ADEPT: Hierarchical Bayes Approach to Personalized Federated Unsupervised Learning

Kaan Ozkara
UCLA
kaan@ucla.edu

Bruce Huang
UCLA
brucehuang@ucla.edu

Ruida Zhou
UCLA
ruida@ucla.edu

Suhas Diggavi
UCLA
suhas@ee.ucla.edu

Abstract

Statistical heterogeneity of clients' local data is an important characteristic in federated learning, motivating personalized algorithms tailored to local data statistics. Though there has been a plethora of algorithms proposed for personalized supervised learning, discovering the structure of local data through personalized unsupervised learning is less explored. We initiate a systematic study of such personalized unsupervised learning by developing algorithms based on optimization criteria inspired by a hierarchical Bayesian statistical framework. We develop adaptive algorithms that discover the balance between using limited local data and collaborative information. We do this in the context of two unsupervised learning tasks: personalized dimensionality reduction (ADEPT-PCA and ADEPT-AE) and personalized diffusion models (ADEPT-DGM). We develop convergence analyses for our adaptive algorithms which illustrate the dependence on problem parameters (*e.g.,* heterogeneity, local sample size). We also develop a theoretical framework for personalized diffusion models, which shows the benefits of collaboration even under heterogeneity. We finally evaluate our proposed algorithms using synthetic and real data, demonstrating the effective sample amplification for personalized tasks, induced through collaboration, despite data heterogeneity.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

One of the goals of unsupervised learning is to discover the underlying structure in data and use this for tasks such as dimensionality reduction and generating new samples from the data distribution. We might want to perform this task on the local data of a client, *e.g.,* data collected from personal sensors or other devices; such data could have heterogeneous statistics across clients. The desired *personalized* task should be tailored to particular distributions of the local data, and hence to discover this structure one might need a significant amount of local data. There might be insufficient local samples for the task, motivating collaboration between clients. Moreover, as argued in the federated learning (FL) paradigm, we would like to leverage data across clients without data sharing McMahan et al. [2017], Kairouz et al. [2021]. In this paper, we initiate a systematic study of personalized federated unsupervised learning, where clients collaborate to discover personalized structure in their local data.

There has been a plethora of personalized learning models proposed in the literature, mostly for *supervised* federated learning Fallah et al. [2020], Dinh et al. [2020], Mansour et al. [2020], Ozkara et al. [2021, 2023a]. These methods were motivated by the statistical heterogeneity of local data, causing a single "global" model to perform poorly on local data. The different personalized federated supervised learning algorithms were unified in Ozkara et al. [2023a] using an empirical/hierarchical Bayes statistical model Efron [2010] [1], which also suggested new *supervised* learning algorithms. However, there has been much less work on personalized federated unsupervised learning. We will build on the statistical approach studied in Ozkara et al. [2023a] for supervised learning, applying it to personalized unsupervised learning. This leads to new federated algorithms for personalized dimensionality reduction

[1] Not to be confused with Bayesian inference/deep learning, our statistical model is being used to model the data, and consequently, to develop personalized adaptive optimization solved using first-order methods.

and personalized diffusion-based generative models for heterogeneous data. Compared to the application of the hierarchical Bayes model to a supervised setting, in our work, there are several differences. For generative models the likelihood function is not available, hence, we utilize ELBO to derive the optimization criteria. Moreover, for all settings, we introduce hyper-prior which is necessary for stable training and meaningful theoretical analysis (e.g. for convergence). Empirical performance is not sensitive to hyper-prior (a small stability parameter) as we show in the ablation studies, yet without the hyper-prior, there is a risk of converging to trivial solutions.

A question is how to use local data and learn from others despite heterogeneity. The hierarchical Bayes framework Efron [2010] suggests using an estimated population distribution effectively as a prior. However, the challenge is to make each client efficiently combine local data and collaborative information for the unsupervised learning task. We do this by *simultaneously* learning a global model (a proxy for the population model) and a local model by adaptively estimating the discrepancy between the global and local information. In Section 2, we define this through a loss function, making it amenable to a distributed gradient descent approach. Our main contributions are as follows.

Unsupervised collaboration learning criteria: Section 2 develops and uses the hierarchical Bayes framework for personalized dimensionality reduction and diffusion (generative) models. We develop an *Adaptive Distributed Emperical-Bayes based Personalized Training* (`ADEPT`) criterion which embeds the balance between local data and collaboration for these tasks (see below). As far as we know, these are the first explicit criteria for such personalized federated unsupervised learning tasks.

Personalized dimensionality reduction: Section 3 develops adaptive personalized algorithms for linear (PCA) (`ADEPT-PCA`) and non-linear (auto-encoders) (`ADEPT-AE`) for dimensionality reduction. We also demonstrate its convergence in Theorems 3.3 and 3.6. In Remark 3.7, we see that these allow us to theoretically examine the impact of heterogeneity, number of local samples, and number of clients, on optimization performance. We believe these are the first adaptive algorithms for personalized dimensionality reduction and their convergence analyses. Finally, in Section 5, we evaluate `ADEPT-PCA` and `ADEPT-AE` and show the benefits of adaptive collaboration. For example, Table 2 shows effective amplification of as much as 20× in local sample complexity through collaboration.

Personalized diffusion models: Section 4 develops an adaptive personalized diffusion generative model (`ADEPT-DGM`) to generate novel samples for local data statistics. We believe that `ADEPT-DGM` is the first algorithm for federated personalized diffusion. We develop a theory for such personalized federated diffusion models through our statistical framework and use this to demonstrate, in Theorem 4.4, conditions when collaboration can improve performance, despite statistical heterogeneity. Finally in Section 5, we evaluate `ADEPT-DGM`, demonstrating the value of adaptive collaboration despite heterogeneity, as well as significant performance benefits for "worst" clients.

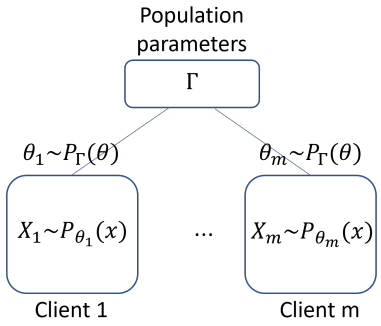

Figure 1: Hierarchical Bayesian model of data distribution

The common goal of dimensionality reduction and generation tasks is due to the aim of modeling the local data generation process $p(x)$. Diffusion models model the generation process to sample new data from the underlying distribution; whereas, in dimensionality reduction, the goal is to discover a low dimensional latent space, for the underlying distribution. We provide related work in Appendix A.

2 Problem Formulation

We present a hierarchical Bayesian framework for personalized unsupervised learning, using it to develop optimization criteria for federated personalized dimensionality reduction and personalized diffusion models. Notation preliminaries can be found in Appendix B.

2.1 Hierarchical Bayes Model for personalized learning

The *statistical model* based on hierarchical Bayes, suitable for federated unsupervised learning, is illustrated in Figure 1. There are m clients. Each client i is associated with a parameter $\boldsymbol{\theta}_i$ and has a local dataset $\boldsymbol{X}_i$ consisting of n data points $(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in})$ i.i.d. from distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$. The parameters obey a population distribution (a.k.a. prior distribution in the Bayesian model) $p(\boldsymbol{\theta}|\Gamma)$ parameterized by Γ. We have a (carefully designed) hyper prior distribution π over Γ, i.e., $\Gamma \sim \pi$ to prevent ill-posedness. This statistical model

defines the joint distribution

$$p_{\Gamma,\{\boldsymbol{\theta}_i\},\{\boldsymbol{X}_i\}}(\Gamma,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m,\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m) = \pi(\Gamma)\prod_{i=1}^{m}p(\boldsymbol{\theta}_i|\Gamma)\prod_{i=1}^{m}\prod_{j=1}^{n}p(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_i). \quad (1)$$

We learn the distribution parameters $\Gamma, \{\boldsymbol{\theta}_i\}$ from data $\{\boldsymbol{X}_i\}$ by maximizing the joint distribution (a.k.a. maximum a posteriori), by minimizing `ADEPT` loss function:

$$f(\{\boldsymbol{\theta}_i\},\Gamma;\{\boldsymbol{X}_i\}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{\theta}_i;\boldsymbol{X}_i) + R(\{\boldsymbol{\theta}_i\},\Gamma),$$

where $f_i(\boldsymbol{\theta}_i;\boldsymbol{X}_i) := -\sum_{j=1}^{n}\log(p(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_i))$ is the local loss function for client i reflecting the likelihood of the local dataset, and $R(\{\boldsymbol{\theta}_i\},\Gamma) = -\frac{1}{m}\log\pi(\Gamma) - \frac{1}{m}\sum_{i=1}^{m}\log p(\boldsymbol{\theta}_i|\Gamma)$ is a regularization allowing the collaboration among clients. When the likelihood function is not easy to optimize, we leverage surrogates such as evidence lower bound (ELBO) instead.

