A STOCHASTIC GRADIENT LANGEVIN DYNAMICS ALGORITHM FOR NOISE INTRINSIC FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-i.i.d data distribution and Differential privacy(DP) protections are two open problems in Federated Learning(FL). We address these two problems by proposing the first noise intrinsic FL training algorithms. In our proposed algorithm, we incorporate a stochastic gradient Langevin dynamices(SGLD) oracle in local node's parameter update phase. Our introduced SGLD oracle would lower generalization errors in local node's parameter learning and provide local node DP protections. We theoretically analyze our algorithm by formulating a min-max objective functions and connects its upper bound with global loss function in FL. The convergence of our algorithm on non-convex function is also given as contraction and coupling rate of two random process defined by stochastic differential equations(SDE) We would provide DP analysis for our proposed training algorithm and provide more experiment results soon.

1 INTRODUCTION

Federated Learning (FL) as a marriage on cloud computing and deep learning are gaining popularity on commercial deployment (Li et al., 2020). It follows a distributed protocol to allow multiple parties to participate on training process on their local side while collaborating and coordinating on the cloud site (Konečnỳ et al., 2015). As a result of an innovative corporation pattern, the consumer node would participate their part of training locally without data publishing, while technical product provider would provide professional service both on the tuning models in training process and expertise inference solutions from their per-trained model warehouses (Li et al., 2020). Federated Learning is especially suitable in the area of medical applications (Sheller et al., 2020). In one way local hospitals maintain and manage the slides of pathology documents such as images and reports. In another way, they are the consumers of computer aided automatic diagnosis products which comes from training on the patterns of these data and documents. Coexists with these promising parts, federated learning has its unique characteristics and challenges.

Firstly, the coordination and communication overheads between distributed nodes and centralized server is significantly higher than that of localized training (Sattler et al., 2019). A direct consequence is that a feasible FL algorithm consists E steps of local SGD updates in parallel (Li et al., 2019c) among than Federated Avgeraging (FedAvg) (McMahan et al., 2017) is the first perhaps the most widely used FL algorithm.

Secondly, the distribution of data is statistically heterogeneous on different devices. The generalization error in each single device's local training is huge. As a result, optimization direction would towards overfit on local data. The shifts in training optimal solution among local devices would cause the stabling point of FedAvg deviates be the non-optimal solution (Li et al., 2019c). One solution for non iid problems would be introducing proximal objective (Li et al., 2018) and dual variables (Zhang et al., 2020; Karimireddy et al., 2019). Xinwei (Zhang et al., 2020) provides an Augmented Lagrange solution on FL learning with non iid data.

Thirdly, the data privacy concerns is frequently encountered issue in Federated learning. Due to the vulnerability properties of internet environment. Information leaking is highly possible. Differential privacy works to incorporating a randomized mechanism such as injecting gradient noise(Dwork et al., 2014; Abadi et al., 2016) and irregular data sampling(Dong et al., 2019) so that the distribution of perturbed results are insensitive to single record change.

In an attempt to handle these challenges, we would bring a Stochastic Gradient MCMC (SG-MCMC) solution into FL settings. SG-MCMC methods as a class of scalable Bayesian sampling algorithm in machine learning has realized significant success recently. We use SG-MCMC in FL settings for its lower generation error bounds (Smith & Le, 2017; Li et al., 2019b) and differential privacy preserving properties (Li et al., 2019a) with appropriately chosen step sizes.

Several existing studies on the extention of SG-MCMC algorithms on improving the generalized performance of parameter learning and preserving differential privacy in Federated learning. Bhardwaj (Bhardwaj, 2019) showed that an adaptive stepsize of Stochastic Gradient Langevine Dynamics(SGLD) could escape local extremes of high generalization error. Chaudhari et al. (Chaudhari et al., 2019) propose a two nested SGD algorithm to perform SGLD in their local loop of optimization. Li et al. (Li et al., 2019a) proved that a practical stepsize of sampling models is realizable to preserve differential privacy. Wang et al. (Wang et al., 2019) gave an bound on empirical risk to measure the error of non-convex local loss under differential privacy. However their works are studying on the case of local training on a single node case.

Motivated by their works, we propose an SGLD algorithm in FL. In our proposed algorithm, each node use SGLD samplings as each node's local gradient update phase. The whole updating follows the protocol in FedPD algorithm (Zhang et al., 2020) except that we take expectations of SGLD sampling on the Augmented Lagrange objective. Next, we analyze our propose algorithm by formulating a joint min-max variational objective functions. The whole learning process in our algorithm would be viewed as a min-max descent in our objective functions. We then prove that our constructed min-max functions is a variational upper bound on the global loss functions where the introduced dual variables closes the gap among local gradient zeros. Finally we study the convergence of our algorithm by using a technique similar in (Eberle et al., 2019) to study the couplings and contraction in Hamilton Monte-Carlo. We prove that the distributio of two process from independent random initialization distributions converges in our designed Wasserstein metric. In this paper, our contributions are two folds.

- We propose an SGLD implementations of FL algorithm where the data distribution is non iid on local nodes.
- We formulate our SGLD implementaion of FL as optimizing on a min-max points of a joint learning objective function. And then we derive two types of variational upper bounds of our learning objectives on global loss functions and connects it with optimal primal-dual conditions in consensus problems. We also study our algorithm's convergence to the stabling point.

2 PRELIMINARIES

2.1 AUGMENTED LAGRANGE FOR FEDERATED LEARNING

In the framework of federated learning, N distributed nodes aim to learn a coherent network mapping model $\nu(\mathbf{x}, \cdot)$ in $\mathbb{R}^m \to \mathbb{R}^n$ parameterized by $\mathbf{x}$ by the loss function $l(\cdot, \cdot)$ in $\mathbb{R}^n, \mathbb{R}^n \to \mathbb{R}$. The data are distributed i.i.d cross N distributed nodes. We use $\mathcal{D}_i$ to denote the dataset on i's distributed node. We denote $\mathcal{D}_{i,q}$ as the qth data in Node i and $\mathcal{Y}_{i,q}$ as the label for qth data in Node i. The learning objective in i's node is defined as the expected loss from the network prediction on a data distribution $\mathcal{P}_i \sim p(\{\mathcal{D}_{i,q}, \mathcal{Y}_{i,q}\} \in \mathcal{D}_i)$

$$F_i(\mathbf{x}) = \mathbb{E}_{\{\mathcal{D}_{i,q}, \mathcal{Y}_{i,q}\} \in \mathcal{P}_i} l(\nu(\mathbf{x}, \mathcal{D}_{i,q}), \mathcal{Y}_{i,q}) \tag{1}$$

For simplicity, we use $\xi_{i,q} \triangleq \{\mathcal{D}_{i,q}, \mathcal{Y}_{i,q}\}$ to denote the combination of qth data and label in i's node. The loss on $\xi_{i,q}$ is denoted as

$$F_i(\mathbf{x}, \xi_{i,q}) \triangleq l(\nu(\mathbf{x}, \mathcal{D}_{i,q}), \mathcal{Y}_{i,q}) \tag{2}$$

The federated learning is aimed as minimizing the averaged loss across all the distributed nodes

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{N} \sum_i F_i(\mathbf{x}) \tag{3}$$

The federated learning process consists of multiple rounds of local distributed training, global aggregation. updating and broadcasts on parameters. In the start of round r, the central node first broadcast its coordinated value of $\mathbf{x}_0^r$ to each distributed node. Each distributed node keeps a copy of $\mathbf{x}_0^r$ as $\mathbf{x}_{0,r}^r$ in their local side. Then at local distributed training phase, each local node optimize their local objective function $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_{0,i}^r, \lambda_i^r)$ in their local optimization oracle. The local objective is an augmented Lagrange $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_{0,i}^r, \lambda_i^r)$ defined as

$$\mathcal{L}_i(\mathbf{x}', \mathbf{x}_0, \lambda_i) \triangleq F_i(\mathbf{x}') + \langle \lambda_i, \mathbf{x}' - \mathbf{x}_0 \rangle + \frac{\gamma}{2} ||\mathbf{x}' - \mathbf{x}_0||_2^2 \tag{4}$$

, where λ_i^r is defined as the dual variable kept at distributed node i that has the same dimension as the parameters $\mathbf{x}$. Then each node returns a $\mathbf{x}_i^{r+1}$ from their local optimization oracle on $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_{0,i}^r, \lambda_i^r)$. Then each distributed node use $\mathbf{x}_i^{r+1}$ and $\mathbf{x}_0^r$ to update its dual variable from λ_i^r to λ_i^{r+1}. Then each distributed node use its updated dual λ_i^{r+1} and parameters $\mathbf{x}_i^{r+1}$ for a new $\mathbf{x}_{0,i}^{r+}$ and send $\mathbf{x}_{0,i}^{r+}$ to centralized coordinate nodes. The centralized nodes aggregates $\mathbf{x}_{0,i}^{r+}$ from all distributed node i and use Fedavg to update for a new $\mathbf{x}_0^{r+1}$.

