ with respect to each component of x. They are just the gradient and hessian with respect to the arguments of the energy function, but we use different notation than ∇ to prevent confusion with $\nabla F(w_0)$ which is the gradient of the loss function. The infinitesimal generator for an Itô diffusion of the form $dX(t) = f(t)dt + \sigma(t)dB(t)$ is defined as

$$\mathscr{A}(\cdot) = \partial_t(\cdot) + \langle \partial_x(\cdot), f(t) \rangle + \frac{1}{2}\text{Tr}(\sigma(t)\sigma(t)^T \partial_{xx}(\cdot)).$$

Following (Orvieto & Lucchi, 2019) and using Itô's lemma and the definition of an Itô diffusion, we obtain Dynkin's formula as

$$\mathbb{E}[\mathcal{E}(X(t),t)] - \mathcal{E}(x_0,0) = \mathbb{E}[\int_0^t \mathscr{A}\mathcal{E}(X(t'),t')dt']. \tag{5}$$

Dynkin's formula is vital to our convergence proofs along with defining the correct energy function $\mathcal{E}(\cdot)$.

We choose the energy function $\mathcal{E}(w_0) : \mathbb{R}^d \to \mathbb{R}_+$ as $\mathcal{E}(w_0) = F(w_0) - F(w_0^*)$. It is important to note that we define the energy function as just an argument of w_0 rather than both w_0 and t, thus the first term in Dynkin's formula is not applicable. This follows the same approach as (Orvieto & Lucchi, 2019).

We then bound the expectation of the stochastic integral of the infinitesimal generator of the process $\{\mathcal{E}(w_0(t))\}_{t\geq 0}$ as

$$\mathscr{A}\mathcal{E}(w_0(t)) = \underbrace{\langle \partial_{w_0}(\mathcal{E}(w_0(t))), \eta_0(t)\hat{M}(t) \rangle}_{B_1} + \underbrace{\frac{1}{2}\text{Tr}(h(\eta_0(t))^2 \hat{V}(t) \partial_{w_0 w_0}(\mathcal{E}(w_0(t))))}_{B_2}.$$

We first bound B_1 (for simplicity we assume $\eta_k(t) = \eta(t)$ for all clients) as

$$\begin{aligned}
B_1 &= \langle \nabla F(w_0(t)), -\eta_0(t) \sum_{k=1}^{Q} p_k \eta(t) (\sum_{i=0}^{E-1} \mathbb{E}[G^k(t,i)]) \rangle \\
&= -\eta_0(t) \sum_{k=1}^{Q} p_k \eta(t) \sum_{i=0}^{E-1} \nabla F(w_0(t))^T \mathbb{E}[G^k(t,i)] \\
&= -\eta_0(t) \sum_{i=0}^{E-1} \eta(t) \nabla F(w_0(t))^T \sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)].
\end{aligned}$$

We first define a function $R(t)$ such that $||\nabla F(w_0(t))||^2 = \nabla F(w_0(t))^T (\sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]) + R(t)$.

We then bound $R(t)$ as

$$\begin{aligned}
R(t) &= ||\nabla F(w_0(t))||^2 - F(w_0(t))^T(\sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]) \\
&= \nabla F(w_0(t))^T(F(w_0(t)) - \sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]) \\
&\leq |\nabla F(w_0(t))^T(\nabla F(w_0(t)) - \sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)])| \\
&\leq \underbrace{||\nabla F(w_0(t))||}_{\leq L} \cdot ||\nabla F(w_0(t)) - \sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]|| \quad \text{(Cauchy-Schwarz)} \\
&\leq L||\nabla F(w_0(t)) - \sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]|| \\
&= L||\mathbb{E}\sum_{k=1}^{Q}[\nabla F(w_0(t)) - p_k \nabla F^k(w^k(t,i))]|| \\
&\leq L\sum_{k=1}^{Q} p_k||\mathbb{E}[\nabla F(w_0(t)) - \nabla F^k(w^k(t,i))]|| \\
&\leq L\sum_{k=1}^{Q} p_k \mathbb{E}||\nabla F(w_0(t)) - \nabla F^k(w^k(t,i))|| \quad \text{(Jensen's Inequality)} \\
&\leq L\mu\sum_{k=1}^{Q} p_k \mathbb{E}||w_0(t) - w^k(t,i)|| \quad (F \text{ is } \mu\text{-smooth}) \\
&\leq L\mu i\eta(t)\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}] \quad \text{(Lemma A.1)}.
\end{aligned}$$

It follows that

$$\begin{aligned}
-\nabla F(w_0(t))^T(\sum_{k=1}^{Q} p_k \mathbb{E}[G^k(t,i)]) &= R(t) - ||\nabla F(w_0(t))||^2 \\
&\leq L\mu i\eta(t)\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}] - ||\nabla F(w_0(t))||^2.
\end{aligned}$$

We return to our bound on B_1 as

$$\begin{aligned}
B_1 &\leq \eta_0(t)\sum_{i=0}^{E-1}\eta(t)\Big[L\mu i\eta(t)\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}] - ||\nabla F(w_0(t))||^2\Big] \\
&= -E\eta_0(t)\eta(t)||\nabla F(w_0(t))||^2 + \eta_0(t)\eta(t)^2 L\mu\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}]\sum_{i=0}^{E-1} i \\
&= -E\eta_0(t)\eta(t)||\nabla F(w_0(t))||^2 + \eta_0(t)\eta(t)^2 L\mu\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}](\frac{1}{2}E(E-1)) \\
&\leq -E\eta_0(t)\eta(t)||\nabla F(w_0(t))||^2 + \frac{1}{2}E^2\eta_0(t)\eta(t)^2 L\mu\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}].
\end{aligned}$$

Now we bound B_2 as

$$B_2 = \frac{1}{2}\text{Tr}(h(\eta_0(t))^2\hat{V}(t)\partial_{w_0 w_0}(\mathcal{E}(w_0(t)))) = \frac{h(\eta_0(t))^2}{2}\text{Tr}\Big(\hat{V}(t)\nabla^2 F(w_0(t))\Big).$$