In this work, we focus on a Gaussian population distribution over an d_θ-dimensional normed metric space $(\Theta,\|\cdot\|)$[2]. Specifically, $\Gamma = (\boldsymbol{\mu},\sigma)$, the parameters $\boldsymbol{\theta},\boldsymbol{\mu}\in\Theta$ and $p(\boldsymbol{\theta}|\Gamma) = \frac{1}{(2\pi\sigma^2)^{d_\theta/2}}\exp(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\mu}\|^2}{2\sigma^2})$, where d is the dimensionality of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ has the same dimension as $\boldsymbol{\theta}$. We assume an improper (non-informative) hyper prior π over $\Gamma = (\boldsymbol{\mu},\sigma)$, as $\boldsymbol{\mu}\sim\mathcal{N}(0,\infty\cdot\boldsymbol{I}_{d_\theta})$ and σ^2 follows an inverse gamma distribution parameterized by a hyper-parameter ξ, i.e., $\pi(\boldsymbol{\mu},\sigma)\propto\exp(\frac{m\xi}{\sigma^2})$ [3]. We thus have the regularization

$$R(\{\boldsymbol{\theta}_i\},\Gamma) = -\frac{1}{m}\log\pi(\Gamma) - \frac{1}{m}\sum_{i=1}^{m}\log p(\boldsymbol{\theta}_i|\Gamma) = \frac{1}{m}\sum_{i=1}^{m}\frac{2\xi+\|\boldsymbol{\mu}-\boldsymbol{\theta}_i\|^2}{2\sigma^2} + d_\theta\log\sigma. \quad (2)$$

The Gaussian population distribution is a natural way of capturing relationships between clients' models by modeling the heterogeneity via covariance. Such a population distribution leads to an ℓ_2 regularization, and hence is a natural one to focus on for unsupervised methods. Hence, given the Gaussian population distribution, heterogeneity can be modeled via variance parameters, which enables adaptive optimization problems.

2.2 Personalized Dimensionality Reduction

Linear Dimensionality Reduction: The linear dimensionality reduction is equivalent to PCA. We extend a personalized PCA formulation that was previously studied in Ozkara et al. [2023b] by introducing adaptivity i.e. optimizing the loss over σ. In this setting, the dataset from client i is $\boldsymbol{X}_i\in\mathbb{R}^{d\times n}$, which contains n samples of d dimensional vectors. Let us denote $\boldsymbol{S}_i = \frac{1}{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top$ as the sample covariance matrix of client i. For notational consistency with canonical PCA notations, we set the parameters $\boldsymbol{\theta}_i = \boldsymbol{U}_i\in\mathbb{R}^{d\times r}$. Similar to Ozkara et al. [2023b], we adopt the probabilistic view of PCA Tipping and Bishop [1999],

$$\boldsymbol{x}_{ij} = \boldsymbol{U}_i\boldsymbol{z}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad (3)$$

where $\boldsymbol{z}_{i1},\ldots,\boldsymbol{z}_{in}\overset{i.i.d.}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_r)$ and $\boldsymbol{\epsilon}_{i1},\ldots,\boldsymbol{\epsilon}_{in}\overset{i.i.d.}{\sim}\mathcal{N}(\boldsymbol{0},\sigma_\epsilon^2\boldsymbol{I}_d)$. This results in the likelihood function $p(\boldsymbol{X}_i|\boldsymbol{\theta}=\boldsymbol{U}_i) = \mathcal{N}(\boldsymbol{0},\boldsymbol{U}_i\boldsymbol{U}_i^\top+\sigma_\epsilon^2\boldsymbol{I})$. Recall that the prior parameter is $\Gamma=(\boldsymbol{\mu},\sigma)$, here for notational consistency we use $\boldsymbol{V}=\boldsymbol{\mu}$. The underlying metric space $(\Theta,\|\cdot\|)$ containing the parameters $\boldsymbol{U}$ and $\boldsymbol{V}$ in the Steifel manifold where $St(d,r) := \{\boldsymbol{U}\in\mathbb{R}^{d\times r}|\boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}\}$. The metric is $d(\boldsymbol{V},\boldsymbol{U}) = \|P_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U})\|$ where $P_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}) = \boldsymbol{U} - \frac{1}{2}\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{U}+\boldsymbol{U}^\top\boldsymbol{V})$ is the projection of $\boldsymbol{U}$ onto the tangent space at $\boldsymbol{V}$. Because computing the (geodesic) distance on $St(d,r)$ is hard, the defined metric first projects a matrix to the tangent space of another matrix and then computes the Frobenius norm. The personalized unsupervised learning is then minimizing the `ADEPT-PCA` loss function, f^{pca} over $\{\boldsymbol{U}_i\},\boldsymbol{V},\sigma$:

$$\frac{1}{m}\Big(\sum_{i=1}^{m}\frac{n}{2}(\log(|\boldsymbol{W}_i|)+\text{tr}(\boldsymbol{W}_i^{-1}\boldsymbol{S}_i))+\frac{2\xi+d^2(\boldsymbol{V},\boldsymbol{U}_i)}{2\sigma^2}\Big) + d_\theta\log\sigma \quad (4)$$

$$\text{s.t.}\boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I},\boldsymbol{U}_i^\top\boldsymbol{U}_i=\boldsymbol{I}\forall i;\boldsymbol{W}_i=(\boldsymbol{U}_i\boldsymbol{U}_i^\top+\sigma_\epsilon^2\boldsymbol{I}).$$

Non-linear dimensionality reduction: While PCA is a good starting point for dimensionality reduction, it cannot capture non-linear relations between the latent variable and the observed space. Hence, one can extend (3) to model non-linearity as follows,

$$X_{ij} = \psi_{\boldsymbol{\theta}_i^{\text{d}}}(\boldsymbol{z}_{ij}) + \boldsymbol{\epsilon}_{ij}, \quad (5)$$

where $\boldsymbol{\theta}_i^{\text{d}}$ parameterizes a non-linear decoding map. We can parameterize the encoder structure such that $\boldsymbol{z}_{ij} = g_{\boldsymbol{\theta}_i^{\text{e}}}(X_{ij})$. Using Gaussian distribution we get the `ADEPT-AE` criterion:

$$\underset{\{\boldsymbol{\theta}_i\},\boldsymbol{\mu},\sigma}{\arg\min} f^{\text{ae}}(\{\boldsymbol{\theta}_i\},\boldsymbol{\mu},\sigma) = \frac{1}{m}\sum_{i=1}^{m}\Big(\|\boldsymbol{X}_i - \psi_{\boldsymbol{\theta}_i^{\text{d}}}(g_{\boldsymbol{\theta}_i^{\text{e}}}(\boldsymbol{X}_i))\|_F^2 + \frac{2\xi+\|\boldsymbol{\mu}-\boldsymbol{\theta}_i\|^2}{2\sigma^2}\Big) + d_\theta\log\sigma \quad (6)$$

where $\boldsymbol{\mu}$ is the global model and $\{\boldsymbol{\theta}_i\}$ are the personalized AEs through concatenation of $\boldsymbol{\theta}_i^{\text{e}}$, $\boldsymbol{\theta}_i^{\text{d}}$.

[2] Note that this general framework can be applied to any general (parametric) population distribution.

[3] Inverse-Gamma distribution is conjugate prior distribution for the variance term.

2.3 Personalized Generation through Diffusion Models

The denoising diffusion model has attracted attention recently due to its capability of generating high-quality images and the theoretical foundation of the stochastic differential equations Rezende and Mohamed [2015], Sohl-Dickstein et al. [2015]. The mathematical model of a diffusion model is the following stochastic differential equation Song et al. [2021]

$$d\boldsymbol{x}_t = d\boldsymbol{w}_t \quad \text{(variance exploding process)}, \tag{7}$$

where $\boldsymbol{w}_t$ is standard d-dimensional Brownian motion. For time $t\in[0,T]$, its time-reversed process is

$$d\boldsymbol{x}_t^{\leftarrow} = \nabla \ln p_{\boldsymbol{x}_{T-t}}(\boldsymbol{x}_t^{\leftarrow})dt + d\boldsymbol{w}_t^{\leftarrow}, \tag{8}$$

where $p_{\boldsymbol{x}_t}$ is the probability density function of $\boldsymbol{x}_t$, $\boldsymbol{w}_t^{\leftarrow}$ is a standard Wiener process. It can be shown that $X_t^{\leftarrow}$ follows the same distribution as $\boldsymbol{x}_{T-t}$. More strongly, suppose $X_0^{\leftarrow}$ has the same distribution as $\boldsymbol{x}_T$, and then the two processes $\{\boldsymbol{x}_{T-t}\}_{t\in[0,T]}$, $\{\boldsymbol{x}_t^{\leftarrow}\}_{t\in[0,T]}$ have the same distribution.