And we define the minibatch loss function $F_i(\mathbf{x}, \xi_{i,\mathcal{B}_{i,t}})$ as

$$F_i(\mathbf{x}, \xi_{i,q}) \triangleq \frac{1}{|\mathcal{B}_{i,t}|} \sum_{b_j \in \mathcal{B}_{i,t}} l(\nu(\mathbf{x}, \mathcal{D}_{i,b_j}), \mathcal{Y}_{i,b_j}) \tag{5}$$

Finally, we define the gradient $h(\mathbf{x}_i^{r,q}, \xi_{i,\mathcal{B}_{i,q}})$ taken at global round r, local round q and node i is defined as

$$h(\mathbf{x}_i^{r,q}, \xi_{i,\mathcal{B}_{i,q}}) = \nabla_{\mathbf{x}'} \mathcal{L}_i(\mathbf{x}_i^{r,q}, \mathbf{x}_{0,i}^r, \lambda_i^r, \xi_{i,\mathcal{B}_{i,q}}) \tag{6}$$

$$= \nabla_{\mathbf{x}} F_i(\mathbf{x}_i^{r,q}, \xi_{i,\mathcal{B}_{i,q}}) + \gamma(\mathbf{x}_i^{r,q} - \mathbf{x}_{0,i}^r) + \lambda_i^r \tag{7}$$

2.2 Stochastic Gradient Langevine Dynamics

Langevine Dynamics is a family of Gaussian noise diffusion on Force Field $\nabla F(F(\mathbf{x}))$. Its continuous time Ito diffusion could be written as

$$d\mathbf{x}_t = -\nabla_{\mathbf{x}} F(\mathbf{x}) dt + \beta^{-\frac{1}{2}} dB_t \tag{8}$$

,where $B_t \in \mathbb{R}_p$ is a p-dimensional Brownian motion. Function F as $F : \mathbb{R}^p \to \mathbb{R}$ are assumed to satisfy Lipschitz continuous condition. Stochastic Gradient Langevine dynamics could be a discrete form of Langevine Dynamics as a Euler-Maruyama approximation of the stochastic ordinary equation(SDE). The discretization has a form of Gaussian Noisy injected Gradient. We write their dicretization in the following form

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nabla_{\mathbf{x}} F(\mathbf{x}) \Delta t + \mathcal{N}(0, \Delta t \beta^{-1} \mathbf{I}) \tag{9}$$

By written $\mathbf{x}^n$ as $\mathbf{x}_{t+1}$, Δt as η_n, we could write the SGLD in the form of step-wise gradient descent plus an Gaussian Noise term to perform Bayesian samplings

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \eta_n \nabla_{\mathbf{x}} F(\mathbf{x}) + \mathcal{N}(0, \eta_n \beta^{-1} \mathbf{I}) \tag{10}$$

By seeing the noise injected descending steps as a Markov chain, the stationary distribution would reduce to the following form

$$p(\mathbf{x}) \propto e^{-\beta F(\mathbf{x})} \tag{11}$$

3 Problem Formulation

3.1 An joint Min-Max Objective for Federated Stochastic Gradient MCMC

We formulate the problem of our federated stochastic gradient MCMC as optimizing the joint min-max function

$$\max_{\mathbf{x}} \min_{\lambda_i} F(\mathbf{x}, \lambda_i) = \sum_{i=1}^{N} \frac{1}{N} \log \int_{\mathbf{x}'} \exp[\beta(-F_i(\mathbf{x}') - < \lambda_i, \mathbf{x}' - \mathbf{x} > - \frac{\gamma}{2} ||\mathbf{x}' - \mathbf{x}||_2^2)] d\mathbf{x}' \tag{12}$$

The gradient of the $F(\mathbf{x}, \lambda_1, \ldots \lambda_n)$ at $\mathbf{x}_0$ could be given by

$$\frac{\delta F}{\delta \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}_0} = \sum_{i=1}^{N} \frac{1}{N}(\mathbb{E}_{P_i(\mathbf{x}'|\mathbf{x}_0)}\mathbf{x}' + \lambda_i - \mathbf{x}_0) \tag{13}$$

where we denote

$$\mathbf{x}_{0,i}^{+} = \mathbf{x}_i + \lambda_i \tag{14}$$

$$\mathbf{x}_i = \mathbb{E}_{P_i(\mathbf{x}'|\mathbf{x}_0)}\mathbf{x}' \tag{15}$$

So we could rewrite the gradient calculation as

$$\frac{\delta F}{\delta \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}_0} = \sum_{i=1}^{N} \frac{1}{N}(\mathbf{x}_{0,i}^{+} - \mathbf{x}_0) \tag{16}$$

The gradient of the $F(\mathbf{x}, \lambda_1, ...\lambda_n)$ at $\mathbf{x}_0, \lambda_i$ could be written as

$$\frac{\delta F}{\delta \lambda_i}\bigg|_{\mathbf{x}=\mathbf{x}_0} = \mathbf{x}_0 - \mathbf{x}_i \tag{17}$$

where the distribution $P_i(\mathbf{x}'|\mathbf{x}_0)$ could be written as

$$P_i(\mathbf{x}'|\mathbf{x}_0) \propto \exp[-\beta \mathcal{L}_i(\mathbf{x}', \mathbf{x}_0, \lambda_i)] \tag{18}$$

where $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_0)$ follows our previous definition as

$$\mathcal{L}_i(\mathbf{x}', \mathbf{x}_0, \lambda_i) \triangleq F_i(\mathbf{x}') + < \lambda_i, \mathbf{x}' - \mathbf{x}_0 > + \frac{\gamma}{2}||\mathbf{x}' - \mathbf{x}_0||_2^2 \tag{19}$$

As a federated learning implementation, the computation of $\delta F/\delta \mathbf{x}$ is distributed among local nodes. In one round of learning, each local node i first use Monte-Carlo estimation of $\mathbf{x}_i$ from the samples along SGLD steps on function $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_0)$ using mini-batch update from its private data. Then each local node i updates its private owned dual variable λ_i by equation by Equation 17. Next, each local node i computes their contributing part of $\mathbf{x}_{0,i}^{+}$ by Eq. 14 and sends it to the server node. Then the server node averages its aggregated $\mathbf{x}_{0,i}^{+}$ from all local nodes for $\delta F/\delta \mathbf{x}$ by Eq. 13 and uses gradient descent to update parameter $\mathbf{x}$. Finally the server node broadcast its update global parameters $\mathbf{x}$ back to each distributed nodes. The algorithm of dual descent on $\mathbf{x}$, $\lambda_1, \ldots, \lambda_N$ is shown in Algorithm 1.

Algorithm 1: Our Federated Stochastic Gradient MCMC Algorithm

Input: : $\mathbf{x}^0, \eta, p, T$
Initialize: $\mathbf{x}_0^0 = x^0$,
for $r = 0, \ldots, T-1$ **do**
  for $i = 1, \ldots, N$ *in parallel* **do**
    Local Update:
    $\mathcal{L}_i(\mathbf{x}', \mathbf{x}_{0,i}^r, \lambda_i^r) = -F_i(\mathbf{x}') - < \lambda_i^r, \mathbf{x}_{0,i}^r - \mathbf{x}' > -\frac{\gamma}{2}||\mathbf{x}' - \mathbf{x}_{0,i}^r||_2^2)$
    $\mathbf{x}_i^{r+1} = \text{SGLD-Oracle}_i(\mathcal{L}_i(\mathbf{x}', \mathbf{x}_{0,i}^r, \lambda_i^r))$
    $\lambda_i^{r+1} = \lambda_i^r + \eta\gamma\left(\mathbf{x}_i^{r+1} - \mathbf{x}_{0,i}^r\right)$
    $\mathbf{x}_{0,i}^{r+} = \mathbf{x}_i^{r+1} + \frac{1}{\gamma}\lambda_i^{r+1}$
  Global Communicate:
    Aggregate:
    $\mathbf{x}_0^{r+1} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{0,i}^{r+}$
    Broadcast:
    $\mathbf{x}_{0,i}^{r+1} = \mathbf{x}_0^r + \eta(\mathbf{x}_0^{r+1} - \mathbf{x}_0^r), i = 1, \ldots, N$

3.2 SGLD as Local Oracle Stochastic Gradient MCMC

In the inner-loop of our federated learning algorithm, each local node computes $\mathbb{E}_{P_i(\mathbf{x}'|\mathbf{x}_0)}$ by taking SG-MCMC steps in their SGLD-Oracle. In their local SGLD-Oracle, the distribution of $P_i(\mathbf{x}'|\mathbf{x}_0)$ is approximated by samplings along the markov chains of SGLD on the objective function of its local augmented Lagarange in Eq. 19 . In our implementation of local SGLD-Oracle, we take several epochs of SGLD without taking sampling in the early burn in period. To have a quick burn in times, we keep the step-size of SGLD fixed in our burn in period. After burn in, we give two SGLD sampling algorithms for $\mathbb{E}_{P_i(\mathbf{x}'|\mathbf{x}_0)}$ with fixed stepsize and decreasing stepsize in a rate of $\eta_T = O(T^{-1/3})$(Teh et al., 2016; Chen et al., 2019) to obtain a optimal mean square error bound.