Using Lemma A.2, we can bound the trace of a product of symmetric matrices by the product of spectral norms. Both $\hat{V}(t)$ and $\nabla^2 F(w_0(t))$ are symmetric by their construction. So we have

$$B_2 \leq \frac{h(\eta_0(t))^2}{2}||\hat{V}(t)||_S||\nabla^2 F(w_0(t))||_S.$$

Then by assumptions 4.1, 4.2, and 4.3, we have

$$B_2 \leq \frac{h(\eta_0(t)\eta(t))^2}{2}||\hat{V}_1(t)||_S||\nabla^2 F(w_0(t))||_S \leq \frac{h(\eta_0(t)\eta(t))^2 V^* L}{2}.$$

We return to the bound on the infinitesimal generator as

$$\mathscr{A}\mathcal{E}(w_0(t)) = \underbrace{\langle \partial_{w_0}(\mathcal{E}(w_0(t))), \eta_0(t)\hat{M}(t)\rangle}_{B_1} + \underbrace{\frac{1}{2}\text{Tr}(h(\eta_0(t))^2\hat{V}(t)\partial_{w_0 w_0}(\mathcal{E}(w_0(t))))}_{B_2}$$
$$\leq -E\eta_0(t)\eta||\nabla F(w_0(t))||^2 + \underbrace{\frac{E^2 L\mu\sum_{k=1}^{Q} p_k[L + \sqrt{\text{Tr}(\Sigma_k)}]}{2}}_{C_1}\eta_0(t)\eta^2 + \frac{h(\eta_0(t)\eta(t))^2 V^* L}{2}.$$

Then we use Dynkin's Formula (Equation 5) and substitute $\eta_0(t) = 1$ to obtain the following expression

$$\mathbb{E}[\mathcal{E}(w_0(t))] - \mathcal{E}(w_0(0)) \leq \mathbb{E}\Big[\int_0^t [-E\eta(t')||\nabla F(w_0(t'))||^2 + C_1\eta(t')^2 + \frac{h\eta(t')^2 V^* L}{2}]dt'\Big]. \tag{6}$$

We follow a similar approach as (Orvieto & Lucchi, 2019) and define $\varphi(t) = \int_0^t \eta(t')dt'$. We can then define the distribution defined by the probability density function $f(t') = \frac{\eta(t')}{\varphi(t)}$. Lemma A.3 shows this is a valid probability density function. We define a random variable $\hat{t} \in [0, t]$ with probability density function $\frac{\eta(t')}{\varphi(t)}$. We then use a similar trick to (Johnson & Zhang, 2013) and (Orvieto & Lucchi, 2019) and use the law of the unconscious statistician to obtain

$$\mathbb{E}_{\hat{t}}||\nabla F(w_0(\hat{t}))||^2 = \frac{1}{\varphi(t)}\int_0^t \eta(t')||\nabla F(w_0(t'))||^2 dt' \tag{7}$$

where $\mathbb{E}_{\hat{t}}$ is the expectation over the random time point $\tilde{t}$.

We substitute Equation 7 into Inequality 6 and get

$$\underbrace{\mathbb{E}[\mathcal{E}(w_0(t))]}_{E_1(t)} - \mathcal{E}(w_0(0)) \leq -E\varphi(t)\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\hat{t}))||^2 + \int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2 V^* L}{2}]dt' \tag{8}$$

where $\mathbb{E}_{\mathcal{G},\tilde{t}}$ is the expectation over the random time point $\tilde{t}$ and the stochastic gradients, $\mathcal{G}$.

We notice that $E_1(t) = F(w_0(t)) - F(w_0^*) \geq 0$, so we can safely drop this term from the inequality.

So we can rewrite Equation 8 as

$$\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\hat{t}))||^2 \leq \frac{\mathcal{E}(w_0(0))}{E\varphi(t)} + \frac{1}{E\varphi(t)}\int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2V^*L}{2}]dt'$$
$$= \frac{F(w_0(0)) - F(w_0^*)}{E\varphi(t)} + \frac{1}{E\varphi(t)}\int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2V^*L}{2}]dt'.$$

□

A.4. Proof of Corollary 4.6

Proof. We first examine the case where $\eta(t) = \frac{1}{t+1}$.

We have

$$\begin{aligned}\varphi(t) &= \int_0^t \eta(t')dt' \\ &= \int_0^t \frac{1}{t'+1}dt' \\ &= \log(t'+1)\Big|_0^t \\ &= \log(t+1).\end{aligned}$$

Therefore we find

$$\begin{aligned}\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\tilde{t}))||^2 &\leq \frac{F(w_0(0)) - F(w_0^*)}{E\varphi(t)} + \frac{1}{E\varphi(t)}\int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2V^*L}{2}]dt' \\ &= \frac{F(w_0(0)) - F(w_0^*)}{E\log(t+1)} + \frac{1}{E\log(t+1)}\left[C_1\underbrace{\int_0^t \eta(t')^2dt'}_{=\frac{t}{t+1}} + \frac{hV^*L}{2}\int_0^t \eta(t')^2dt'\right] \\ &\leq \frac{F(w_0(0)) - F(w_0^*)}{E\log(t+1)} + \frac{C_1 + \frac{hV^*L}{2}}{E\log(t+1)} \\ &= \frac{F(w_0(0)) - F(w_0^*) + C_1 + \frac{hV^*L}{2}}{E}\frac{1}{\log(t+1)}.\end{aligned}$$