Suppose that the score function $\nabla \ln p_{\boldsymbol{x}_t}(\boldsymbol{x})$ can be represented by a neural network $\boldsymbol{\phi}(\boldsymbol{x};\boldsymbol{\theta},t)\in\mathbb{R}^d$ for some parameter $\boldsymbol{\theta}$. The data generation is then integration over the time-revised process (8) with the drift function $\boldsymbol{\phi}(\boldsymbol{x};\boldsymbol{\theta},t)$ and the starting distribution $\boldsymbol{x}_0^{\leftarrow}\sim\mathcal{N}(\boldsymbol{0},\sigma_0^2\boldsymbol{I}_d)$. The generated distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ is then implicitly defined without a closed form. [Kingma and Gao, 2023, Eq (3)] shows that

$$\begin{aligned}&\ln p(\boldsymbol{x}|\boldsymbol{\theta}) \geq ELBO_{\boldsymbol{\theta}}(\boldsymbol{x}) \\ &= -\frac{1}{2}\int_{t=0}^{T} \mathbb{E}\|\boldsymbol{\phi}(\boldsymbol{x}_t;\boldsymbol{\theta}_j,t) - \nabla_{\boldsymbol{x}_t}\ln p_{\boldsymbol{x}_t|\boldsymbol{x}_0}(\boldsymbol{x}_t|\boldsymbol{x})\|^2 dt + C\end{aligned} \tag{9}$$

where $\boldsymbol{x}_t\sim\mathcal{N}(\boldsymbol{x},\boldsymbol{I}_d)$ and C is some constant factor independent of $\boldsymbol{\theta}$. Accordingly, we use ELBO as a surrogate function for the negative log-likelihood; thus, the personalized loss function is $f_i(\boldsymbol{\theta}_i;\boldsymbol{X}_i) = -ELBO_{\boldsymbol{\theta}_i}(\boldsymbol{X}_i) = -\sum_{j=1}^n ELBO_{\boldsymbol{\theta}_i}(\boldsymbol{x}_{ij})$. The personalized data generation is then minimizing the `ADEPT-DGM` loss function

$$\begin{aligned}&\min_{\{\boldsymbol{\theta}_i\},\boldsymbol{\mu},\sigma^2} f^{\mathrm{df}}(\{\boldsymbol{\theta}_i\},\boldsymbol{\mu},\sigma;\{\boldsymbol{X}_i\}) \\ &:= \frac{1}{m}\sum_{i=1}^m\left(-ELBO_{\boldsymbol{\theta}_i}(\boldsymbol{X}_i) + \frac{2\xi+\|\boldsymbol{\mu}-\boldsymbol{\theta}_i\|^2}{2\sigma^2}\right) + d_\theta\log\sigma.\end{aligned} \tag{10}$$

3 Personalized Federated Dimensionality Reduction

We present our algorithms and convergence results on personalized dimensionality reduction.

3.1 Personalized Adaptive PCA: `ADEPT-PCA`

We introduce `ADEPT-PCA` (Algorithm 2 deferred to Appendix C) to train adaptive personalized PCA. To show the convergence of the algorithm we need the following standard assumption and naturally occurring lower bound on σ.

Assumption 3.1. For each client i, the operator and Frobenius norms of $\boldsymbol{S}_i$ are bounded by

$$\|\boldsymbol{S}_i\|_F \leq G_{i,F} \quad \text{and} \quad \|\boldsymbol{S}_i\|_{op} \leq G_{i,op},$$

and $G_{max,F} := \max_{i\in[m]} G_{i,F}$,$G_{max,op} := \max_{i\in[m]} G_{i,op}$. The assumption implies the Lipschitz smoothness properties of the loss function w.r.t. personalized PCs $\{\boldsymbol{U}_i\}$.

Lemma 3.2 (A lower bound on σ_t)**.** *Given any $\omega\in(0,1)$. Let the learning rate $\eta_3\leq(1-\omega)\frac{2\xi}{d_\theta^2}$ and the initialization $\sigma_0\geq\omega\sqrt{\frac{2\xi}{d_\theta}}$. Then, for all $t\in[T]$, we have $\sigma_t\geq\omega\sqrt{\frac{2\xi}{d_\theta}}$.*

See Appendix C.3.1 for the proof. We will fix some $\omega\in(0,1)$ for the rest of the paper and initialize σ_0 accordingly so that we can utilize the lower bound in Lemma 3.2. The bound is due to ξ and is necessary to guarantee that the loss does not explode due to vanishing σ. Let us now define $\boldsymbol{g}_{i,t}^U = P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}}(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1},\boldsymbol{V}_{t-1},\sigma_{t-1}))$, $\boldsymbol{g}_t^V = \frac{1}{m}\sum_{i=1}^m P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}}(\nabla_{\boldsymbol{V}_{t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t},\boldsymbol{V}_{t-1},\sigma_{t-1}))$, and $g_t^\sigma = \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})$. Then, Algorithm 2 has the following convergence bound for finding a first-order stationary point.

Theorem 3.3 (Convergence of `ADEPT-PCA` Algorithm 2)**.** *Let $G_t = \left(\frac{1}{m}\sum_{i=1}^m\left\|\boldsymbol{g}_{i,t}^U\right\|^2\right) + \left\|\boldsymbol{g}_t^V\right\|^2 + (g_t^\sigma)^2$. By choosing $\eta_1 = \min\{\frac{1}{3C_{\eta_1}},1\}$, $\eta_2 = \min\{\frac{1}{3C_{\eta_2}},1\}$, and $\eta_3 = \min\left\{\frac{\eta_1}{3(L_U^{(\sigma)})^2},\frac{\eta_2}{3(L_V^{(\sigma)})^2},\frac{1}{L_\sigma}\right\}$, we have*

$$\frac{1}{T}\sum_{t=1}^T G_t \leq \frac{3\Delta_T^{\mathrm{pca}}}{T\min\{\eta_1,\eta_2,\eta_3\}},$$

where T is number of total iterations, $\Delta_T^{\mathrm{pca}} = f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,0}\}_i,\boldsymbol{V}_0,\sigma_0) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,T}\}_i,\boldsymbol{V}_T,\sigma_T)$, and η_1,η_2,η_3 are the learning rates for updating $\boldsymbol{U}_i$, $\boldsymbol{V}$, and σ respectively. The constants are defined such that $f^{pca}(\cdot)$ isL_σ-smooth w.r.t. σ and $\boldsymbol{g}_{i,t}^U,\boldsymbol{g}_t^V$ are $L_U^{(\sigma)},L_V^{(\sigma)}$ continuous w.r.t. σ respectively. C_{η_1},C_{η_2} are defined in Appendix C and depend on smoothness w.r.t. $\boldsymbol{U},\boldsymbol{V}$.

We provide a detailed comment on the factors impacting the convergence rate in Remark 3.7.

Remark 3.4 (The Polar Retraction). In Algorithm 2, we use the polar retraction to map the updated $\boldsymbol{V}$ and $\boldsymbol{U}_i$ back to the Steifel manifold. We define the polar retraction in Appendix C in detail.

Proof outline of Theorem 3.3. We show that a lower bound on σ_t is obtained with proper initialization of

σ_0 and we can further derive the Lipschitz constants of the loss function (4). The sufficient decrease of the loss function w.r.t. $\boldsymbol{U}_i$, $\boldsymbol{V}$, and σ (Lemma C.8) are derived individually using non-expansiveness of polar retraction (Lemma C.1) and Lipschitz type inequality (Lemma C.2). After the sufficient decrease, we show that carefully choosing the learning rates results in an upper bound on the squared norm of gradients. The detailed proof is in Appendix C. Technical challenges in this proof are to control the error due to projections, utilize the lower bound on σ to avoid non-smoothness, and use the Lipschitz continuity of the gradients to combine updates on PCs with the update on σ.

3.2 Personalized Adaptive AEs: ADEPT-AE

Algorithm 1 ADEPT-AE Algorithm

Input: Number of iterations T, learning rates (η_1, η_2, η_3), number of local iterations τ

1: **Init** local models $\{\boldsymbol{\theta}_{i,0}\}_{i=1}^m$, global model $\boldsymbol{\mu}_0$, and σ_0.
2: **On server:**
3: Broadcast $\boldsymbol{\mu}_0$, σ_0 to all clients
4: **for** $t = 1$ **to** T **do**
5: **On Clients:**
6: **for** $i = 1$ **to** m **do**
7: **if** τ divides $t-1$ **then**
8: Receive $\boldsymbol{\mu}_{t-1}$, σ_{t-1}
9: **end if**
10: $\boldsymbol{\theta}_{i,t}=\boldsymbol{\theta}_{i,t-1}-\eta_1\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
11: **if** τ divides t **then**
12: $\boldsymbol{\mu}_{i,t} = \boldsymbol{\mu}_{t-1} - \eta_2\nabla_{\boldsymbol{\mu}_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
13: $\sigma_{i,t} = \sigma_{t-1} - \eta_3\nabla_{\sigma_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
14: Send $\boldsymbol{\mu}_{i,t}, \sigma_{i,t}$ to server
15: **else**
16: $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1}$, $\sigma_t = \sigma_{t-1}$
17: **end if**
18: **end for**
19: **At the Server:**
20: **if** τ divides t **then**
21: Receive $\{\boldsymbol{\mu}_{i,t}\}_{i=1}^m$ and $\{\sigma_{i,t}\}_{i=1}^m$
22: $\boldsymbol{\mu}_t = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\mu}_{i,t}$, $\sigma_t = \frac{1}{m}\sum_{i=1}^m \sigma_{i,t}$
23: Broadcast $\boldsymbol{\mu}_t$, σ_t to all clients
24: **end if**
25: **end for**

Output: Personalized autoencoders $\{\boldsymbol{\theta}_{1,T}, \ldots, \boldsymbol{\theta}_{m,T}\}$.