Algorithm 2: SGLD-Oracle

Input: :Local Dataset ξ, number of local iterations Q, clip norm length L, base step-size $\{\eta\}$
Initialize: $\mathbf{x}_i^{r+1,0} = x_{0,i}^{r+1}, \eta_t = \eta \mathbf{x}_i^{r+1} = 0,$
for $q = 0, \ldots, Q$ **do**
  Sample a mini-batch $\xi_{i,\mathcal{B}_{i,q}}$
  Calculate gradients $h(\mathbf{x}_i^{r,q}, \xi_{i,\mathcal{B}_{i,q}}) = \nabla_{\mathbf{x}'} \mathcal{L}_i(\mathbf{x}_i^{r,q}, \cdot, \cdot, \xi_{i,\mathcal{B}_{i,q}})$
  Clip norm : $\hat{h}(\cdot) = h(\cdot)/\max(1, \frac{||h(\cdot)||_2}{L})$
  if $q > Q_0$*(Decreasing Steps)* **then**
    $\eta_t = (q - Q_0)^{-1/3}\eta$
  Noisy Gradient Descent: $\mathbf{x}_i^{r,q+1} = \mathbf{x}_i^{r,q} - \eta_t \hat{h}(\mathbf{x}_i^{r,q}, \xi_{i,\mathcal{B}_{i,q}}) + \sqrt{\eta_t}\epsilon\mathcal{N}(0, \mathbf{I})$
  if $q = Q_0$ **then**
    $\mathbf{x}_i^{r+1} = \mathbf{x}_i^{r,q+1}$
  if $q > Q_0$ **then**
    $\mathbf{x}_i^{r+1} = \sigma\mathbf{x}_i^{r+1} + (1 - \sigma)\mathbf{x}_i^{r,q+1}$

Return: $\mathbf{x}_i^{r+1}$

3.3 DIFFERENTIAL PRIVACY ANALYSIS

In each node's local optimization oracle, Q rounds of mini-batch gradient descents are taken. In round q, node i samples a minibatch $\mathcal{B}_{i,t} \triangleq \{b_1, b_2, \ldots, b_{|\mathcal{B}_{i,t}|} | \forall j, b_j \in \{1, \ldots |\mathcal{D}_i|\}\}$. The subsample follows Poisson sampling method, which is defined as follows.

Definition 3.1. (**PoissonS ample**). Given a dataset X, the procedure PoissonSample outputs a subset of the data $\{x_i \mid \sigma_i = 1, i \in [n]\}$ by sampling $\sigma_i \sim \mathrm{Ber}(p)$ independently for $i = 1, \ldots, n$.

Definition 3.2. (**Gradient Clipping**). The clipping operation is defined as

$$\mathrm{CL}(g; C) \triangleq \frac{g}{\max\left(1, \frac{\|g\|}{C}\right)}.$$

Hence, $\|g\| \leq C$.

4 A STUDY ON MIN-MAX VARIATIONAL BOUND

Theorem 4.1. *$F(\mathbf{x}, \lambda_i)$ is an upper bound on* $-\frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{x})$

Proof. Using the second order Taylor approximation of $F_i(\mathbf{x}')$ around $\mathbf{x}$, we have

$$F_i(\mathbf{x}') \approx F_i(\mathbf{x}) + \nabla F_i(\mathbf{x}) < \mathbf{x}' - \mathbf{x} > + \frac{1}{2}||\mathbf{x}' - \mathbf{x}||_{H_i} \tag{20}$$

where $H_i = \nabla_2 F_i(\mathbf{x})$ is the Hessian matrix. Using the above equation, we have

$$\log \int_{\mathbf{x}'} \exp(-F_i(\mathbf{x}') - < \lambda_i, \mathbf{x}' - \mathbf{x} > - \frac{\gamma}{2}||\mathbf{x}' - \mathbf{x}||_2^2) d\mathbf{x}' \tag{21}$$

$$\approx \log \int_{\mathbf{x}'} \exp(-F_i(\mathbf{x}) - < -\lambda_i - \nabla F_i(\mathbf{x}), \mathbf{x} - \mathbf{x}' > - \frac{1}{2}||\mathbf{x}' - \mathbf{x}||_{H_i + \gamma\mathbf{I}}) d\mathbf{x}' \tag{22}$$

$$= \log \int_{\mathbf{x}'} \exp[-F_i(\mathbf{x}) + \frac{1}{2}||-\lambda_i - \nabla F_i(\mathbf{x})||_{[H_i + \gamma\mathbf{I}]^{-1}} \tag{23}$$

$$
\begin{aligned}
&-\frac{1}{2}||\mathbf{x}-\mathbf{x}'+[H_i+\gamma\mathbf{I}]^{-1}[-\lambda_i-\nabla F_i(\mathbf{x})]^T||_{H_i+\gamma\mathbf{I}}]d\mathbf{x}' && (24)\\
=&-F_i(\mathbf{x})+\frac{1}{2}||-\lambda_i-\nabla F_i(\mathbf{x})||_{[H_i+\gamma\mathbb{I}]^{-1}}+\frac{M}{2}\log\pi-\frac{1}{2}\log\det|H_i+\gamma\mathbf{I}|\\
&+\int_{\mathbf{x}}' \mathcal{N}(\mathbf{x}';\mathbf{x}-[H_i+\gamma\mathbf{I}]^{-1}[\nabla F_i(\mathbf{x})+\lambda_i]^T,[H_i+\gamma\mathbf{I}]^{-1})d\mathbf{x}' && (25)\\
=&-F_i(\mathbf{x})+\frac{1}{2}||-\lambda_i-\nabla F_i(\mathbf{x})||_{[H_i+\gamma\mathbf{I}]^{-1}}+\frac{M}{2}\log\pi-\frac{1}{2}\log\det|H_i+\gamma\mathbf{I}| && (26)\\
\geq&-F_i(\mathbf{x})+\frac{M}{2}\log\pi-\frac{1}{2}\log\det|H_i+\gamma\mathbf{I}| && (27)\\
=&-F_i(\mathbf{x})+\text{const} && (28)
\end{aligned}
$$

So we have

$$
\begin{aligned}
&F(\mathbf{x},\lambda_i) && (29)\\
=&\sum_{i=1}^{N}\frac{1}{N}\log\int_{\mathbf{x}'}\exp(-F_i(\mathbf{x}')-<\lambda_i,\mathbf{x}'-\mathbf{x}>-\frac{\gamma}{2}||\mathbf{x}'-\mathbf{x}||_2^2)d\mathbf{x}' && (30)\\
\geq&-\sum_{i=1}^{N}\frac{1}{N}F_i(\mathbf{x})+\text{const} && (31)
\end{aligned}
$$

□

In the above theorem, we find that our federated learning algorithm's joint min-max objective $F(\mathbf{x},\lambda)$ is an upper bound on the averages of loss functions on all nodes. And the saddle point $\mathbf{x}^\star,\lambda_i^\star$ satisfies the condition that $\lambda_i+\nabla F_i(\mathbf{x})=0$. This condition is in accordance with the optimal primal-dual conditions in augmented Lagrange where the gap between local gradient zeros and global gradient zeros are closed by the dual parameters λ_i.

Theorem 4.2. *If* $\sum_{i=1}^{N}\lambda_i=0$*,* $F(\mathbf{x},\lambda_i)$ *is an upper bound on* $\log\int_{\mathbf{x}'}\exp(\frac{1}{N}\sum_{i=1}^{N}-F_i(\mathbf{x}')-\frac{\gamma}{2}||\mathbf{x}'-\mathbf{x}||_2^2)d\mathbf{x}')$

Proof. From Eq. 26, we have

$$
\begin{aligned}
&F(\mathbf{x},\lambda_i)\\
\approx&\frac{1}{N}\sum_{i=1}^{N}-F_i(\mathbf{x})+\frac{1}{2}||\lambda_i+\nabla F_i(\mathbf{x})||_{[H_i+\gamma\mathbf{I}]^{-1}}+\frac{M}{2}\log\pi-\frac{1}{2}\log\det|H_i+\gamma\mathbf{I}| && (32)
\end{aligned}
$$

From Cauchy–Schwarz inequality, we have

$$
\frac{1}{N}\sum_{i=1}^{N}||\lambda_i+\nabla F_i(\mathbf{x})||_{[H_i+\gamma\mathbf{I}]^{-1}}\geq||\frac{1}{N}\sum_{i=1}^{N}\lambda_i+\nabla F_i(\mathbf{x})||_{[H+\gamma\mathbf{I}]^{-1}} \qquad (33)
$$

where $H=\frac{1}{N}\sum_{i=1}^{N}H_i$

By substituting inequaliy 33 into Eq. 32, we have

$$
\begin{aligned}
&F(\mathbf{x},\lambda_i)\\
\geq&\frac{1}{N}\sum_{i=1}^{N}+F_i(\mathbf{x})+\frac{1}{2}||\frac{1}{N}\sum_{i=1}^{N}\lambda_i-\nabla F_i(\mathbf{x})||_{[H+\gamma\mathbf{I}]^{-1}}+\frac{M}{2}\log\pi-\frac{1}{2}\log\det|H_i+\gamma\mathbf{I}| && (34)\\
=&\frac{1}{N}\sum_{i=1}^{N}-F_i(\mathbf{x})+\frac{1}{2}||\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x})||_{[H+\gamma\mathbf{I}]^{-1}}+\frac{M}{2}\log\pi-\det|H+\gamma\mathbf{I}|\\
&+\det|H+\gamma\mathbf{I}|-\det|H_i+\gamma\mathbf{I}| && (35)\\
\approx&\log\int_{\mathbf{x}'}\exp(\frac{1}{N}\sum_{i=1}^{N}-F_i(\mathbf{x}')-\frac{\gamma}{2}||\mathbf{x}'-\mathbf{x}||_2^2)d\mathbf{x}')\\
& && (36)
\end{aligned}
$$

$$-\frac{1}{N}\sum_{i=1}^{N}\det|H_i+\gamma\mathbf{I}|+\det|H+\gamma\mathbf{I}| \tag{37}$$

$$=\log\int_{\mathbf{x}'}\exp(\frac{1}{N}\sum_{i=1}^{N}-F_i(\mathbf{x}')-\frac{\gamma}{2}||\mathbf{x}'-\mathbf{x}||_2^2)d\mathbf{x}')+\text{const} \tag{38}$$

□

In the above theorem, we find that by introducing dual parameters λ_i, the averages of our local FL objectives have an upper bound of the same function that the local loss is replace the global loss $\frac{1}{N}\sum_i F_i(\mathbf{x})$. The upper bound is achieved in either of two conditions. The Hessian matrix H_i of different nodes have the same value. Or the zeros gradient gap among $\nabla F_i(\mathbf{x})$ is closed by the duality parameters λ_i.