Now we examine $\eta(t) = \frac{1}{\sqrt{t+1}}$ We have $\varphi(t) = \int_0^t \eta(t')dt' = 2\sqrt{t+1} - 2$, and we find

$$\begin{aligned}\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\tilde{t}))||^2 &\leq \frac{F(w_0(0)) - F(w_0^*)}{E\varphi(t)} + \frac{1}{E\varphi(t)}\int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2V^*L}{2}]dt' \\ &= \frac{F(w_0(0)) - F(w_0^*)}{E(2\sqrt{t+1}-2)} + \frac{1}{E(2\sqrt{t+1}-2)}\left[C_1\underbrace{\int_0^t \eta(t')^2dt'}_{=\log(t+1)} + \frac{hV^*L}{2}\int_0^t \eta(t')^2dt'\right] \\ &\leq \frac{F(w_0(0)) - F(w_0^*)}{E(2\sqrt{t+1}-2)} + \frac{[C_1 + \frac{hV^*L}{2}]\log(t+1)}{E(2\sqrt{t+1}-2)}.\end{aligned}$$

Now we examine the asymptotic rates. This proof is the same as in (Orvieto & Lucchi, 2019). The general form of the inequality is

$$\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\tilde{t}))||^2 \leq \frac{F(w_0(0)) - F(w_0^*)}{E\varphi(t)} + \frac{1}{E\varphi(t)}\int_0^t [C_1\eta(t')^2 + \frac{h\eta(t')^2V^*L}{2}]dt'.$$

The first term on the RHS of the inequality is the deterministic term and has rate $\mathcal{O}(t^{b-1})$ for $0 < b < 1$ and $\mathcal{O}(1/\log(t))$ for $b = 1$. The stochastic term is $\mathcal{O}(t^{-b})$ for $b \in (0, 1/2) \cup (1/2, 1)$, $\mathcal{O}(\frac{\log(t)}{\sqrt{t}})$ for $b = 1/2$, and $\mathcal{O}(\frac{1}{\log t})$ for $b = 1$.

□

A.5. Proof of Corollary 4.7

Proof. We examine the convergence for the choice of client learning rate, $\eta(t) = \eta_c$, and server learning rate, $\eta_0(t) = \frac{1}{t+1}$. Most of the proof follows the same steps as the proof for Theorem 4.5, but we return to the inequality of the infinitesimal generator as

$$\begin{aligned}\mathscr{A}\mathcal{E}(w_0(t)) &\leq -E\eta_0(t)\eta||\nabla F(w_0(t))||^2 + \underbrace{\frac{E^2L\mu\sum_{k=1}^{Q}p_k[L+\sqrt{\text{Tr}(\Sigma_k)}]}{2}}_{C_1}\eta_0(t)\eta^2 + \frac{h(\eta_0(t)\eta(t))^2V^*L}{2} \\ &= -E\eta_0(t)\eta_c||\nabla F(w_0(t))||^2 + \underbrace{\frac{E^2L\mu\sum_{k=1}^{Q}p_k[L+\sqrt{\text{Tr}(\Sigma_k)}]}{2}}_{C_1}\eta_0(t)\eta_c^2 + \frac{h(\eta_0(t)\eta_c)^2V^*L}{2}.\end{aligned}$$

Use Dynkin's Formula, 5, to obtain

$$\mathbb{E}[\mathcal{E}(w_0(t))] - \mathcal{E}(w_0(0)) \leq \mathbb{E}\bigg[\int_0^t [-E\eta_c\eta_0(t')||\nabla F(w_0(t'))||^2 + C_1\eta_0(t')\eta_c^2 + \frac{h\eta_0(t')^2\eta_c^2V^*L}{2}]dt'\bigg]. \tag{9}$$

Now we follow steps similar to the final steps of the proof for Theorem 4.5.

We set $\varphi(t) = \int_0^t \eta_0(t')dt'$. We then substitute Equation 7 into Inequality 6 and get

$$\underbrace{\mathbb{E}[\mathcal{E}(w_0(t))]}_{E_1(t)} - \mathcal{E}(w_0(0)) \leq -E\eta_c\varphi(t)\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\hat{t}))||^2 + \eta_c^2\int_0^t [C_1\eta_0(t') + \frac{h\eta_0(t')^2V^*L}{2}]dt'. \tag{10}$$

We notice that $E_1(t) = F(w_0(t)) - F(w_0^*) \geq 0$, so we can safely drop this term from the inequality.

So we can rewrite Equation 10 as

$$\begin{aligned}\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\hat{t}))||^2 &\leq \frac{\mathcal{E}(w_0(0))}{E\eta_c\varphi(t)} + \frac{\eta_c}{E\varphi(t)}\int_0^t [C_1\eta_0(t') + \frac{h\eta_0(t')^2V^*L}{2}]dt' \\ &= \frac{F(w_0(0)) - F(w_0^*)}{E\eta_c\varphi(t)} + \frac{\eta_c}{E\varphi(t)}\int_0^t [C_1\eta_0(t') + \frac{h\eta_0(t')^2V^*L}{2}]dt'.\end{aligned}$$

Now we substitute $\eta_0(t) = \frac{1}{t+1}$ and get

$$\begin{aligned}\mathbb{E}_{\mathcal{G},\tilde{t}}||\nabla F(w_0(\hat{t}))||^2 &\leq \frac{F(w_0(0)) - F(w_0^*)}{E\eta_c\log(t+1)} + \frac{\eta_c}{E\log(t+1)}[C_1\log(t+1) + \frac{hV^*L}{2}]. \\ &= \frac{F(w_0(0)) - F(w_0^*) + \eta_c^2hV^*L/2}{E\eta_c\log(t+1)} + \frac{C_1\eta_c}{E} \\ &= \frac{F(w_0(0)) - F(w_0^*) + \eta_c^2hV^*L/2}{E\eta_c\log(t+1)} + \frac{E\eta_cL\mu\sum_{k=1}^{Q}p_k[L+\sqrt{\text{Tr}(\Sigma_k)}]}{2}.\end{aligned}$$

□

A.6. Proof of Theorem 4.9

Proof. This proof follows similarly to the proof of Theorem 4.5 where we find a suitable energy function, bound the infinitesimal generator, then use Dynkin's Formula to complete the bound.

For this case, we choose the energy function $\mathcal{E}(w_0) : \mathbb{R}^d \to \mathbb{R}_+$ as $\mathcal{E}(w_0) = \frac{1}{2}||w_0 - w_0^*||^2$.