Algorithm 1 shows the alternating gradient descent training procedure for personalized adaptive AEs. At the beginning of local iterations (**line 8**) the clients receive the global model and σ[4] terms then do local updates on $\boldsymbol{\theta}_i$ (**line 10**). At the end of local iterations, the client updates the global model and variance term using its personalized model **lines 12,13** and sends them to the server where it is aggregated.

Assumption 3.5. The loss function $f_i^{(ae)}(\{\boldsymbol{\theta}_i\}, \boldsymbol{\mu}, \sigma)$ is $L_{\boldsymbol{\theta}}$-smooth w.r.t. individual $\{\boldsymbol{\theta}_i\}$, $L_{\boldsymbol{\mu}}$-smooth w.r.t. $\boldsymbol{\mu}$ and L_σ-smooth w.r.t. σ. Note that only the first one is an assumption, second and third ones are derived from the fact that σ is lower bounded when initialized properly (Appendix D).

Let $\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$, $\boldsymbol{g}_t^{\boldsymbol{\mu}} = \nabla_{\boldsymbol{\mu}_{t-1}} f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}_i, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}), g_t^\sigma = \frac{\partial f(\{\boldsymbol{\theta}_{i,t}\}_i, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})}{\partial \sigma_{t-1}}$. For Algorithm 1, we obtain the following convergence bound for finding a first-order stationary point.

Theorem 3.6 (Convergence of ADEPT-AE (Algorithm 1))**.** *Let us define* $G_t = \left(\frac{1}{m}\sum_{i=1}^m \|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\|^2\right) + \|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + (g_t^\sigma)^2$. *Then, by choosing* $\eta_1 = \frac{1}{L_{\boldsymbol{\theta}}}$, $\eta_2 = \frac{1}{L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}}$, *and* $\eta_3 = \min\{1, \frac{1}{L_{\boldsymbol{\mu}}}\}$, *we have* $\min_{t\in[T], \tau|t}\{G_t\}$ *is upper bounded by*

$$\frac{\max\{L_{\boldsymbol{\theta}}, L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}, L_{\boldsymbol{\mu}}, 1\}\Delta_T^{\text{ae}}}{R},$$

where $R = T/\tau$ *is the number of communication rounds,* $\Delta_T^{\text{ae}} = f^{\text{ae}}(\{\boldsymbol{\theta}_{i,0}\}_i, \boldsymbol{\mu}_0, \sigma_0) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,T}\}_i, \boldsymbol{\mu}_T, \sigma_T)$, *and the constants can be found in Lemma D.1.*

Remark 3.7. By examining the multiplicative constants within the bounds specified in Theorems 3.3 and 3.6, we note a consistent observation: σ exhibits an inverse relationship with convergence speed. A higher σ, whether resulting from a large value of ξ or inherent heterogeneity in the setting, can expedite the convergence process. Essentially, a large σ diminishes collaboration, allowing the model to fit quickly due to a reduced effective number of samples. Conversely, a smaller σ promotes collaboration and may augment the effective sample count. This theoretical observation persists in the experiments as well, as can be seen from the convergence plots in Appendix F. Note that faster convergence does not necessarily imply a superior generalization error. Consequently, in our experiments, opting for a high value of σ_0 facilitates fast convergence in the initial stages, while still allowing flexibility for adjustments to yield a superior generalization error.

Proof outline of Theorem 3.6. Similar to our proof for Theorem 3.3, we utilize the lower bound on σ_t and look for the sufficient decrease w.r.t. $\boldsymbol{\theta}_i$, $\boldsymbol{\mu}$, and σ. However, now we have to deal with the fact that we are doing τ local iterations after each communication round. Thus, we consider the two cases of whether it is a communication round or not separately and come to our Theorem in the end. The proof is in Appendix D.

[4] In the experiments we use individual variance terms for each weight that is $\boldsymbol{\sigma} \in \Theta$ which only has a constant effect on convergence (see Appendix D).

4 Personalized Generation through Adaptive Diffusion Models: ADEPT-DGM

The ADEPT-DGM algorithm is given in Algorithm 3 (deferred to Appendix E). It is based on optimizing the criterion given in (10). The adaptation method to discover the balance between local data and collaborative information is similar to that in ADEPT-AE Algorithm 1.

As an illustration of the effectiveness of personalized diffusion model learning, we analyze a simple personalized Gaussian distribution generation problem, i.e., client-i's target distribution is a Gaussian distribution $p(\boldsymbol{x}|\boldsymbol{\theta}_i) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\theta}_i, \sigma_0^2 \boldsymbol{I}_d)$. In the following, we first introduce some details of the canonical denoising diffusion model for Gaussian target distribution and analyze the improvement by collaboration in the proposed personalized algorithm.

Diffusion model with Gaussian target distribution: When the desired distribution is $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\theta}, \sigma_0^2 I_d)$, the diffusion process (7) is a Gaussian process and the drift term for the reverse-time process (8) is a linear function, i.e., $\nabla \ln p_{\boldsymbol{x}_{T-t}}(\boldsymbol{x}_t^{\leftarrow}) = -\frac{\boldsymbol{x}_t^{\leftarrow} - \boldsymbol{\theta}}{\sigma_0^2 + t}$. The time-reversed process is

$$d\boldsymbol{x}_t^{\leftarrow} = \nabla \ln p_{\boldsymbol{x}_{T-t}}(\boldsymbol{x}_t^{\leftarrow}) dt + d\boldsymbol{w}_t^{\leftarrow}, \boldsymbol{x}_0^{\leftarrow} \sim \mathcal{N}(\boldsymbol{\theta}, (\sigma_0^2 + T)\boldsymbol{I}_d). \tag{11}$$

Without the knowledge of $\boldsymbol{\theta}$, we approximate the score function by a neural network of $\boldsymbol{\phi}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}, t) = -\frac{\boldsymbol{x} - \hat{\boldsymbol{\theta}}}{\sigma_0^2 + t}$ and approximate the initial distribution of the time-reversed process $\mathcal{N}(\boldsymbol{\theta}, (\sigma_0^2 + T)\boldsymbol{I}_d)$ by $\mathcal{N}(\boldsymbol{0}, (\sigma_0^2 + T)\boldsymbol{I}_d)$. The learned/approximated time-reversed process is then

$$d\boldsymbol{x}_t^{\leftarrow} = -\frac{\boldsymbol{x} - \hat{\boldsymbol{\theta}}}{\sigma_0^2 + t} dt + d\boldsymbol{w}_t^{\leftarrow}, \boldsymbol{x}_0^{\leftarrow} \sim \mathcal{N}(\boldsymbol{0}, (\sigma_0^2 + T)\boldsymbol{I}_d). \tag{12}$$

The following lemma characterizes the difference between the generation and target distributions.

Lemma 4.1. *The output distribution $p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}}$ of the learned reversed-time process* (12) *satisfies,*

$$D_{KL}(p_{\boldsymbol{x}|\boldsymbol{\theta}} || p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}}) = \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} + \frac{\sigma_0^2}{\sigma_0^2 + T} \hat{\boldsymbol{\theta}} \right\|^2, \tag{13}$$

where D_{KL} is the Kullback-Leibler divergence and $p_{\boldsymbol{x}|\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\theta}_i, \sigma_0^2 \boldsymbol{I}_d)$ is the target distribution.

The lemma shows that the KL-divergence between target distribution and the learned distribution is measured by the difference between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ a bias term $\frac{\sigma_0^2}{\sigma_0^2 + T} \hat{\boldsymbol{\theta}}$ due to the initial distribution mismatch of the time-reversed process. It is straightforward to verify that for $\boldsymbol{\phi}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}, t) = -\frac{\boldsymbol{x} - \hat{\boldsymbol{\theta}}}{\sigma_0^2 + t}$, training with dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ by maximizing ELBO (9) has a closed-form sample-mean solution, i.e., $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} ELBO_\theta(\boldsymbol{X}) = \frac{\sum_{j=1}^n \boldsymbol{x}_j}{n}$. Since the data are i.i.d. sampled from $\mathcal{N}(\boldsymbol{\theta}, \sigma_0^2 \boldsymbol{I}_d)$, we have $\mathbb{E}[D_{KL}(p_{\boldsymbol{x}|\boldsymbol{\theta}} || p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}})] = \frac{\sigma_0^2}{n}$, when omitting the initial distribution bias by $T \to \infty$. It can be viewed as personalized training without collaboration and in the following, we show that the proposed personalized training with collaboration can improve over this.

Personalized denoising diffusion model with Gaussian targets: The local dataset $\boldsymbol{X}_i$ of client-i sampled i.i.d. from a target Gaussian distribution $P(\boldsymbol{x}|\boldsymbol{\theta}_i) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\theta}_i, \sigma_0^2 \boldsymbol{I}_d)$, and the population distribution is also Gaussian with $P(\boldsymbol{\theta}|\Gamma_*) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_*, \sigma_*^2 \boldsymbol{I}_d)$. Note that we analyze a fixed unknown population distribution parameterized by $\Gamma_* = (\boldsymbol{\mu}_*, \sigma_*)$, though the proposed loss function and algorithm follows a Hierarchical Bayesian model as in Section 2.3.