5 Convergence Analysis

In this section, we study the convergence properties of our algorithm. In our analyse, we first see the whole SG-MCMC Federated learning process as a homogenization of a stochastic differential equations(SDE) in the limit of step size variables $\epsilon \to 0$. Then we use a technique similar as(Eberle et al., 2019) to analyze the couplings and contraction of two independent randomly initialized stochastic process. And we derive a exponential bound of convergence of any two process on time in the metric of our defined Wasserstein Distance.

5.1 Homogenization of SDE systems

Theorem 5.1. *Consider of the SDE system given by*

$$d\mathbf{x}_0(t)=-[\mathbf{x}_0-\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i+\frac{1}{\gamma}\lambda_i)]dt \tag{39}$$

$$d\lambda_i(t)=-\gamma_1(\mathbf{x}_0-\mathbf{x}_i)dt \tag{40}$$

$$d\mathbf{x}_i(t)=-\frac{1}{\epsilon}[\nabla F_i(\mathbf{x}_i)+\gamma(\mathbf{x}_i-\mathbf{x}_0)+\lambda_i]dt+\sqrt{\frac{\beta}{\epsilon}}dB_t \tag{41}$$

It follows that in the limit of $\epsilon \to 0$, the dynamics of $d\mathbf{x}_0(t)$ and $d\lambda_i(t)$ converges to

$$d\mathbf{x}_0(t)=-[\mathbf{x}_0-\frac{1}{N}\sum_{i=1}^{N}(\int_{\mathbf{x}_i}\mathbf{x}_iP_i(d\mathbf{x}_i,\mathbf{x}_0(t))+\frac{1}{\gamma}\lambda_i)]dt \tag{42}$$

$$d\lambda_i(t)=-\gamma_1(\mathbf{x}_0-\int_{\mathbf{x}_i}\mathbf{x}_iP_i(d\mathbf{x}_i,\mathbf{x}_0(t))dt \tag{43}$$

$$\tag{44}$$

Proof. The proof follows Sec. 4.1 in(Chaudhari et al., 2018) □

In above theorem, we would assume our SG-MCMC Federated learning process as a discretization and homogenization of a stochastic differential equations(SDE).

5.2 Contraction and Coupling rate of SDE

Let probability measures $\mu(\mathbf{x}_i(0),\mathbf{x}_0(0),\lambda_i(0))$ and $\mu'(\mathbf{x}'_i(0),\mathbf{x}'_0(0),\lambda'_i(0))$ be any two probability measures on the initial distribution of $\mathbf{x}_i, x_0, \lambda_i$. And we denote μp_t as the distribution of $\mu(\mathbf{x}_i(t),\mathbf{x}_0(t),\lambda_i(t))$ of the process defined in SDE(40, 41, 39) with its initial distribution as μ. And we have the following theorem on the exponential rate couplings and contractions of two process

Theorem 5.2. *There exists a contant c, and a metric $\rho((\mathbf{x}'_i, \mathbf{x}'_0, \lambda'_i), (\mathbf{x}_i, \mathbf{x}_0, \lambda_i))$ such that for any $t \geq 0$ and any probability measure*

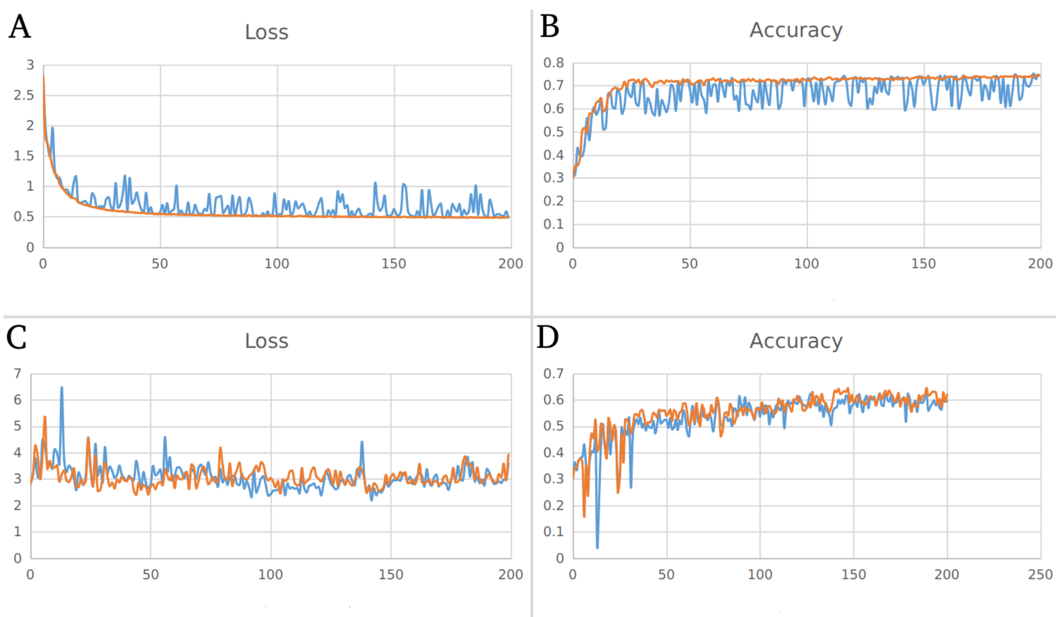
$$\mathcal{W}_\rho(\mu p_t, \mu' p_t) \leq c^{-ct} \mathcal{W}_\rho(\mu, \mu') \tag{45}$$

$\mathcal{W}_\rho$ is a Wassertein distance defined on the metric $\rho((\mathbf{x}'_i, \mathbf{x}'_0, \lambda'_i), (\mathbf{x}_i, \mathbf{x}_0, \lambda_i))$

Proof. The proof appears in our Appendix.7.1 ☐

6 EXPERIMENTS

In this section, we run simulations on Federated Learning Benchmark in (Shamir et al., 2014; Li et al., 2018) to verify our algorithms. The data is heterogeneously distributed among devices. We test our algorithm by comparing it with baseline in both low noise and high noise case. Our result is shown in Fig.6. The performance of our methods is shown in red line while the baseline method is in blue line. The high noise case is shown in the lower section. And the low noise case is shown in the upper section.



6.0.1 SYNTHETIC DATA

In particular, for each device k, we generate data with a generation distribution of $y = \arg\max(\textit{softmax}(Wx + b))$. We model $W_k \sim \mathcal{N}(u_k, 1), b_k \sim \mathcal{N}(u_k, 1), u_k \in \mathcal{N}(0, \alpha), x_k \sim (v_k, \Sigma), v_k \sim (0, \beta + 1)$.

6.0.2 LOW NOISE CASE

In this case, we inject a tiny noise of $\beta = 10^{-4}$ and compares our algorithm with the baseline where no noise is injected in local FedPD optimizations. Our algorithms have a significantly lower training loss error and with a much smoother training curve. Because local nodes run SGLD to infer parameters with lower generalization error bounds.

6.1 HIGH NOISE CASE

In this case, we inject a significant amount of noise $\beta = 0.5$ and compares our algorithm with the baseline where the same amount of noise is injected in local update without sampling in SGLD.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Chandrasekaran Anirudh Bhardwaj. Adaptively preconditioned stochastic gradient langevin dynamics. *arXiv preprint arXiv:1906.04324*, 2019.

Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

Changyou Chen, Wenlin Wang, Yizhe Zhang, Qinliang Su, and Lawrence Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient mcmc. *Science China Information Sciences*, 62(1):12101, 2019.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Andreas Eberle, Arnaud Guillin, Raphael Zimmer, et al. Couplings and quantitative contraction rates for langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

Bai Li, Changyou Chen, Hao Liu, and Lawrence Carin. On connecting stochastic gradient mcmc and differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 557–566. PMLR, 2019a.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019c.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 2019.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008, 2014.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pp. 6526–6535, 2019.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

7 Appendix

7.1 An Contraction and Coupling Rate Analysis on Convergence

The Fokker-Plank equation of the SDE. (39,41, 40) could be written as

$$\begin{aligned}\mathcal{L} &= \frac{\beta}{2\epsilon}\sum_i \triangle_{\mathbf{x}_i} - \frac{1}{\epsilon}\sum_i [\nabla F_i(\mathbf{x}_i) + \gamma(\mathbf{x}_i - \mathbf{x}_0) + \lambda_i]\cdot\nabla_{\mathbf{x}_i} \\ &\quad - \sum_i \gamma_1(\mathbf{x}_0 - \mathbf{x}_i)\cdot\nabla_{\lambda_i} - [\mathbf{x}_0 - \frac{1}{N}\sum_{i=1}^N(\mathbf{x}_i + \frac{1}{\gamma}\lambda_i)]\cdot\nabla_{\mathbf{x}_0} \end{aligned} \tag{46}$$

We consider the following Lyapunov as

$$\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) = \sum_i F_i(\mathbf{x}_i) + \frac{A}{2}|\mathbf{x}_i|^2 + \frac{B}{2}|\mathbf{x}_i + \zeta\lambda_i|^2 + \frac{C}{2\epsilon}|\mathbf{x}_0|^2 \tag{47}$$

Following the line of the work(Eberle et al., 2019), we make the following assumptions on functions $F_i(\mathbf{x})$

Assumption A1.

$$F_i(\mathbf{x}) \geq 0 \tag{48}$$

$$|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}| \tag{49}$$

$$\mathbf{x}\cdot\nabla F_i(\mathbf{x})/2 \geq \kappa(F_i(\mathbf{x}) + z|\mathbf{x}|^2/4) - F \tag{50}$$

$$|F_i(\mathbf{x})| \leq G \tag{51}$$

Then we have the following lemma

Lemma 7.1. *If the above assumption holds, then* $\mathcal{L}\mathcal{V} \leq \frac{1}{\epsilon}(M\beta(A+B+L) + DG + (A+B)F - \kappa(A+B+\gamma)\mathcal{V})$

Proof. By applying Fokker-Plank Eq. 46, we have

$$\mathcal{L}F_i(\mathbf{x}_i) = \frac{\beta}{\epsilon} \triangle_{\mathbf{x}_i} F_i(\mathbf{x}_i) - \frac{1}{\epsilon}[|\nabla F_i(\mathbf{x}_i)|^2 + \gamma \nabla F_i(\mathbf{x}_i)\cdot \mathbf{x}_i - \gamma \nabla F_i(\mathbf{x}_i)\cdot \mathbf{x}_0 + \nabla F_i(\mathbf{x}_i)\cdot \lambda_i] \tag{52}$$

$$\mathcal{L}\frac{1}{2}|\mathbf{x}_i|^2 = \frac{M\beta}{\epsilon} - \frac{1}{\epsilon}[\nabla F_i(\mathbf{x}_i)\cdot \mathbf{x}_i + \gamma|\mathbf{x}_i|^2 - \gamma \mathbf{x}_i\cdot \mathbf{x}_0 + \mathbf{x}_i \cdot \lambda_i] \tag{53}$$

$$\mathcal{L}\frac{1}{2}|\mathbf{x}_0|^2 = -[|\mathbf{x}_0|^2 - \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i\cdot \mathbf{x}_0 + \frac{1}{\gamma}\lambda_i \cdot \mathbf{x}_0)] \tag{54}$$

$$\begin{aligned}\mathcal{L}\frac{1}{2}|\mathbf{x}_i + \zeta\lambda_i|^2 = & \frac{M\beta}{\epsilon} - \frac{1}{\epsilon}[\nabla F_i(\mathbf{x}_i)\cdot \mathbf{x}_i + \gamma|\mathbf{x}_i|^2 - \gamma \mathbf{x}_i\cdot \mathbf{x}_0 + \mathbf{x}_i\cdot\lambda_i \\ & + \zeta\nabla F_i(\mathbf{x}_i)\cdot\lambda_i + \zeta\gamma \mathbf{x}_i\cdot\lambda_i - \zeta\gamma\lambda_i\cdot \mathbf{x}_0 + \zeta|\lambda_i|^2] \qquad (55) \\ & - \gamma_1\zeta(\mathbf{x}_0\cdot \mathbf{x}_i - |\mathbf{x}_i|^2) - \gamma_1\zeta^2(\mathbf{x}_0\cdot\lambda_i - \mathbf{x}_i\cdot\lambda_i) \qquad (56)\end{aligned}$$

As

$$\triangle_{\mathbf{x}_i} F_i(\mathbf{x}_i) \le L, \quad |F_i(\mathbf{x}_i)| \le G^2, \quad \mathbf{x}_i\cdot\nabla F_i(\mathbf{x}_i)/2 \ge \kappa(F_i(\mathbf{x}_i) + z|\mathbf{x}_i|^2/4) - F \tag{57}$$

Then we have

$$\begin{aligned}&\mathcal{L}(\sum_i F_i(\mathbf{x}_i) + \frac{A}{2}|\mathbf{x}_i|^2 + \frac{B}{2}|\mathbf{x}_i + \zeta\lambda_i|^2 + \frac{C}{2\epsilon}|\mathbf{x}_0|^2)) \qquad (58)\\ &\le \frac{1}{\epsilon}M((B+L)/\beta) + DG + (A+B)F] \\ &- \sum_i \frac{1}{\epsilon}\{\frac{(A+\gamma+B)z\kappa}{4} + [(A+B)\gamma - B\gamma_1\epsilon\zeta]|\mathbf{x}_i|^2 + B\zeta|\lambda_i|^2 + \frac{C}{N}|\mathbf{x}_0|^2 + (1+D)\nabla F_i(\mathbf{x}_i)^2 \\ &- [(A+B)\gamma - B\gamma_1\zeta\epsilon + \frac{C}{N}]\mathbf{x}_i\cdot \mathbf{x}_0 - [B\zeta\gamma + \frac{C}{\gamma N} - B\gamma_1\epsilon\zeta^2]\mathbf{x}_0\cdot\lambda_i \\ &+ [A + B + B\zeta\gamma - B\gamma_1\zeta^2\epsilon]\mathbf{x}_i\cdot\lambda_i + \nabla F_i(\mathbf{x}_i)\cdot[(1+B\zeta)\lambda_i - \gamma\mathbf{x}_0]\} \qquad (59)\end{aligned}$$

By choosing the proper values of $A, B, C, D, \zeta, z, \gamma_1$, we could let the following equality holds

$$\mathcal{L}\mathcal{V} \le \frac{1}{\epsilon}(M\beta(A+B+L) + DG + (A+B)F - \kappa(A+B+\gamma)\mathcal{V}) \tag{60}$$

Here is one set of $A, B, C, D, \zeta, z, \gamma_1$ satisfying the above inequality.

$$A = B = \gamma, \quad C = \frac{7}{3}\gamma^2 N, \quad D = \frac{3}{4}\kappa\gamma^2, \quad \zeta = \frac{2}{\gamma} \tag{61}$$

where $\kappa, \gamma, \gamma_1, z$satisfying the following constraints

$$\begin{aligned}\gamma_1\epsilon &\le \frac{1}{6}\gamma^2 \\ \kappa &\le \frac{77}{150}N \\ \frac{4}{\gamma} &\ge 2 + 6\kappa\gamma + \frac{3}{2}\kappa\gamma^4 \\ \frac{3\kappa z\gamma}{4} - \frac{3\kappa}{2} &\ge \frac{8}{25}[2\gamma^2 - 2\gamma_1\epsilon]\end{aligned} \tag{62}$$

□

7.1.1 Couplings of two Process

Let $\mathbf{X}_t = [\mathbf{x}_1^T(t), \mathbf{x}_2^T(t)\dots\mathbf{x}_N^T(t)]$, $\lambda_t = [\lambda_1^T(t), \lambda_2^T(t)\dots\lambda_N^T(t)]$. We consider two coupling process $\mathbf{X}_t, \lambda_t, \mathbf{x}_0(t)$ and $\mathbf{X}'_t, \lambda'_t, \mathbf{x}'_0(t)$ with different initialization. We compose their brownian motions in the direction of synchronized drift and reflection drift which we would give conditions.