We then bound the expectation of the stochastic integral of the infinitesimal generator of the process $\{\mathcal{E}(w_0(t))\}_{t\geq 0}$ as

$$\mathscr{A}\mathcal{E}(w_0(t)) = \underbrace{\langle \partial_{w_0}(\mathcal{E}(w_0(t))), \eta_0(t)\hat{M}(t)\rangle}_{B_1} + \underbrace{\frac{1}{2}\text{Tr}(h(\eta_0(t))^2\hat{V}(t)\partial_{w_0 w_0}(\mathcal{E}(w_0(t))))}_{B_2}.$$

First examine B_1 as

$$\begin{aligned} B_1 &= \langle w_0 - w_0^*, -\eta_0(t)\sum_{k=1}^{Q} p_k \eta(t) \sum_{i=0}^{E-1} \mathbb{E}[G^k(t,i)]\rangle \\ &= -\eta_0(t)\sum_{k=1}^{Q} p_k \eta(t) \sum_{i=0}^{E-1} (w_0(t) - w_0^*)^T \mathbb{E}[\nabla F^k(w^k(t,i))]. \end{aligned}$$

We define $R(t)$ such that

$$(w_0(t) - w_0^*)^T \mathbb{E}[\nabla F^k(w^k(t,i))] + R(t) = (w_0(t) - w_0^*)^T (\nabla F^k(w_0(t))). \tag{11}$$

Now we bound $R(t)$ as

$$\begin{aligned} R(t) &= (w_0(t) - w_0^*)^T(\nabla F^k(w_0(t)) - \mathbb{E}[\nabla F^k(w^k(t,i))]) \\ &\leq |(w_0(t) - w_0^*)^T(\nabla F^k(w_0(t)) - \mathbb{E}[\nabla F^k(w^k(t,i))])| \\ &\leq ||w_0(t) - w_0^*|| \cdot ||\nabla F^k(w_0(t)) - \mathbb{E}[\nabla F^k(w^k(t,i))]|| \quad \text{(Cauchy-Schwarz)} \\ &\leq ||w_0(t) - w_0^*|| \cdot ||\mathbb{E}[\nabla F^k(w_0(t)) - \nabla F^k(w^k(t,i))]|| \\ &\leq ||w_0(t) - w_0^*|| \cdot \mathbb{E}||\nabla F^k(w_0(t)) - \nabla F^k(w^k(t,i))|| \quad \text{(Jensen's Inequality)} \\ &\leq \mu||w_0(t) - w_0^*|| \cdot \mathbb{E}||w_0(t) - w^k(t,i)|| \quad (\mu\text{-smooth}) \\ &\leq \mu i\eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]||w_0(t) - w_0^*|| \quad \text{(Lemma A.1)} \\ &= \mu i\eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]||w_0(0) + (w_0(t) - w_0(0)) - w_0^*||. \end{aligned}$$

Now we need to find a bound on $||w_0(0) + (w_0(t) - w_0(0)) - w_0^*||$. Specifically, we need to bound $||w_0(t) - w_0(0)||$. We bound $||w_0(t)||$ as

$$\begin{aligned} ||w_0(t) - w_0(0)|| &= ||w_0(0) - w_0(0) + \int_0^t \eta_0(t')\hat{M}(t')dt' + \int_0^t \sqrt{h}\eta_0(t')\hat{V}^{1/2}(t')dB(t')|| \\ &\leq \int_0^t \eta_0(t')||\hat{M}(t')||dt' + ||\int_0^t \sqrt{h}\eta_0(t')\hat{V}^{1/2}(t')dB(t')||. \end{aligned}$$

We assume constant server learning rate $\eta_0(t) = \eta_0$. We examine $\int_0^t \eta_0(t')||\hat{M}(t')||dt'$ as follows

$$\begin{aligned} \int_0^t \eta_0(t')||\hat{M}(t')||dt' &= \int_0^t \eta_0(t')||\eta(t')\sum_{k=1}^{Q} p_k \sum_{i=0}^{E-1} \mathbb{E}[G^k(t',i)]||dt' \\ &\leq \int_0^t \eta_0(t')\eta(t')\sum_{k=1}^{Q} p_k \sum_{i=0}^{E-1} \mathbb{E}||\nabla F^k(w_k(t',i))||dt' \\ &\leq \int_0^t \eta_0(t')\eta(t')LEdt' \\ &= \eta_0 LE\int_0^t \eta(t')dt'. \end{aligned}$$

Therefore we have the following

$$||w_0(t) - w_0(0)|| \leq \eta_0 LE \int_0^t \eta(t')dt' + \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||.$$

Returning back to $R(t)$, we have

$$R(t) \leq \mu i \eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^t \eta(t')dt' + \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||\Bigg].$$

Returning to Equation 11, it follows that

$$\begin{aligned}
-(w_0(t) - w_0^*)^T \mathbb{E}[\nabla F^k(w^k(t,i))] &= R(t) - (w_0(t) - w_0^*)^T(\nabla F^k(w_0(t))) \\
&\leq \mu i \eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^t \eta(t')dt' \\
&+ \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||\Bigg] - (w_0(t) - w_0^*)^T(\nabla F^k(w_0(t))) \\
&\leq \mu i \eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^t \eta(t')dt' \\
&+ \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||\Bigg] - \tau(F(w_0(t)) - F(w_0^*)) \quad \text{(Assumption 4.8)}.
\end{aligned}$$

Returning back to the original bound on B_1, we have

$$\begin{aligned}
B_1 \leq \eta_0 \sum_{k=1}^{Q} p_k \eta(t) \sum_{i=0}^{E-1} \Bigg(&\mu i \eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^t \eta(t')dt' + \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||\Bigg] \\
&- \tau(F(w_0(t)) - F(w_0^*))\Bigg).
\end{aligned}$$

Now we examine B_2 as

$$\begin{aligned}
B_2 &= \frac{1}{2}\text{Tr}(h(\eta_0(t))^2 \hat{V}(t) \partial_{w_0 w_0}(\mathcal{E}(w_0(t)))) &(12)\\
&= \frac{h\eta_0^2}{2}\text{Tr}(\hat{V}(t)). &(13)
\end{aligned}$$