Lemma 4.2 (Personalized estimation)**.** *For the parameterized score function $\boldsymbol{\phi}(\boldsymbol{x}; \boldsymbol{\theta}, t) = -\frac{\boldsymbol{x} - \boldsymbol{\theta}}{\sigma_0^2 + t}$, the optimal solution to (10) is*

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^m \sum_{j=1}^n \boldsymbol{x}_{ij}}{mn}, \hat{\boldsymbol{\theta}}_i = \frac{n\alpha\hat{\sigma}^2}{n\alpha\hat{\sigma}^2 + 1} \Big(\frac{\sum_{j=1}^n \boldsymbol{x}_{ij}}{n} + \hat{\boldsymbol{\mu}} \Big),$$

where $\alpha = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_0^2 + T}$ and $\hat{\sigma}^2$ satisfies $\hat{\sigma}^2 = \frac{2\xi}{d} + s^2 (\frac{n\alpha\hat{\sigma}^2}{n\alpha\hat{\sigma}^2 + 1})^2$ with $s^2 = \frac{\sum_{i=1}^m \left\| \hat{\boldsymbol{\mu}} - \frac{1}{n} \sum_{j=1}^n \boldsymbol{x}_{ij} \right\|^2}{md}$.

Remark 4.3. We note that the global optimum model is the average of average samples of each client. The personalized optimum model is an interpolation of the local estimate and the true global model. The interpolation coefficient depends on the heterogeneity (large σ skews the result towards the local estimate and low σ skews it towards the true global model). A large amount of local samples n decreases the reliance on $\boldsymbol{\mu}$ These observations are in parallel with findings in Ozkara et al. [2023a] on mean estimation.

Theorem 4.4 (Condition for performance improvement)**.** *Consider an asymptotic regime that $m \to \infty$ and $T \to \infty$. Compared to training without collaboration, the solution of $(\{\hat{\boldsymbol{\theta}}_i\}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2)$ in Lemma 4.2 improves the averaged KL-divergence $\frac{1}{m} \sum_{i=1}^m D_{KL}(p_{\boldsymbol{x}|\boldsymbol{\theta}_i} || p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}_i})$ by a factor of $\left(\frac{2\hat{\sigma}^2 + \sigma_0^2/n - \sigma_*^2}{\hat{\sigma}^2 + \sigma_0^2/n} \right) \frac{\sigma_0^2/n}{\hat{\sigma}^2 + \sigma_0^2/n} \frac{\sigma_0^2}{n}$, when $\hat{\sigma}^2 > \frac{\sigma_*^2}{2} - \frac{\sigma_0^2}{2n}$.*

Corollary 4.5. *Under the same setting as in Theorem 4.4, choosing $\xi \geq \frac{3d\sigma_0^2}{2n}$ guarantees improvement of collaboration for any population distribution $\mathcal{N}(\boldsymbol{\mu}_*, \sigma_* \boldsymbol{I}_d)$.*

Remark 4.6. Note that if our estimate $\hat{\sigma}^2$ of the population variance is accurate, *i.e.,* $\hat{\sigma}^2 = \sigma_*^2$, then collaboration always improves over only using local data for a personalized generation; and the collaboration gain is the largest. In this case the gain is larger when the number of local samples is small. However, if the estimate

is inaccurate, one could be better off without collaboration. However, by setting the hyperparameter $\xi \geq \frac{3d\sigma_0^2}{2n}$, we can ensure that our estimate $\hat{\sigma}^2 \geq \frac{1}{2}\left(\sigma_*^2 - \sigma_0^2/n\right)$, ensuring that collaboration is useful.

5 Experiments

5.1 Experimental Setting

In our experiments, we compare our adaptive personalized unsupervised algorithms with global training (FedAvg, FedAvg+fine-tuning), local training (training individual models without collaboration), and competitive baselines in terms of testing performance under different heterogeneous scenarios. While the supervised personalized FL methods that are based on architectural changes, such as shared representations Collins et al. [2021], are not applicable to unsupervised personalized FL due to symmetric model architecture; optimization-based methods can be applied to unsupervised settings, without providing theoretical insights that our framework provides. From such methods, we compare to pFedMe Dinh et al. [2020] applied them in the AE problem. Experiments for `ADEPT-PCA` are in Appendix F.1. Additional experiments, and details hyper-parameters can be found in Appendix F.

ADEPT-AE: We do synthetic and real data experiments for AEs. We use 50 clients ($m = 50$) for all AE experiments. In the synthetic experiments, from a zero mean $\sigma_\mu = 0.1$ standard deviation Gaussian distribution, we sample weights for a single layer decoder $\mu^{d,*}$ with 5 latent dimensionality and 64 output dimensionality. Then by perturbing the weights with another zero mean Gaussian with σ^* we obtain the true personalized decoders, which are used as in (5) to generate 10 local samples across 50 clients. Heterogeneity among clients will depend on σ^* and is quantified in terms of signal-to-noise ratio as $20\log_{10}\frac{\sigma_\mu}{\sigma^*}$ dB. For real data experiments, we use MNIST, Fashion MNIST (F. MNIST), and CIFAR-10. To introduce heterogeneity, we distribute the samples such that each client has access to samples from a single class which simulates distinct data distributions for each client (referred to as pathological heterogeneity McMahan et al. [2017]). For MNIST and F. MNIST, each client has access to 120 training samples; and for CIFAR-10, 250 samples. In the experiments, we update σ in the first iteration instead of the last one and update the global model locally in each local iteration. Details of the models are in Appendix F.2.

ADEPT-DGM: In the experiments, we let the number of clients be $m = 50, 30, 6$ respectively for MNIST, Fashion MNIST (F. MNIST), CIFAR-10 due to the increased size of the models. For MNIST, we use a 6-layer U-Net from Hugging Face. For F. MNIST and CIFAR-10 we use 1.7M and 9.4M parameter models respectively. The model with size 1.7M has 3 downsampling and upsampling blocks, and the model with size 9.4M has 4 of them; each of the blocks consists of multiple attention and residual layers. More details are shown in the Appendix F. For the MNIST experiments, every client has access to 600 or 1200 samples from one class depending on the experiment; and for CIFAR-10 each client has 2000 samples sampled from 2 classes. We compare to FedAvg+fine-tuning since FedAvg cannot exclusively generate samples from the client's target distribution. We do the same modifications to Algorithm 3 as we did for `ADEPT-AE`. The results for F. MNIST, and more results with different numbers of samples or heterogeneity levels are deferred to Appendix F.2.

Table 1: Total energy captured % averaged over samples and clients in the synthetic experiments, where heterogeneity(Het.) is the std of noise and SNR.

Het.	Method	Energy captured %
0.05 (6dB)	Baseline	88.3 ± 0.5
	`ADEPT-AE`	$\mathbf{87.3} \pm 1.1$
	Global Training	81.3 ± 0.1
	Local Training	83.2 ± 1.9
0.025 (12dB)	Baseline	95.4 ± 0.8
	`ADEPT-AE`	$\mathbf{95.9} \pm 0.1$
	Global Training	94.3 ± 0.2
	Local Training	83.2 ± 1.9
0.01 (20dB)	Baseline	98.6 ± 0.1
	`ADEPT-AE`	$\mathbf{98.7} \pm 0.1$
	Global Training	98.4 ± 0.4
	Local Training	83.2 ± 1.9

5.2 Results

ADEPT-AE: The results are in terms of the percentage of total energy captured per sample, which is $100(1 - \|x - \hat{x}\|^2/\|x\|^2)$ for each sample x. The results are averaged over 3 runs and reported together with the standard deviation. Results on synthetic data are shown in Table 1, the baseline denotes that each client trains a personalized AE whose decoder part is the true data generating decoder. Our method outperforms local and global training and even the competitive baseline when heterogeneity is smaller. The result is similar to Figure 3, showing our method outperforms both local and global training in the regimes in which they are strong alternatives. For results on real datasets, the competitive baseline in Table 2 is evaluated when 10 clients maintain all the training data from their corresponding classes ($n = 5000$ per client). Remarkably, our method ($n = 250$ per client) matches the baseline

Table 2: Total energy captured % averaged over samples and clients in the real dataset experiments.