Each is governed by the following SDE

$$
\begin{aligned}
d\mathbf{X}_t &= -\frac{1}{\epsilon}[\nabla F_i(\mathbf{X}_t)dt + \gamma\mathbf{X}_t dt - \gamma\mathbf{x}_0(t)\hat{\mathbf{1}}^T dt + \lambda_t dt] \\
&\quad + \sqrt{\beta/2\epsilon} rc(Z_t, W_t, Y_t)dB_t^{rc} + \sqrt{\beta/2\epsilon} sc(Z_t, W_t, Y_t)dB_t^{sc} \\
d\mathbf{x}_0(t) &= -[\mathbf{x}_0(t)dt - \frac{1}{N}\hat{\mathbf{1}}(\mathbf{X}_t + \frac{1}{\gamma}\lambda_t)dt] \\
d\lambda_t &= -[\mathbf{x}_0(t)\hat{\mathbf{1}}dt - \mathbf{X}_t dt] \qquad (63) \\
d\mathbf{X}'_t &= -\frac{1}{\epsilon}[\nabla F_i(\mathbf{X}'_t)dt + \gamma\mathbf{X}'_t dt - \gamma\mathbf{x}_0(t)\hat{\mathbf{1}}^T dt + \lambda'_t dt] \\
&\quad + \sqrt{\beta/2\epsilon} rc(Z_t, W_t, Y_t)(\mathbf{I} - 2e_t e_t^T)dB_t^{rc} + \sqrt{\beta/2\epsilon} sc(Z_t, W_t, Y_t)dB_t^{sc} \\
d\mathbf{x}'_0(t) &= -[\mathbf{x}'_0(t)dt - \frac{1}{N}\hat{\mathbf{1}}(\mathbf{X}'_t + \frac{1}{\gamma}\lambda'_t)dt] \\
d\lambda'_t &= -[\mathbf{x}'_0(t)\hat{\mathbf{1}}^T dt - \mathbf{X}_t dt] \qquad (64)
\end{aligned}
$$

,where $\hat{\mathbf{1}}$ is defined as

$$
\hat{\mathbf{1}}_{m,n} = \begin{cases} 1 & (m-1)M + 1 \leq n \leq mM \\ 0 & \text{else} \end{cases} \qquad (65)
$$

The existence and uniqueness of decomposition holds by Levy's characterization. Then we write the differentiation of the two process as $Z_t = \mathbf{X}_t - \mathbf{X}'_t, W_t = \mathbf{x}_0(t) - \mathbf{x}'_0(t), Y_t = \lambda_t - \lambda'_t$. Moreover, we define we define $rc, sc : \mathrm{R} \to [0,1]$ are Lipschitz continuous functions such that $rc^2 + sc^2 = 1$ as a function of Z_t, W_t and Y_t

$$
rc = 0 \quad \text{if} \quad |W_t| = 0, |Y_t| = 0 \quad \text{or} \quad |Z_t| + \alpha_1|W_t| + \alpha_2|Y_t| \geq R_1 + \xi \qquad (66)
$$

$$
rc = 1 \quad \text{if} \quad \alpha_1|W_t| + \alpha_2|Y_t| \geq \xi \quad \text{and} \quad |Z_t| + \alpha_1|W_t| + \alpha_2|Y_t| \leq R_1 \qquad (67)
$$

We also define e_t as an unit length vector in the direction of Z_t and e_t shrinks at $|Z_t| = 0$

$$
e_t = Z_t/|Z_t| \quad \text{if} \quad Z_t \neq 0 \quad \text{and} \quad e_t = 0 \quad \text{if} \quad Z_t = 0 \qquad (68)
$$

The process of (Z_t, W_t, Y_t) could be written as

$$
\begin{aligned}
dZ_t &= -[\sum_i \nabla F_i(\mathbf{x}_i(t)) - F_i(\mathbf{x}'_i(t)) + \gamma Z_t dt - \gamma W_t \hat{\mathbf{1}}^T dt + Y_t dt] \\
&\quad + \sqrt{2\beta/\epsilon} rc(Z_t, W_t, Y_t)dB_t^{rc} \qquad (69) \\
dW_t &= -[W_t dt - \frac{1}{N}\hat{\mathbf{1}}(Z_t + \frac{1}{\gamma}Y_t)] \qquad (70) \\
dY_t &= -[W_t - Z_t\hat{\mathbf{1}}^T] \qquad (71)
\end{aligned}
$$

The derivative of $|W_t|$ and $|Y_t|$ could be write in the form of

$$
\begin{aligned}
\frac{d}{dt}|W_t| &= \frac{W_t}{|W_t|} \cdot -[W_t - \frac{1}{N}\hat{\mathbf{1}}(Z_t + \frac{1}{\gamma}Y_t)] \qquad (72) \\
\frac{d}{dt}|Y_t| &= \frac{Y_t}{|Y_t|} \cdot -[W_t\hat{\mathbf{1}}^T - Z_t] \qquad (73)
\end{aligned}
$$

We set

$$
\begin{aligned}
r_t &= r((\mathbf{X}_t, \mathbf{x}_0(t), \lambda_t), (\mathbf{X}'_t, \mathbf{x}'_0(t), \lambda'_t)) = |Z_t| + \alpha_1|W_t| + \alpha_2|Y_t| \qquad (74) \\
\rho_t &= \rho((\mathbf{X}_t, \mathbf{x}_0(t), \lambda_t), (\mathbf{X}'_t, \mathbf{x}'_0(t), \lambda'_t)) = f(r_t)G_t \qquad (75) \\
G_t &= 1 + \nu\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i) + \nu\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i) \qquad (76)
\end{aligned}
$$

Then we have the following lemmas

Lemma 7.2. *There exists a R_1 if $r_t \geq R_1$ such that*

$$
\mathcal{L}\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i) + \mathcal{L}\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) \leq -\frac{\kappa(A + B + \gamma)}{6\epsilon}(\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i) + \mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i)) \qquad (77)
$$

, where R_1 is give by

$$R_1 \leq [\frac{12}{5}(\frac{2}{A}(1+\frac{\alpha_2}{\zeta})^2+\frac{2\alpha_2^2}{B\zeta^2}+\frac{2\epsilon\alpha_1^2}{C})(M\beta(A+B+L)+DG+(A+B)F)/(\kappa(A+B+\gamma))]^{1/2} \tag{78}$$

Proof. We have

$$\begin{aligned} r_t &= |Z_t|+\alpha_1|W_t|+\alpha_2|Y_t| \\ &= |\mathbf{X}_t-\mathbf{X}_t'|+\alpha_1|\mathbf{x}_0(t)-\mathbf{x}_0'(t)1|+\alpha_2|\lambda_t-\lambda_t'| \\ &\leq |\mathbf{X}_t|+\alpha_1|\mathbf{x}_0(t)|+\alpha_2|\lambda_t|+|\mathbf{X}_t'|+\alpha_1|\mathbf{x}_0'(t)|+\alpha_2|\lambda_t'| \\ &\leq |\mathbf{X}_t|+\alpha_1|\mathbf{x}_0(t)|+\alpha_2(\frac{1}{\zeta})(|\mathbf{X}_t|+|\mathbf{X}_t+\zeta\lambda_t|)+|\mathbf{X}_t'|+\alpha_1|\mathbf{x}_0'(t)|+\alpha_2(\frac{1}{\zeta})(|\mathbf{X}_t'|+|\mathbf{X}_t'+\zeta\lambda_t'|) \\ &\leq (1+\frac{\alpha_2}{\zeta})|\mathbf{X}_t|+\frac{\alpha_2}{\zeta}|\mathbf{X}_t+\zeta\lambda_t||+\alpha_1|\mathbf{x}_0(t)|+(1+\frac{\alpha_2}{\zeta})|\mathbf{X}_t'|+\frac{\alpha_2}{\zeta}|\mathbf{X}_t'+\zeta\lambda_t'||+\alpha_1|\mathbf{x}_0'(t)| \end{aligned} \tag{79}$$

In Cauchy-Swartz inequality, we have

$$\begin{aligned} &\sum[\frac{A}{2}|\mathbf{x}_i|^2+\frac{B}{2}|\mathbf{x}_i+\zeta\lambda_i|^2+\frac{C}{2\epsilon}|\mathbf{x}_0^2|][\frac{2}{A}(1+\frac{\alpha_2}{\zeta})^2+\frac{2\alpha_2^2}{B\zeta^2}+\frac{2\epsilon\alpha_1^2}{C}] \\ &\geq [(1+\frac{\alpha_2}{\zeta})|\mathbf{X}_t|+\frac{\alpha_2}{\zeta}|\mathbf{X}_t+\zeta\lambda_t||+\alpha_1|\mathbf{x}_0(t)|]^2 \end{aligned} \tag{80}$$

So we have

$$r_t^2 \leq [\frac{2}{A}(1+\frac{\alpha_2}{\zeta})^2+\frac{2\alpha_2^2}{B\zeta^2}+\frac{2\epsilon\alpha_1^2}{C}](\mathcal{V}(\mathbf{x}_0',\mathbf{x}_i',\lambda_i')+\mathcal{V}(\mathbf{x}_0,\mathbf{x}_i,\lambda_i)) \tag{81}$$

So we have

$$\begin{aligned} &\mathcal{V}(\mathbf{x}_0',\mathbf{x}_i',\lambda_i')+\mathcal{V}(\mathbf{x}_0,\mathbf{x}_i,\lambda_i)) \\ &\geq \frac{12}{5}(M\beta(A+B+L)+DG+(A+B)F)/(\kappa(A+B+\gamma)) \end{aligned} \tag{82}$$

And thus

$$\mathcal{L}\mathcal{V}(\mathbf{x}_0',\mathbf{x}_i',\lambda_i')+\mathcal{L}\mathcal{V}(\mathbf{x}_0,\mathbf{x}_i,\lambda_i) \leq -\frac{\kappa(A+B+\gamma)}{6\epsilon}(\mathcal{V}(\mathbf{x}_0',\mathbf{x}_i',\lambda_i')+\mathcal{V}(\mathbf{x}_0,\mathbf{x}_i,\lambda_i)) \tag{83}$$