From Lemma 3 in the supplementary materials of (Orvieto & Lucchi, 2019) and the same approach that we use in the proof of Theorem 4.5, we have that $\text{Tr}(\hat{V}(t)) \leq d\eta(t)^2 V^*$ where d is the dimensionality of our weights w_0. Therefore we have

$$B_2 \leq \frac{dh\eta_0^2 \eta(t)^2 V^*}{2}. \tag{14}$$

We return to the bound on the infinitesimal generator as

$$\begin{aligned}
\mathscr{A}\mathcal{E}(w_0(t)) \leq \eta_0 \sum_{k=1}^{Q} p_k \eta(t) \sum_{i=0}^{E-1} \Bigg(&\mu i \eta(t)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^t \eta(t')dt' + \sqrt{h}\eta_0 || \int_0^t \hat{V}^{1/2}(t')dB(t')||\Bigg] \\
&- \tau(F(w_0(t)) - F(w_0^*))\Bigg) + \frac{dh\eta_0^2 \eta(t)^2 V^*}{2}.
\end{aligned}$$

As in the proof of Theorem 4.5, we now use Dynkin's Formula to get

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(w_0(t))] - \mathcal{E}(w_0(0)) &\leq \mathbb{E}\Bigg[\int_0^t \Bigg[\eta_0 \sum_{k=1}^{Q} p_k \eta(s) \sum_{i=0}^{E-1} \Bigg(\mu i \eta(s)[L + \sqrt{\text{Tr}(\Sigma_k)}]\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^s \eta(t')dt' \\
&+ \sqrt{h}\eta_0 || \int_0^s \hat{V}^{1/2}(t')dB(t')||\Bigg] - \tau(F(w_0(s)) - F(w_0^*))\Bigg) + \frac{dh\eta_0^2 \eta(s)^2 V^*}{2}\Bigg] ds\Bigg] \\
&\leq \mathbb{E}\Bigg[\int_0^t \Bigg[\eta_0 \eta(s)\Bigg(\underbrace{\mu E^2 \sum_{k=1}^{Q} p_k [L + \sqrt{\text{Tr}(\Sigma_k)}]}_{C_2} \eta(s)\Bigg[||w_0(0) - w_0^*|| + \eta_0 LE \int_0^s \eta(t')dt' \\
&+ \sqrt{h}\eta_0 || \int_0^s \hat{V}^{1/2}(t')dB(t')||\Bigg] - \tau(F(w_0(s)) - F(w_0^*))\Bigg) + \frac{dh\eta_0^2 \eta(s)^2 V^*}{2}\Bigg] ds\Bigg] \\
&= \mathbb{E}\Bigg[\int_0^t \Bigg[\eta_0 C_2 \eta(s)^2 \Bigg(\eta_0 LE \int_0^s \eta(t')dt' + \sqrt{h}\eta_0 || \int_0^s \hat{V}^{1/2}(t')dB(t')||\Bigg)\Bigg] ds\Bigg] \\
&+ \underbrace{\Bigg[\frac{dh\eta_0^2 V^*}{2} + \eta_0 C_2 ||w_0(0) - w_0^*||\Bigg]}_{C_3} \mathbb{E}\Bigg[\int_0^t \eta(s)^2 ds\Bigg] - \mathbb{E}\Bigg[\int_0^t \Bigg[\eta_0 \eta(s) \tau(F(w_0(s)) - F(w_0^*))\Bigg] ds\Bigg] \\
&= \eta_0^2 C_2 \mathbb{E}\Bigg[\int_0^t \Bigg[\eta(s)^2 \Bigg(LE \int_0^s \eta(t')dt' + \sqrt{h}|| \int_0^s \hat{V}^{1/2}(t')dB(t')||\Bigg)\Bigg] ds\Bigg] \\
&+ C_3 \mathbb{E}\Bigg[\int_0^t \eta(s)^2 ds\Bigg] - \tau\eta_0 \mathbb{E}\Bigg[\int_0^t \Bigg[\eta(s)(F(w_0(s)) - F(w_0^*))\Bigg] ds\Bigg].
\end{aligned}$$

We examine the term with $\hat{V}^{1/2}(t')$ as

$$\mathbb{E}\Bigg[\int_0^t \eta(s)^2 \sqrt{h} || \int_0^s \hat{V}^{1/2}(t')dB(t')||ds\Bigg] = \int_0^t \eta(s)^2 \sqrt{h} \mathbb{E}\Bigg[|| \int_0^s \hat{V}^{1/2}(t')dB(t')||\Bigg] ds,$$

and we can make this switch because $\eta(s)^2\sqrt{h}\mathbb{E}\Big[|| \int_0^s \hat{V}^{1/2}(t')dB(t')||\Big] < \infty$. From multivariate Ito isometry and Jensen's

inequality for concave functions (such as the square root function), we have that

$$\begin{aligned}
\mathbb{E}||\int_0^s \hat{V}^{1/2}(t')dB(t')|| &\leq \sqrt{\mathbb{E}||\int_0^s \hat{V}^{1/2}(t')dB(t')||^2} \\
&\leq \sqrt{\mathbb{E}\int_0^s ||\hat{V}^{1/2}(t')||^2 dt'} \\
&= \sqrt{\mathbb{E}\int_0^s ||\eta(t')\hat{V_1}^{1/2}(t')||_F^2 dt'} \\
&= \sqrt{\mathbb{E}\int_0^s \eta(t')^2||\hat{V_1}^{1/2}(t')||_F^2 dt'} \\
&\leq \sqrt{\int_0^s \eta(t')^2 (V^*)^2 dt'} \\
&= V^*\sqrt{\int_0^s \eta(t')^2 dt'},
\end{aligned}$$

where $||\cdot||_F$ is the Frobenius norm.