Dataset	Method	Latent dimensionality	
		Low	*High*
MNIST	Baseline	78.9 ± 0.5	85.7 ± 0.4
	`ADEPT-AE`	$\mathbf{70.8} \pm 0.5$	$\mathbf{77.7} \pm 0.1$
	FedAvg	66.2 ± 0.9	75.9 ± 0.7
	Local Training	67.0 ± 0.8	69.1 ± 1.1
	pFedMe	70.5 ± 0.5	76.5 ± 0.8
F. MNIST	Baseline	88.5 ± 0.2	91.3 ± 0.1
	`ADEPT-AE`	$\mathbf{83.9} \pm 0.2$	$\mathbf{85.6} \pm 0.2$
	FedAvg	81.2 ± 0.2	84.9 ± 0.2
	Local Training	76.9 ± 2.0	77.1 ± 0.5
	pFedMe	83.5 ± 0.6	85.2 ± 0.9
CIFAR-10	Baseline	88.7 ± 0.5	93.3 ± 0.1
	`ADEPT-AE`	$\mathbf{88.4} \pm 0.5$	$\mathbf{93.3} \pm 0.2$
	FedAvg	87.4 ± 0.2	91.2 ± 0.1
	Local Training	87.7 ± 0.2	92.2 ± 0.2
	pFedMe	87.7 ± 0.3	92.4 ± 0.3

on CIFAR-10 (for high latent dimensions), indicating that adaptive personalized collaboration results in 20× effective sample size. For other datasets, our method consistently outperforms FedAvg and local training by an important margin regardless of latent dimensionality. Our method reduces reconstruction error by as much as $\sim 35\%$ and $\sim 25\%$ compared to local training and FedAvg respectively. In MNIST and F. MNIST experiments, `ADEPT-AE` also outperforms all the other methods except the strong baseline we implemented.

Comparison to fine-tuning of regularization parameter: Many of the previous works propose personalization through regularization as outlined in Appendix A. Here we compared to pFedMe as one instance, and showed that our method with adaptivity can outperform pFedMe with offline fine-tuning. **`ADEPT-DGM`:** For diffusion models, we use KID Bińkowski et al. [2018] metric to quantitatively measure the quality of generated dataset (see Table 3), instead of the commonly used FID Heusel et al. [2017]; since FID is biased, it tends be unreliable when the generated datasets are small Bińkowski et al. [2018] as in a personalized federated setting. At the end of the training, each client generates 200 samples using the model with the lowest validation loss and compares it to the local test dataset to compute the metrics. Our method consistently results in generated samples with better quality and improves upon other methods by $5\% - 22\%$. Moreover, in Figure 2 we depict the resulting KID values of the clients. We see that our method brings equity, that is, the worst-performing client is much better compared to the worst clients of other methods,

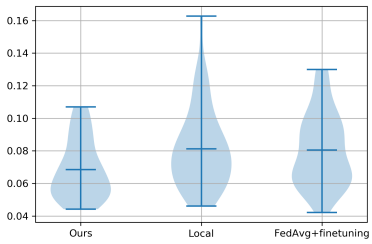

Figure 2: Violin plot of KID values of clients.

and the performance variance of the clients is lower. We also illustrate randomly chosen sample images in Figure 6 in Appendix F.3 and estimate the amount of noise using Immerkær [1996]. Compared to `ADEPT-DGM`, we observe missing features and inconsistent hallucinations in images obtained using FedAvg+fine-tuning. On the other hand, local training outputs images with significantly more background noise in images (e.g. 1st from the last row and 2nd from the first row), and there is a 1.5× increase in noise level in terms of noise standard deviation ($\sigma = 0.032$ for local training vs $\sigma = 0.024$ for adaptive personalized method).

Table 3: Diffusion model generation quality for generating MNIST samples using U-Net model.

Method	MNIST	CIFAR-10
Baseline	0.062 ± 0.003	–
`ADEPT-DGM`	$\mathbf{0.069} \pm 0.002$	$\mathbf{0.058} \pm 0.003$
FedAvg+fine-tuning	0.090 ± 0.009	0.071 ± 0.006
Local Training	0.083 ± 0.003	0.064 ± 0.004

Additional results in Appendix F: In the Appendix, we provide convergence plots under different heterogeneity for `ADEPT-AE` and `ADEPT-PCA` that validate our theoretical findings. Moreover, we provide additional results for `ADEPT-DGM` in different settings alongside with generated images. We also do ablation studies for sensitivity to hyper-prior ξ, which is low.

6 Conclusion

We developed, `ADEPT`, a hierarchical Bayes framework for personalized federated unsupervised learning; leading to new criteria for linear (`ADEPT-PCA`), non-linear (`ADEPT-AE`) dimensionality reduction, and personalized federated diffusion models (`ADEPT-DGM`). Each of our algorithms included adaptation for the heterogeneity during training which resulted in novel theoretical interpretations and superior empirical performance. Open questions include extensions with information constraints such as communication and privacy.

References

Durmus Alp Emre Acar, Yue Zhao, Ruizhao Zhu, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Debiasing model updates for improving personalized federated training. In *International Conference on Machine Learning*, pages 21–31. PMLR, 2021.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Xingjian Cao, Gang Sun, Hongfang Yu, and Mohsen Guizani. Perfed-gan: Personalized federated learning via generative adversarial networks. *IEEE Internet of Things Journal*, 10(5):3749–3762, 2023. doi: 10.1109/JIOT.2022.3172114.

Huili Chen, Jie Ding, Eric William Tramel, Shuang Wu, Anit Kumar Sahu, Salman Avestimehr, and Tao Zhang. Self-aware personalized federated learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=EqJ5_hZSqgy.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. Decentralized riemannian gradient descent on the stiefel manifold. In *International Conference on Machine Learning*, pages 1594–1605. PMLR, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010. doi: 10.1017/CBO9780511761362.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, 2020.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

John Immerkær. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300–302, 1996. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1996.0060. URL https://www.sciencedirect.com/science/article/pii/S1077314296900600.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, 2019.

Diederik P Kingma and Ruiqi Gao. Understanding the diffusion objective as a weighted integral of elbos. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Nikita Yurevich Kotelevskii, Maxime Vono, Alain Durmus, and Eric Moulines. Fedpop: A bayesian approach for personalised federated learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=KETwimTQexH.

Qi Le, Enmao Diao, Xinran Wang, Ali Anwar, Vahid Tarokh, and Jie Ding. Personalized federated recommender systems with private and partially federated autoencoders, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020, MLSys*, 2020.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34, 2021.

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*, pages 1273–1282. PMLR, 2017.

Taehong Moon, Moonseok Choi, Gayoung Lee, Jung-Woo Ha, and Juho Lee. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Kaan Ozkara, Navjot Singh, Deepesh Data, and Suhas Diggavi. Quped: Quantized personalization via distillation with applications to federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Kaan Ozkara, Antonious Girgis, Deepesh Data, and Suhas Diggavi. A statistical framework for personalized federated learning and estimation: Theory, algorithms, and privacy. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=FUiDMCr_W4o.

Kaan Ozkara, Bruce Huang, and Suhas Diggavi. Personalized PCA for federated heterogeneous data. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 168–173. IEEE, 2023b.

Chao Peng, Yiming Guo, Yao Chen, Qilin Rui, Zhengfeng Yang, and Chenyang Xu. Fedgm: Heterogeneous federated learning via generative learning and mutual distillation. In *Euro-Par 2023: Parallel Processing: 29th International Conference on Parallel and Distributed Computing, Limassol, Cyprus, August 28 – September 1, 2023, Proceedings*, page 339–351, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-39697-7. doi: 10.1007/978-3-031-39698-4_23. URL https://doi.org/10.1007/978-3-031-39698-4_23.

Mirko Polato. Federated variational autoencoder for collaborative filtering. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. doi: 10.1109/IJCNN52387.2021.9533358.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

Naichen Shi and Raed Al Kontar. Personalized pca: Decoupling shared and unique features. *arXiv preprint arXiv:2207.08041*, 2022.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pages 4424–4434, 2017.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Ye Lin Tun, Chu Myaet Thwal, Ji Su Yoon, Sun Moo Kang, Chaoning Zhang, and Choong Seon Hong. Federated learning with diffusion models for privacy-sensitive vision tasks. In *2023 International Conference on Advanced Technologies for Communications (ATC)*, pages 305–310. IEEE, 2023.

Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*, pages 509–517. PMLR, 2017.

Lei Yang, Jiaming Huang, Wanyu Lin, and Jiannong Cao. Personalized federated learning on non-iid data via group-based meta-learning. *ACM Transactions on Knowledge Discovery from Data*, 17(4):1–20, 2023.

Valentina Zantedeschi, Aurélien Bellet, and Marc Tommasi. Fully decentralized joint learning of personalized models and collaboration graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 864–874. PMLR, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, included in Sections 3 and 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, we include convergence analyses in Section 3]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, included with supplementary material]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, in Sections 3 and 4 and Appendices D and E]
   (b) Complete proofs of all theoretical results. [Yes, in Appendix]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, as part of the supplemental material]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, in Section 5 and Appendix F]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Related Works, Limitations, and Broader Impact

As mentioned earlier, there have been several recent works on personalized *supervised* learning, and we give a more detailed accounting of it in Appendix A. There has been much less attention given to personalized federated unsupervised learning. The closest work to ours on personalized dimensionality reduction is Shi and Kontar [2022], Ozkara et al. [2023b] which studies personalized PCA algorithms. Shi and Kontar [2022] has a restrictive assumption that principal components for global and local models are non-overlapping. Ozkara et al. [2023b] develops a criterion for personalized PCA; however, the authors assume heterogeneity of the setting is known; in contrast, `ADEPT-PCA` learns and adapts the solution accordingly. There is some literature on specific tasks such as training recommender systems Le et al. [2022], Polato [2021], grouping clients based on latent representations Yang et al. [2023], generating data to improve performance of personalized supervised learning Cao et al. [2023], Peng et al. [2023]. However, our approach to developing personalized dimensionality reduction for heterogeneous data is distinct from theirs. We are not aware of any FL work for personalized diffusion generative models. There is work using pre-trained diffusion models and then fine-tuning them Ruiz et al. [2023], Zhang et al. [2023], Moon et al. [2022], Ma et al. [2023]. However, these do not fit into the federated learning paradigm and require data collection from clients to obtain the pre-trained model. One way to view our approach is to *simultaneously* build such "global" models (akin to pre-trained models) and individual (personalized) models, while not sharing data.