□

Lemma 7.3. *Let c,ν and suppose that $f:[0,\infty)\to[0,\infty)$ is continuous, non-decreasting, concave and C^2 except for finitely many points. The we have*

$$e^{ct}\rho_t \leq \rho_0+\int_0^t e^{cs}K_s ds+M_t \tag{84}$$

where M_t is a local continuous martingale , and K_t could be written as

$$\begin{aligned} K_t &= cf(r_t)G_t+(\frac{1}{\epsilon}L\sqrt{N}-\frac{1}{\epsilon}\gamma+\gamma_1\alpha_2+\frac{\alpha_1}{\sqrt{N}})|Z_t|+((\frac{1}{\epsilon}\gamma+\gamma_1\sqrt{N}-\alpha_1)|W_t| \\ &+(\frac{\alpha_1}{\gamma}+\frac{1}{\epsilon})|Y_t|)f_-'(r_t))G_t+\frac{\beta}{\epsilon}rc(Z_t,W_t,Y_t)^2f''(r_t)G_t \\ &+\frac{\nu}{\epsilon}f(r_t)(\frac{2}{\epsilon}(M\beta(B+L))+DG+(A+B)F-\kappa(1+\gamma)\mathcal{V}-\kappa(1+\gamma)\mathcal{V}')) \\ &+\nu\beta/\epsilon\max(L+A+B,B\zeta/\alpha_2)r_tf_-'(r_t)rc(Z_t,W_t,Y_t)^2 \end{aligned} \tag{85}$$

Proof. We apply Ito's formula on the process of $|Z_t|$

$$|Z_t|=|Z_0|+A_t^Z+\tilde{M}_t^Q \tag{86}$$

where (A_t^Q) and $(\hat{M}_t^Q)$ is absolute continuous process and martingale given by

$$A_t^Q = -\frac{1}{\epsilon}\int_0^t e_s^T \cdot (\sum_i(\nabla F_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}'_i)) + \gamma Z_t - \gamma W_t + Y_t) \tag{87}$$

$$\hat{M}_t^Q = \sqrt{2\beta/\epsilon}\int_0^t rc(Z_t, W_t, Y_t) e_s^T dB_s^{rc} \tag{88}$$

Because $\delta^2_{z/|z|}|z|= 0$, there is no Ito's correlation. By the Lipschitz continuous on , we could have

$$A_t^Q \leq \frac{1}{\epsilon}\int_0^t \sum_i L||\mathbf{x}_i - \mathbf{x}_i||-\gamma|Z_t|+\gamma|Y_t|+|W_t|dt \tag{89}$$

$$= \frac{1}{\epsilon}\int_0^t (L\sqrt{N} - \gamma)|Z_t|+\gamma|W_t|+|Y_t| \tag{90}$$

And then we write the semimartingale decompostion of r_t

$$r_t = |Q_0|+\alpha_1|W_t|+\alpha_2|Y_t| \tag{91}$$

Similarly, we have the following bound on $d|W_t|$ and $d|Y_t|$

$$\frac{d}{dt}|W_t| \leq -|W_t|+\frac{1}{\sqrt{N}}(|Z_t|+\frac{1}{\gamma}|Y_t|)] \tag{92}$$

$$\frac{d}{dt}|Y_t| \leq \gamma_1[\sqrt{N}|W_t|+|Z_t|] \tag{93}$$

Since by assumption, f is concave and C^2, we can now apply Ito-Tanaka formula to $f(r_t)$. Let f' and f'' denote the left-sided first derivative and almost everywhere defined second order derivative. We obtain the following semimartingale decompostion bound on $e^{ct}f(r_t)$

$$e^{ct}f(r_t) = f(r_0) + \tilde{A}_t + \tilde{M}_t \tag{94}$$

with the martingale part

$$\tilde{M}_t = \sqrt{2\beta/\epsilon}\int_0^t e^{cs} f'_-(r_s) rc(Z_t, W_t, Y_t) e_s^T dB_s^{rc} \tag{95}$$

and a continuous finite-variation process $(\tilde{A}_t)$ is bounded by

$$d\tilde{A}_t \leq (cf(r_t) + (\frac{1}{\epsilon}L\sqrt{N} - \frac{1}{\epsilon}\gamma + \gamma_1\alpha_2 + \frac{\alpha_1}{\sqrt{N}})|Z_t|+((\frac{1}{\epsilon}\gamma + \gamma_1\alpha_2\sqrt{N} - \alpha_1)|W_t| \tag{96}$$

$$+ (\frac{\alpha_1}{\gamma} + \frac{1}{\epsilon})|Y_t|) f'_-(r_t)) e^{ct} dt + \frac{\beta}{\epsilon} rc(Z_t, W_t, Y_t)^2 f''(r_t) e^{ct} dt \tag{97}$$

Now we bound on the integration of the process's evolution on time $G_t = 1 + \nu\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) + \nu\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i)$, by applying Ito's formula we have

$$\begin{aligned} dG_t &= \nu(\mathcal{L}\mathcal{V})(\mathbf{x}_0, \mathbf{x}_i, \lambda_i)dt + \nu(\mathcal{L}\mathcal{V})(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i)dt \\ &+\nu\sqrt{\beta/2\epsilon}(\nabla_{\mathbf{x}_i}\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) - \nabla_{\mathbf{x}'_i}\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i)) e_t e_t^T rc(Z_t, W_t, Y_t)^2 dB_t^{rc} \\ &+\nu\sqrt{\beta/2\epsilon}(\nabla_{\mathbf{x}'_i}\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) + \nabla_{\mathbf{x}_i}\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i))(\mathbf{I} - e_t e_t^T) rc(Z_t, W_t, Y_t)^2 dB_t^{rc} \\ &+\nu\sqrt{\beta/2\epsilon}(\nabla_{\mathbf{x}'_i}\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) + \nabla_{\mathbf{x}_i}\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i)) sc(Z_t, W_t, Y_t)^2 dB_t^{sc} \end{aligned} \tag{98}$$

Hence by Ito's formula, we obtain the following semi-martingale decomposition

$$e^{ct}\rho_t = e^{ct}f(r_t)G_t = \rho_0 + M_t + A_t \tag{99}$$

where (M_t) is a continuous local martingale, and

$$\begin{aligned} dA_t &= G_t d\tilde{A}_t + \nu e^{ct} f(r_t)((\mathcal{L}\mathcal{V})(\mathbf{x}_0, \mathbf{x}_i, \lambda_i)dt + (\mathcal{L}\mathcal{V})(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i))dt \\ &+ e^{ct}\nu\beta/\epsilon f'_-(r_t) rc(Z_t, W_t, Y_t)^2 (\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) - \nabla\mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i))dt \end{aligned} \tag{100}$$

Now recall that by Lemma 7.1, we have

$$\mathcal{L}\mathcal{V} \leq \frac{1}{\epsilon}(\beta M(A+B+L) + DG + (A+B)F - \kappa(A+B+\gamma)\mathcal{V}) \tag{101}$$

Furthermore, $|\nabla_{\mathbf{x}_i'}\mathcal{V}(\mathbf{x}_0', \mathbf{x}_i', \lambda_i')|$ is bounded by

$$\begin{aligned}
&|\nabla_{\mathbf{x}_i'}\mathcal{V}(\mathbf{x}_0', \mathbf{x}_i', \lambda_i') - \nabla\mathcal{V}_{\mathbf{x}_i'}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i)| = |\sum_i(\nabla F_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i') + (A+B)(\mathbf{x}_i - \mathbf{x}_i') + B\zeta(\lambda_i - \lambda_i')| \\
&\leq (L+A+B)|Z_t| + B\zeta|W_t| \\
&\leq \max(L+A+B, B\zeta/\alpha_2)r_t
\end{aligned} \tag{102}$$

By combining , we finally obtain $dA_t \leq e^{ct}K_t dt$, where

$$\begin{aligned}
K_t = {} & cf(r_t)G_t + (\frac{1}{\epsilon}L\sqrt{N} - \frac{1}{\epsilon}\gamma + \gamma_1\alpha_2 + \frac{\alpha_1}{\sqrt{N}})|Z_t| + ((\frac{1}{\epsilon}\gamma + \gamma_1\alpha_2\sqrt{N} - \alpha_1)|W_t| \\
& + (\frac{\alpha_1}{\gamma} + \frac{1}{\epsilon})|Y_t|)f_-'(r_t))G_t + \frac{\beta}{\epsilon}rc(Z_t, W_t, Y_t)^2 f''(r_t)G_t \\
& + \frac{\nu}{\epsilon}f(r_t)(2(M\beta(A+B+L)) + DG + (A+B)F - \kappa(A+B+\gamma)\mathcal{V} - \kappa(A+B+\gamma)\mathcal{V}')) \\
& + \nu\beta/\epsilon\max(L+A+B, B\zeta/\alpha_2)r_t f_-'(r_t)rc(Z_t, W_t, Y_t)^2
\end{aligned} \tag{103}$$

□

Lemma 7.4. *By choosing the following ν and $f(r)$, the continuous evolving process K_t vanishes as $\xi \to 0$*

$$f(r) = \int_0^{r\wedge R_1} \varphi(s)g(s)ds \tag{104}$$

$$\varphi(s) = exp(-\frac{C_1 s^2}{2}), \tag{105}$$

$$g(r) = 1 - C_2\int_0^r \phi(s)\varphi(s)^{-1}dr, \quad \textit{with} \quad \phi(s) = \int_0^s \varphi(x)dx \tag{106}$$

$$C_1 = \nu\max(L+A+B, B\zeta/\alpha_2) + \max(\gamma + \gamma_1\alpha_2\sqrt{N}\epsilon - \epsilon\alpha_1)/\beta\alpha_1, (\frac{\alpha_1\epsilon}{\gamma} + 1)/\beta\alpha_2) \tag{107}$$

$$C_2 = \frac{9c\epsilon}{\beta} \tag{108}$$

$$4c\epsilon = \nu(M\beta(A+B+L) + DG + (A+B)F) \tag{109}$$

$$\tag{110}$$

and we assume that

$$C_4 = \frac{1}{\epsilon}L\sqrt{N} - \frac{1}{\epsilon}\gamma + \gamma_1\alpha_2 + \frac{\alpha_1}{\sqrt{N}} < 0 \tag{111}$$