Returning to the bound from Dynkin's Formula we have

$$\begin{aligned}
\underbrace{\mathbb{E}[\mathcal{E}(w_0(t))]}_{\geq 0} - \mathcal{E}(w_0(0)) \leq \eta_0^2 C_2 \int_0^t \left[\eta(s)^2\left(LE\int_0^s \eta(t')dt' + \sqrt{h}V^*\sqrt{\int_0^s \eta(t')^2 dt'}\right)\right]ds \\
+ C_3\int_0^t \eta(s)^2 ds - \tau\eta_0\mathbb{E}\left[\int_0^t \left[\eta(s)(F(w_0(s)) - F(w_0^*))\right]ds\right].
\end{aligned}$$

Rearranging we obtain

$$\begin{aligned}
\mathbb{E}\left[\int_0^t \left[\eta(s)(F(w_0(s)) - F(w_0^*))\right]ds\right] \leq \frac{\mathcal{E}(w_0(0))}{\tau\eta_0} + \frac{\eta_0^2 C_2}{\tau\eta_0}\int_0^t \left[\eta(s)^2\left(LE\int_0^s \eta(t')dt' + \sqrt{h}V^*\sqrt{\int_0^s \eta(t')^2 dt'}\right)\right]ds \\
+ \frac{C_3}{\tau\eta_0}\int_0^t \eta(s)^2 ds.
\end{aligned}$$

Using the same trick as in the proof of Theorem 4.5, where we create random variable $\hat{t} \in [0, t]$ with probability density function $\frac{\eta(t')}{\varphi(t)}$, we can substitute

$$\varphi(t)\mathbb{E}_{\hat{t}}[(F(w_0(\hat{t})) - F(w_0^*))] = \int_0^t \eta(t')(F(w_0(t')) - F(w_0^*))dt',$$

where $\varphi(t) = \int_0^t \eta(t')dt'$.

We finally find

$$\begin{aligned}
\mathbb{E}_{\mathcal{G},\hat{t}}[(F(w_0(\hat{t})) - F(w_0^*))] \leq \frac{||w_0(0) - w_0^*||}{\tau\eta_0\varphi(t)} + \frac{\eta_0^2 C_2}{\tau\eta_0\varphi(t)}\int_0^t \left[\eta(s)^2\left(LE\int_0^s \eta(t')dt' + \sqrt{h}V^*\sqrt{\int_0^s \eta(t')^2 dt'}\right)\right]ds \\
+ \frac{C_3}{\tau\eta_0\varphi(t)}\int_0^t \eta(s)^2 ds.
\end{aligned}$$

□

A.7. Proof of Corollary 4.10

Proof. We have $\varphi(t) = \log(t+1)$.

We examine the case where $\eta(t) = \frac{1}{t+1}$ as

$$\begin{aligned}
\mathbb{E}_{\mathcal{G},\tilde{t}}[(F(w_0(\hat{t})) - F(w_0^*))] &\leq \frac{||w_0(0) - w_0^*||}{\tau\eta_0\varphi(t)} + \frac{\eta_0^2 C_2}{\tau\eta_0\varphi(t)} \int_0^t \left[\eta(s)^2 \left(LE \int_0^s \eta(t')dt' + \sqrt{h}V^* \sqrt{\int_0^s \eta(t')^2 dt'}\right)\right] ds \\
&+ \frac{C_3}{\tau\eta_0\varphi(t)} \int_0^t \eta(s)^2 ds \\
&= \frac{||w_0(0) - w_0^*||}{\tau\eta_0 \log(t+1)} + \frac{\eta_0^2 C_2}{\tau\eta_0 \log(t+1)} \int_0^t \left[\left(LE\frac{\log(s+1)}{(s+1)^2} + \frac{\sqrt{h}V^*}{(s+1)^2}\sqrt{\frac{s}{(s+1)}}\right)\right] ds \\
&+ \frac{C_3}{\tau\eta_0 \log(t+1)} \int_0^t \eta(s)^2 ds \\
&\leq \frac{||w_0(0) - w_0^*|| + C_3 + \eta_0^2 C_2 \sqrt{h} V^*}{\tau\eta_0 \log(t+1)} + \frac{\eta_0^2 C_2 LE}{\tau\eta_0} \cdot \frac{t - \log(t+1)}{t\log(t+1)}.
\end{aligned}$$

□

A.8. Proof of Theorem 5.4

Proof. **Lyapunov CLT** We first describe the Lyapunov Central Limit Theorem. Assume we have a sequence of random variables $X_1, X_2, ..., X_n$ which are independent but not necessarily identically distributed. These random variables have possible different means μ_i and variances σ_i^2. To satisfy the Lyapunov condition, we need to show that for some $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|X_i - \mu_i|^{2+\delta}] = 0 \tag{15}$$

where $s_n^2 = \sum_{i=1}^n \sigma_i^2$.

The problem we are studying is of $A(t)$, which is formulated as

$$A(t) = \sum_{k=1}^{Q} \underbrace{p_k [\sum_{i=0}^{E-1} \eta_k(t)(N_{t-E+i}^k - G_{t-E+i}^k)]}_{A_k} \tag{16}$$

where $N_{t+i}^k \sim \mathcal{N}(0, \Sigma_k(w_{t+i}^k))$ and $G_{t+i}^k = \nabla F^k(w_{t+i}^k)$.

For simplicity, the proof below assumes A_k is one-dimensional. However, from Assumption 5.1, we have that $\Sigma_k(w_{t+i}^k)$ is diagonal. This means we can apply this approach to each component of a vector A_k, and thus each component of A_k tends to a normal distribution that are independent of each other because $\Sigma_k(w_{t+i}^k)$ is diagonal. Since the components are independent and normally distributed, we have that A_k tends to a multivariate normal distribution.