Related Works: Some of the related works beyond those given in Section 1 are as follows: Personalized Federated Learning (FL) has seen recent advances with diverse approaches for learning personalized models. These approaches encompass meta-learning-based methods [Fallah et al., 2020, Acar et al., 2021, Khodak et al., 2019], regularization techniques [Deng et al., 2020, Mansour et al., 2020, Hanzely and Richtárik, 2020], clustered FL [Zhang et al., 2021, Mansour et al., 2020, Ghosh et al., 2020, Marfoq et al., 2021], knowledge distillation strategies [Li et al., 2020, Ozkara et al., 2021], multi-task learning [Dinh et al., 2020, Smith et al., 2017, Vanhaesebrouck et al., 2017, Zantedeschi et al., 2020], and the utilization of common representations [Du et al., 2021, Tian et al., 2020, Collins et al., 2021]. Additionally, recently there have been works on using a hierarchical Bayesian view to derive novel personalized supervised FL algorithms [Ozkara et al., 2023a, Chen et al., 2022, Kotelevskii et al., 2022]. Adaptation is also considered in [Ozkara et al., 2023a, Chen et al., 2022]. In [Chen et al., 2022], variance estimation is performed for supervised learning to estimate the heterogeneity within the local models and the estimated variance is used to form an initialization for each local model to be trained on. However, if we apply variance estimation to our criterion, we observe an early stopping issue during the training. In contrast, ours is based on a standard gradient descent. Moreover, they do not examine convergence which we do in our unsupervised algorithms. In [Ozkara et al., 2023a], the authors consider an adaptation for supervised learning based on KL-divergence criterion and do not have hyper prior. Moreover, a convergence analysis is not explored in their algorithm. In that case, the σ is inclined to smaller values or even vanishes during the training, and our `ADEPT` criterion resolves the issue by introducing the hyper prior. [Tun et al., 2023] investigated local features of globally trained diffusion models through FedAvg, but they do not consider personalized generation. We did not find any other works on personalized federated generation models.

Limitations: As we mentioned in the conclusion, the introduction of information constraints such as privacy is an important future work. Without formal privacy guarantees, an adversarial client can utilize a global model to generate data similar to other clients' data. Being the first such result, we have a simplified setting for the theoretical analysis of `ADEPT-DGM`; our analysis can be extended to more complex settings.

Broader Impact: Discovering the structure of local data is important in many applications. As we argued, collaboration becomes critical when there is insufficient local data for doing such analysis. Such applications occur in many domains including cyber-physical systems, autonomous systems, sensing, medical devices, etc., where one wants to discover the structure of heterogeneous local data. These applications are becoming widespread and therefore solutions to these problems could have a broader impact in these applications, especially where one wants to do this without explicit data sharing. Our methods for personalization can benefit users with limited data and resources to discover their particular structure of local data. Not explicitly sharing data is a first step towards privacy; however, as mentioned earlier, building explicit privacy mechanisms is needed to give privacy guarantees. This is an important aspect that has been studied in federated learning, but has received less attention in personalized learning and is an important topic of future exploration. By incorporating this as well, many more applications with societal benefits could be realized.

B Preliminaries

For the notations used in this paper:

- We use bold lowercase letters (such as $\boldsymbol{u}$, $\boldsymbol{v}$) to denote vectors and we use bold uppercase letters (such as $\boldsymbol{U}$, $\boldsymbol{V}$) to denote matrices.
- Given a composite function $f(\boldsymbol{u}, \boldsymbol{v})$, we denote $\nabla f(\boldsymbol{u}, \boldsymbol{v})$ or $\nabla_{(\boldsymbol{u},\boldsymbol{v})} f(\boldsymbol{u}, \boldsymbol{v})$ as the gradient; $\nabla_{\boldsymbol{u}} f(\boldsymbol{u}, \boldsymbol{v})$ and $\nabla_{\boldsymbol{v}} f(\boldsymbol{u}, \boldsymbol{v})$ as the partial gradients with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$.
- For a vector $\boldsymbol{u}$, $\|\boldsymbol{u}\|$ denotes the ℓ_2-norm $\|\boldsymbol{u}\|_2$. For a matrix $\boldsymbol{U}$, $\|\boldsymbol{U}\|_2$ and $\|\boldsymbol{U}\|_{op}$ both denote the ℓ_2-norm (operator norm) of the matrix and $\|\boldsymbol{U}\|, \|\boldsymbol{U}\|_F$ denotes the Frobenius norm of the matrix.
- We use $\{\boldsymbol{U}_i\}_{i=1}^m$ or $\{\boldsymbol{U}_i\}$ (when the context is clear) to denote the collection $\{U_1, \ldots, U_m\}$. When there are multiple indices, we use $\{\boldsymbol{U}_{i,t}\}_i \coloneqq \{\boldsymbol{U}_{1,t}, \ldots, \boldsymbol{U}_{m,t}\}$ to denote the collection over index i.

C Proofs for Adaptive PCA: `ADEPT-PCA`

Algorithm 2 `ADEPT-PCA` Algorithm

Input: Number of iterations T and learning rates (η_1, η_2, η_3).

Initialize local PCs $\{\boldsymbol{U}_{i,0}\}_{i=1}^m$, global PC $\boldsymbol{V}_0$, and σ_0.
Broadcast $\boldsymbol{V}_0, \sigma_0$ to the clients
for $t = 1$ **to** T **do**
  On the clients:
  for $i = 1$ **to** m: **do**
    Receive $\boldsymbol{V}_{t-1}, \sigma_{t-1}$
    $\sigma_{i,t} = \sigma_{t-1} - \eta_3 \frac{\partial}{\partial \sigma_{t-1}} f_i^{\text{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, \sigma_{t-1})$
    $\boldsymbol{g}_{i,t}^{\boldsymbol{U}} = P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}}(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\text{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, \sigma_{t-1}))$
    $\boldsymbol{U}_{i,t} \leftarrow \mathcal{R}_{\boldsymbol{U}_{i,t-1}}(-\eta_1 \boldsymbol{g}_{i,t}^{\boldsymbol{U}})$
    $\boldsymbol{g}_{i,t}^{\boldsymbol{V}} = P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}}(\nabla_{\boldsymbol{V}_{t-1}} f_i^{\text{pca}}(\boldsymbol{U}_{i,t}, \boldsymbol{V}_{t-1}, \sigma_{t-1}))$
    $\boldsymbol{V}_{i,t} \leftarrow \boldsymbol{V}_{t-1} - \eta_2 \boldsymbol{g}_{i,t}^{\boldsymbol{V}}$
    Send $\boldsymbol{V}_{i,t}$, $\sigma_{i,t}$ to the server
  end for
  At the server:
  Receive $\{\boldsymbol{V}_{i,t}\}_{i=1}^m$ and $\{\sigma_{i,t}\}_{i=1}^m$
  $\boldsymbol{V}_t = \mathcal{R}_{\boldsymbol{V}_{t-1}}\left(\frac{1}{m}\sum_{i=1}^m \boldsymbol{V}_{i,t} - \boldsymbol{V}_{t-1}\right)$
  $\sigma_t = \frac{1}{m}\sum_{i=1}^m \sigma_{i,t}$
  Broadcast $\boldsymbol{V}_t, \sigma_t$ to the clients
end for

Output: Personalized PCs $\{\boldsymbol{U}_{1,T}, \ldots, \boldsymbol{U}_{m,T}\}$.

C.1 Proof Summary

First, we define the polar retraction used in `ADEPT-PCA` and show the non-expansiveness and Lipschitz properties under the polarization retraction (Lemma C.1, Lemma C.2). Next, we establish a lower bound on σ_t given a proper initialization σ_0 (Lemma 3.2). With the lower bound on σ_t, we are able to derive the Lipschitz smoothness constants and gradients bounds with respect to $\boldsymbol{U}$, $\boldsymbol{V}$ (Lemma C.5 and Lemma C.6). Also, we derive the Lipschitz continuous constants of the derivative $\frac{\partial}{\partial \sigma} f_i^{\text{pca}}$ with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ (Lemma C.7). With the Lipschitz constants, we derive the sufficient decrease of the loss function for $\boldsymbol{U}$, $\boldsymbol{V}$, and σ (Lemma C.8). Finally, we combine them and get to our Theorem 3.3.