Proof. To bound K_t, we consider different region to achieve up to an error term which vanishes as $\xi \to 0$

(i) $\alpha_1|Z_t| + \alpha_2|W_t| \geq \xi$ and $r_t \leq R_1$
Here we have $rc(Z_t, W_t, Y_t) = 1$. Therefore, since $G_t \geq 1, |W_t| \geq 0, |Z_t| \geq 0$ and $|Y_t| \geq 0$. We have

$$\begin{aligned}
K_t \leq {} & \frac{\beta}{\epsilon}f''(r_t)G_t + \frac{1}{\epsilon}(\nu\beta\max(L+A+B, B\zeta/\alpha_2) \\
& + \max(\gamma + \gamma_1\alpha_2\sqrt{N}\epsilon - \epsilon\alpha_1)\alpha_1, (\frac{\alpha_1\epsilon}{\gamma} + 1)\alpha_2))r_t G_t f_-'(r_t) + 9cf(r_t)G_t
\end{aligned} \tag{112}$$

Then we have

$$\frac{\beta}{\epsilon}\varphi'(r_t) + \frac{1}{\epsilon}(\nu\beta\max(L+A+B, B\zeta/\alpha_2) + \max(\gamma + \gamma_1\alpha_2\sqrt{N}\epsilon - \epsilon\alpha_1)\alpha_1, (\frac{\alpha_1\epsilon}{\gamma} + 1)\alpha_2))r_t\varphi(r_t) = 0$$

Hense we have

$$K_t \leq 9c(\int_0^{r_t} 2\varphi(s)g(s)G_t ds - \int_0^{r_t} \varphi(s)G_t) \tag{113}$$

In order to ensure $g(r) \geq 1/2$ for $r < R_1$, we have to assume

$$c \leq 2\beta/(9\epsilon \int_0^{R_1} \phi(s)\varphi(s)^{-1}ds) \tag{114}$$

(ii) $\alpha_1|Z_t|+\alpha_2|W_t|< \xi$ and $r_t \leq R_1$
With the same choise of f and $g \geq \frac{1}{2}$, similarly we derive a bound on K_t as

$$\begin{aligned} K_t \leq & (\frac{\beta}{\epsilon}f''(r_t)G_t + \frac{\beta}{\epsilon}\nu \max(L+A+B, B\zeta/\alpha_2))rc(Z_t, W_t, Y_t)^2 \\ & + (\frac{1}{\epsilon}L\sqrt{N} - \frac{1}{\epsilon}\gamma + \gamma_1\alpha_2 + \frac{\alpha_1}{\sqrt{N}})r_t f'(r_t) + 9cf(r_t)G_t + C_3\xi f(r_t)G_t \end{aligned} \tag{115}$$

where the constant C_3 is given by

$$C_3 = \max(\frac{L\sqrt{N}}{\epsilon}+\frac{1}{\epsilon}\gamma-\gamma_1\alpha_2-\frac{\alpha_1}{\sqrt{N}}+\frac{\gamma}{\epsilon\alpha_1}+\frac{\gamma_1\alpha_2\sqrt{N}}{\alpha_1}-1, \frac{L\sqrt{N}}{\epsilon}+\frac{1}{\epsilon}\gamma-\gamma_1\alpha_2-\frac{\alpha_1}{\sqrt{N}}+\frac{\alpha_1}{\gamma\alpha_2}+\frac{1}{\epsilon\alpha_2})) \tag{116}$$

In order to ensure that the upper bound converges to 0 as $\xi \to 0$, we assume

$$c \leq \frac{1}{18}(\frac{-1}{\epsilon}L\sqrt{N} + \frac{1}{\epsilon}\gamma - \gamma_1\alpha_2 - \frac{\alpha_1}{\sqrt{N}}) \inf_{r\in(0,R_1]} \frac{r\varphi(r)}{\phi(r)} \tag{117}$$

(iii) $r_t \geq R_1$. Here $f'_-(r_t) = 0$.
Let $C_5 = (M(A+B+L) + DG/\beta + (A+B)F/\beta)$
Hence we have

$$\begin{aligned} K_t = & \frac{\nu\beta}{\epsilon}[2C_5 + \frac{c\epsilon}{\nu\beta} - (\kappa(A+B+\gamma) - c\epsilon/\beta)(\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) + \mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i))]f(r_t) \\ \leq & [\frac{9}{4}C_5 - \frac{15}{16}(\kappa(A+B+\gamma) - c\epsilon/\beta)(\mathcal{V}(\mathbf{x}_0, \mathbf{x}_i, \lambda_i) + \mathcal{V}(\mathbf{x}'_0, \mathbf{x}'_i, \lambda'_i))]f(r_t) \\ \leq & 0 \end{aligned} \tag{118}$$

provided we assume

$$c \leq \frac{\beta\kappa(A+B+\gamma)}{16\epsilon} \tag{119}$$

□

We finally introduce Wassertein distance on our defined metric ρ on the probability space of our two distributions of $\mu(\mathbf{X}_t, \mathbf{x}_0(t), \lambda_t)$ and $\mu'(\mathbf{X}'_t, \mathbf{x}'_0(t), \lambda'_t)$

Definition 7.1. For probability measures $\mu(\mathbf{X}_t, \mathbf{x}_0(t), \lambda_t)$ and $\mu'(\mathbf{X}'_t, \mathbf{x}'_0(t), \lambda'_t)$ on $\mathbb{R}^{2M}$, we define

$$\mathcal{W}_\rho(\mu, \mu') = \inf_{\Gamma\in\Pi(\mu,\mu')} \rho((\mathbf{X}, \mathbf{x}_0, \lambda), (\mathbf{X}', \mathbf{x}'_0, \lambda'))\Gamma(d(\mathbf{X}, \mathbf{x}_0, \lambda), d(\mathbf{X}', \mathbf{x}'_0, \lambda')) \tag{120}$$

,where Γ is a coupling of μ and μ', our defined Wassertein distance is taking the infimum of ρ metrics over all couplings.

Then we conclude on our final theorem

Theorem 7.5. *For a positive constant c such that*

$$c \leq \min(2\beta/(9\epsilon \int_0^{R_1} \phi(s)\varphi(s)^{-1}ds), \frac{1}{18}(\frac{-1}{\epsilon}L\sqrt{N}+\frac{1}{\epsilon}\gamma-\gamma_1\alpha_2-\frac{\alpha_1}{\sqrt{N}}) \inf_{r\in(0,R_1]} \frac{r\varphi(r)}{\phi(r)}, \frac{\beta\kappa(A+B+\gamma)}{16\epsilon}) \tag{121}$$

Moreover, let $f : [0, \infty, \to [0, \infty)$ be defined above. Then for any $t \geq 0$ and for any probabilty measure μ, μ' on $\mathbb{R}^{2M}$,

$$\mathcal{W}_\rho(\mu p_t, \mu' p_t) \leq e^{-ct}\mathcal{W}_\rho(\mu, \mu') \tag{122}$$

Proof. Let Γ be a coupling of two probability measures μ and μ' such that $\mathcal{W}_\rho(\mu p_t, \mu' p_t) < \infty$. We consider two coupling process $(\mathbf{X}, \mathbf{x}_0, \lambda), (\mathbf{X}', \mathbf{x}_0', \lambda')$satisfying the initial optimal couplings $(\mathbf{X}, \mathbf{x}_0, \lambda), (\mathbf{X}', \mathbf{x}_0', \lambda') \in \Gamma$. in each of the cases conditions considered above, we obtain $K_t \leq C_3 \xi G_t$. Therefore we apply lemma7.3 and taking expectations, we have

$$\mathbb{E}[\rho_t] \leq e^{-ct}\mathbb{E}[\rho_0] + C_3\xi \int_0^t e^{c(s-t)}\mathbb{E}[G_s]ds \tag{123}$$

Note that $\mathbb{E}[G_s]$ is finite. So at the limit of $\xi \to 0$, we have

$$\mathbb{E}[\rho_t] \leq e^{-ct}\mathbb{E}[\rho_0] \tag{124}$$

Since $(\mathbf{X}_t, \mathbf{x}_0(t), \lambda_t), (\mathbf{X}_t', \mathbf{x}_0'(t), \lambda_t')$ is a coupling of μp_t and $\mu' p_t$, we have $\mathcal{W}_\rho(\mu p_t, \mu' p_t) \leq \mathbb{E}[\rho_t]$. As the initial optimal couplings conditions, we have

$$\mathbb{E}[\rho_0] = \int \rho d\Gamma = \mathcal{W}_\rho(\mu, \mu') \tag{125}$$

So we conclude

$$\mathcal{W}_\rho(\mu p_t, \mu' p_t) \leq e^{-ct}\mathcal{W}_\rho(\mu, \mu') \tag{126}$$

$\square$