To satisfy the Lyapunov condition, we need to show that for some choice of δ the following holds

$$\lim_{Q\to\infty} \frac{1}{s_Q^{2+\delta}} \sum_{k=1}^{Q} \mathbb{E}\left[\left|A_k - \mathbb{E}[A_k]\right|^{2+\delta}\right] = 0.$$

We choose $\delta = 2$ and examine $\mathbb{E}\Big[\big|A_k - \mathbb{E}[A_k]\big|^{2+\delta}\Big]$ as

$$\begin{aligned}
\mathbb{E}\Big[\big|A_k - \mathbb{E}[A_k]\big|^{2+\delta}\Big] &= \mathbb{E}\Big[\Big|p_k[\sum_{i=0}^{E-1}\eta_k(t)(N^k_{t-E+i} - G^k_{t-E+i})] - \mathbb{E}[p_k[\sum_{i=0}^{E-1}\eta_k(t)(N^k_{t-E+i} - G^k_{t-E+i})]]\Big|^4\Big] \\
&= p_k^4\eta_k(t)^4\mathbb{E}\Big[\Big|\underbrace{\sum_{i=0}^{E-1}(N^k_{t-E+i})}_{=N_k} + \underbrace{\sum_{i=0}^{E-1}(\mathbb{E}[G^k_{t-E+i}] - G^k_{t-E+i})}_{=R_k}\Big|^4\Big] \\
&= p_k^4\eta_k(t)^4\mathbb{E}\Big[\Big(N_k + R_k\Big)^4\Big] \\
&= p_k^4\eta_k(t)^4\mathbb{E}\Big[N_k^4 + 4N_k^3R_k + 6N_k^2R_k^2 + 4N_kR_k^3 + R_k^4\Big] \\
&= p_k^4\eta_k(t)^4\Big[\mathbb{E}[N_k^4] + 4\mathbb{E}[N_k^3R_k] + 6\mathbb{E}[N_k^2R_k^2] + \mathbb{E}[4N_kR_k^3] + \mathbb{E}[R_k^4]\Big].
\end{aligned}$$

We have that $s_Q^4 = \Big(\sum_{k=1}^Q \mathbb{E}\Big[\big(A_k - \mathbb{E}[A_k]\big)^2\Big]\Big)^2 = \Big(\sum_{k=1}^Q p_k^2\eta_k(t)^2\mathbb{E}\big[(N_k + R_k)^2\big]\Big)^2$.

We can use the following formula which states that for positive x_i,

$$\sum_{i=1}^n x_i^2 = \frac{(\sum_{i=1}^n x_i)^2 + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(x_i - x_j)^2}{n}.$$

We quickly prove this equation as follows

$$\begin{aligned}
\frac{(\sum_{i=1}^n x_i)^2 + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(x_i - x_j)^2}{n} &= \frac{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n\sum_{j\neq i} x_ix_j + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(x_i^2 - 2x_ix_j + x_j^2)}{n} \\
&= \frac{(n+1)\sum_{i=1}^n x_i^2 + \sum_{i=1}^n\sum_{j\neq i} x_ix_j - \sum_{i=1}^n\sum_{j=1}^n x_ix_j}{n} \\
&= \frac{(n+1)\sum_{i=1}^n x_i^2 + \sum_{i=1}^n(\sum_{j\neq i} x_ix_j - \sum_{j=1}^n x_ix_j)}{n} \\
&= \frac{(n+1)\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i^2}{n} \\
&= \sum_{i=1}^n x_i^2.
\end{aligned}$$

Using this formula and substituting $x_i = p_k^2\eta_k(t)^2\mathbb{E}\big[(N_k + R_k)^2\big]$ we get

$$\begin{aligned}
\Big(\sum_{k=1}^Q p_k^2\eta_k(t)^2\mathbb{E}\big[(N_k + R_k)^2\big]\Big)^2 &= Q\sum_{k=1}^Q\Big(p_k^2\eta_k(t)^2\mathbb{E}\big[(N_K + R_k)^2\big]\Big)^2 \\
&\quad - \frac{1}{2}\sum_{i=1}^Q\sum_{j=1}^Q\Big(p_i^2\eta_i(t)^2\mathbb{E}\big[(N_i + R_i)^2\big] - p_j^2\eta_j(t)^2\mathbb{E}\big[(N_j + R_j)^2\big]\Big)^2.
\end{aligned}$$

From our assumptions on the similarity of variance between clients, we have for some $C > 0$

$$\begin{aligned}
\Big(\sum_{k=1}^Q p_k^2\eta_k(t)^2\mathbb{E}\big[(N_K + R_k)^2\big]\Big)^2 &= Q\sum_{k=1}^Q C \\
&= CQ^2.
\end{aligned}$$

We have

$$\begin{aligned}
\frac{1}{s_Q^4}\sum_{k=1}^{Q}\mathbb{E}\Big[\big|A_k-\mathbb{E}[A_k]\big|^4\Big] &= \frac{1}{CQ^2}\sum_{k=1}^{Q}\mathbb{E}\Big[\big|A_k-\mathbb{E}[A_k]\big|^4\Big] \\
&= \frac{1}{CQ^2}\sum_{k=1}^{Q}p_k^4\eta_k(t)^4\Big[\mathbb{E}[N_k^4]+4\mathbb{E}[N_k^3R_k]+6\mathbb{E}[N_k^2R_k^2]+\mathbb{E}[4N_kR_k^3]+\mathbb{E}[R_k^4]\Big] \\
&\le \frac{1}{CQ^2}\sum_{k=1}^{Q}p_k^4\eta_k(t)^4\Big[|\mathbb{E}[N_k^4]|+4|\mathbb{E}[N_k^3R_k]|+6|\mathbb{E}[N_k^2R_k^2]|+4|\mathbb{E}[N_kR_k^3]|+|\mathbb{E}[R_k^4]|\Big] \\
&\le \frac{1}{CQ^2}\sum_{k=1}^{Q}p_k^4\eta_k(t)^4\Big[D+4D+6D+4D+D\Big] \\
&\le \frac{16DQ}{CQ^2}.
\end{aligned}$$

Since, $0\le\frac{1}{s_Q^4}\sum_{k=1}^{Q}\mathbb{E}\Big[\big|A_k-\mathbb{E}[A_k]\big|^4\Big]\le\frac{16D}{CQ}$, and $\lim_{Q\to\infty}0=0$ and $\lim_{Q\to\infty}\frac{16D}{CQ}=0$, by the squeeze theorem, we know that $\lim_{Q\to\infty}\frac{1}{s_Q^4}\sum_{k=1}^{Q}\mathbb{E}\Big[\big|A_k-\mathbb{E}[A_k]\big|^4\Big]=0$. Thus, the Lyapunov condition holds.

□

A.9. Proof of Theorem 6.2

Proof. We wish to show that after E local iterations of SGD, the local weights evolve as

$$w_{t+E}^k\sim w_t^0-\eta\sum_{j=0}^{E-1}(I-\eta U_k)^jU_k(w_t^0-a_k)-\eta\sum_{j=0}^{E-1}\sum_{i=1}^{j}(I-\eta U_k)^{j-i}U_k\mathcal{N}(0,\Sigma_k)+E\mathcal{N}(0,\Sigma_k).$$

First observe the evolution for a few iterations as

$$\begin{aligned}
w_{t+1}^k &\sim w_t^0-\eta\mathcal{N}(U_k(w_t^0-a_k),\Sigma_k) \\
w_{t+2}^k|w_{t+1}^k &\sim w_{t+1}^k-\eta\mathcal{N}(U_k(w_{t+1}^k-a_k),\Sigma_k) \\
&\sim w_t^0-\eta U_k(w_t^0-a_k)-\eta U_k(w_{t+1}^k-a_k)+2\eta\mathcal{N}(0,\Sigma_k) \\
w_{t+3}^k|(w_{t+2}^k,w_{t+1}^k) &\sim w_t^0-\eta U_k(w_t^0-a_k)-\eta U_k(w_{t+1}^k-a_k)+2\eta\mathcal{N}(0,\Sigma_k)-\eta U_k(w_{t+2}^k-a_k)+\eta\mathcal{N}(0,\Sigma_k).
\end{aligned}$$

Extend to E number of time steps forward as

$$w_{t+E}^k|(w_{t+1}^k,...,w_{t+E-1}^k)\sim w_t^0-\eta\sum_{j=0}^{E-1}U_k(w_{t+j}^k-a_k)+E\eta\mathcal{N}(0,\Sigma_k). \tag{17}$$

In general we have

$$U_k(w_{t+j}^k-a_k)=(1-\eta U_k)U_k(w_{t+j-1}^k-a_k)+U_k\mathcal{N}(0,\Sigma_k).$$

Starting from $j=1$, we have

$$\begin{aligned}
U_k(w_{t+1}^k-a_k) &= (1-\eta U_k)U_k(w_t^0-a_k)+U_k\mathcal{N}(0,\Sigma_k) \\
&= (U_k-\eta U_k^2)w_t^0+(-U_k+\eta U_k^2)a_k+U_k\mathcal{N}(0,\Sigma_k).
\end{aligned}$$

Now for $j=2$, we have

$$U_k(w_{t+2}^k-a_k)=(1-\eta U_k)[(1-\eta U_k)U_k(w_t^0-a_k)+U_k\mathcal{N}(0,\Sigma_k)]+U_k\mathcal{N}(0,\Sigma_k).$$

Expanding out to $j = E$, we have

$$U_k(w_{t+j}^k - a_k) = (I - \eta U_k)^j U_k(w_t^0 - a_k) + \sum_{i=1}^{j}(I - \eta U_k)^{j-i} U_k \mathcal{N}(0, \Sigma_k). \tag{18}$$

Combining equations 17 and 18, we have

$$w_{t+E}^k \sim w_t^0 - \eta \sum_{j=0}^{E-1}(I - \eta U_k)^j U_k(w_t^0 - a_k) - \eta \sum_{j=0}^{E-1}\sum_{i=1}^{j}(I - \eta U_k)^{j-i} U_k \mathcal{N}(0, \Sigma_k) + E\eta \mathcal{N}(0, \Sigma_k).$$

□

A.10. Global loss function form for quadratic assumption

With the quadratic assumption, we can reformat the equation for the server loss function as

$$F(w) = \frac{1}{2}(w - a)^T (\sum_{k=1}^{Q} p_k U_k)(w - a).$$

We must find a. We have that $\min_w F(w) = a$, $\nabla F(w) = \sum_{k=1}^{Q} p_k U_k (w - a_k)$ and $\nabla_w F(a) = 0$. We find

$$\begin{aligned}\nabla_w F(a) = 0 &= \sum_{k=1}^{Q} p_k U_k (a - a_k) \\ &= \sum_{k=1}^{Q} p_k U_k a - \sum_{k=1}^{Q} p_k U_k a_k.\end{aligned}$$

Now we solve for a as

$$\begin{aligned}\sum_{k=1}^{Q} p_k U_k a &= \sum_{k=1}^{Q} p_k U_k a_k \\ a &= (\sum_{k=1}^{Q} p_k U_k)^{-1} \sum_{k=1}^{Q} p_k U_k a_k.\end{aligned}$$

Therefore, the server loss function is also represented by a quadratic form with local minimum at $w_0^* = (\sum_{k=1}^{Q} p_k U_k)^{-1} \sum_{k=1}^{Q} p_k U_k a_k$ and Hessian $H_0 = (\sum_{k=1}^{Q} p_k U_k)$.

A.11. Proof of Theorem 6.3

For a linear SDE of the form, $dX(t) = (a(t) + A(t)X(t))dt + b(t)dB(t)$, the solution is normally distributed as $\mathcal{N}(m(t), v(t))$. The mean $m(t)$ is the solution of the ordinary differential equation (ODE) $\frac{dm(t)}{dt} = A(t)m(t) + a(t)$ with initial condition $m(0) = X(0)$. The variance $v(t)$ is the solution of the ODE $\frac{dv(t)}{dt} = 2A(t)v(t) + b(t)^2$ with initial condition $v(0) = 0$ (Movellan, 2011). These ODEs are straightforward to solve and we obtain the solution in Theorem 6.3.