C.2 Lemmas

For any point $\boldsymbol{U} \in St(d, r)$, a retraction at a point $\boldsymbol{U} \in St(d, r)$ is a map $\mathcal{R}_{\boldsymbol{U}} : \mathcal{T}_{\boldsymbol{U}} \to St(d, r)$ that induces local coordinates on the Stiefel manifold. In this work, we use polar retraction that is defined as $\mathcal{R}_{\boldsymbol{U}}(\boldsymbol{V}) =$

$(\boldsymbol{U}+\boldsymbol{V})(\boldsymbol{I}+\boldsymbol{V}^\top\boldsymbol{V})^{-\frac{1}{2}}$. The polar retraction is a second-order retraction that approximates the exponential mapping up to second-order terms. Consequently, it possesses the following non-expansiveness property and we state the Lemmas and properties that we use throughout the proof before showing the proofs.

Lemma C.1 (Non-expansiveness of polar retraction Chen et al. [2021]). *Let $\boldsymbol{V} \in St(d,r)$, for any point $\boldsymbol{U} \in \mathcal{T}_{\boldsymbol{V}}$ with bounded norm, $\|\boldsymbol{U}\|_F \leq M$, there exists $C \in \mathbb{R}$ such that*

$$\|\mathcal{R}_{\boldsymbol{V}}(\boldsymbol{U}) - (\boldsymbol{V}+\boldsymbol{U})\|_F \leq C\|\boldsymbol{U}\|_F^2. \tag{14}$$

Lemma C.2 (Lipschitz type inequality Chen et al. [2021]). *Let $\boldsymbol{U},\boldsymbol{V} \in St(d,r)$. If a function ψ is L-Lipschitz smooth in $\mathbb{R}^{d\times r}$, the following inequality holds:*

$$|\psi(\boldsymbol{V}) - (\psi(\boldsymbol{U}) + \langle P_{\mathcal{T}_{\boldsymbol{U}}}(\nabla\psi(\boldsymbol{U})), \boldsymbol{V}-\boldsymbol{U}\rangle)| \leq \frac{L_g}{2}\|\boldsymbol{V}-\boldsymbol{U}\|_F^2$$

where $L_g = L + G$ with $G \coloneqq \max_{\boldsymbol{U}\in St(d,r)} \|\nabla\psi(\boldsymbol{U})\|_2$.

Assumption C.3. For each client i, the operator and Frobenius norms of $\boldsymbol{S}_i$ are bounded by

$$\|\boldsymbol{S}_i\|_F \leq G_{i,F} \quad \text{and} \quad \|\boldsymbol{S}_i\|_{op} \leq G_{i,op},$$

and we define $G_{max,F} \coloneqq \max_{i\in[m]} G_{i,F}$ and $G_{max,op} \coloneqq \max_{i\in[m]} G_{i,op}$.

For the remaining part of the paper, we fix some $\omega \in (0,1)$ and initialize σ_0 corresponding to the ω in order to obtain the lower bound in Lemma 3.2.

Lemma C.4 (Lipschitz smoothness and bounded gradients with respect to σ). *For all $i \in [m]$ and $\boldsymbol{U}_i, \boldsymbol{V} \in St(d,r)$, within the domain $\sigma \in [\omega\sqrt{\frac{2\xi}{d}}, \infty]$, the function $f_i^{\text{pca}}(\boldsymbol{U}_i,\boldsymbol{V},\sigma)$ is L_σ-Lipschitz smooth with respect to σ with constants*

$$L_\sigma \coloneqq \frac{d^2}{2\xi\omega^2} + \frac{3d^2}{2\xi\omega^4} + \frac{3d^2}{\xi^2\omega^4}.$$

Lemma C.5 (Lipschitz smoothness and bounded gradients with respect to $\boldsymbol{U}_i$). *The function $f_i^{\text{pca}}(\boldsymbol{U}_i,\boldsymbol{V},\sigma)$ is L_U-Lipschitz smooth with respect to $\boldsymbol{U}_i$ and $\|\nabla f_i^{\text{pca}}(\boldsymbol{U}_i,\boldsymbol{V},\sigma)\|_2 \leq G_U$ for all $i \in [m]$ with constants*

$$L_U \coloneqq \frac{n}{2}\left(\frac{1}{\sigma_\epsilon^2} + \frac{G_{max,op}}{\sigma_\epsilon^4} + \left(1 + \frac{2G_{max,op}}{\sigma_\epsilon^2}\right)\frac{2}{\sigma_\epsilon^4}\right) + \frac{d}{\xi\omega^2},$$
$$G_U \coloneqq \frac{n}{2}\left(\frac{G_{max,op}}{\sigma_\epsilon^4} + \frac{1}{\sigma_\epsilon^2}\right) + \frac{d}{\xi\omega^2}.$$

Lemma C.6 (Lipschitz smoothness and bounded gradients with respect to $\boldsymbol{V}$). *The function $f^{\text{pca}}(\{\boldsymbol{U}_i\}_i,\boldsymbol{V},\sigma)$ is L_V-Lipschitz smooth with respect to $\boldsymbol{V}$ and $\|\nabla f^{\text{pca}}(\{\boldsymbol{U}_i\}_i,\boldsymbol{V},\sigma)\|_2 \leq G_V$ with constants*

$$L_V \coloneqq \frac{12d}{\xi\omega^2},$$
$$G_V \coloneqq \frac{3d}{\xi\omega^2}.$$

Lemma C.7 (Lipschitz continuity of $\frac{\partial}{\partial\sigma}f_i^{\text{pca}}(\boldsymbol{U},\boldsymbol{V},\sigma)$ with respect to $\boldsymbol{U},\boldsymbol{V}$). *The function $\frac{\partial}{\partial\sigma}f_i^{\text{pca}}(\boldsymbol{U},\boldsymbol{V},\sigma)$ is $L_U^{(\sigma)}$-Lipschitz continuous with respect to $\boldsymbol{U}$ and $L_V^{(\sigma)}$-Lipschitz continuous with respect to $\boldsymbol{V}$ with*

$$L_U^{(\sigma)} = \frac{\sqrt{2d^3}}{\omega^3\sqrt{\xi^3}},$$
$$L_V^{(\sigma)} = \frac{2\sqrt{d^3}}{\omega^3\sqrt{2\xi^3}}.$$

Before showing the convergence results, we define the following terms. Let

$$C_{\eta_1} = C_1G_1 + \frac{(L_U+G_U)(C_1^2G_1^2+1)}{2},$$

$$C\eta_2 = C_2G_2 + \frac{(L_V + G_V)(C_2^2G_2^2+1)}{2},$$
$$G_1 = 2G_U\sqrt{d},$$
$$G_2 = 2G_V\sqrt{d}. \tag{15}$$

with some constants C_1, C_2 given by Lemma C.1 and G_U, G_V given in Lemma C.5, C.6.

Lemma C.8 (Sufficient Decrease)**.** *At any iteration t, we have*

$$\begin{aligned}
&f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_t) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \\
\leq &\left(-\eta_1 + C_{\eta_1}\eta_1^2 + \eta_3(L_U^{(\sigma)})^2\right)\frac{1}{m}\sum_{i=1}^m \|P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}}\left(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\text{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, , \sigma_{t-1})\right)\|_F^2 \\
&+ \left(-\eta_2 + C\eta_2\eta_2^2 + \eta_3(L_V^{(\sigma)})^2\right)\|P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}}\left(\nabla_{\boldsymbol{V}_{t-1}} f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right)\|_F^2 \\
&+ \left(\frac{-\eta_3 + \eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2.
\end{aligned}$$

Theorem C.9. *By choosing $\eta_1 = \min\{\frac{1}{3C_{\eta_1}}, 1\}$, $\eta_2 = \min\{\frac{1}{3C\eta_2}, 1\}$, and $\eta_3 = \min\left\{\frac{\eta_1}{3(L_U^{(\sigma)})^2}, \frac{\eta_2}{3(L_V^{(\sigma)})^2}, \frac{1}{bL_\sigma}\right\}$, we have*

$$\sum_{t=1}^T\left(\left(\frac{1}{m}\sum_{i=1}^m\|\boldsymbol{g}_{i,t}^{\boldsymbol{U}}\|_F^2\right) + \|\boldsymbol{g}_t^{\boldsymbol{V}}\|_F^2 + (\boldsymbol{g}_t^\sigma)^2\right) \leq \frac{3\Delta_T}{\min\{\eta_1, \eta_2, \eta_3\}}$$

where

$$\begin{aligned}
\boldsymbol{g}_t^{\boldsymbol{V}} &= P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}}\left(\nabla_{\boldsymbol{V}_{t-1}} f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right), \\
\boldsymbol{g}_{i,t}^{\boldsymbol{U}} &= P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}}\left(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\text{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, , \sigma_{t-1})\right), \\
\boldsymbol{g}_t^\sigma &= \frac{\partial}{\partial\sigma_{t-1}} f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}), \\
\Delta_T &= f^{\text{pca}}(\{\boldsymbol{U}_{i,0}\}_i, \boldsymbol{V}_0, \sigma_0) - f^{\text{pca}}(\{\boldsymbol{U}_{i,T}\}_i, \boldsymbol{V}_T, \sigma_T).
\end{aligned}$$

C.3 Proofs

Fact C.10. The gradients of the local loss function with respect to the local and global PC's and σ are given as

$$\begin{aligned}
\nabla_{\boldsymbol{U}_i} f_i^{\text{pca}}(\boldsymbol{U}_i, \boldsymbol{V}, \sigma) &= -\frac{n}{2}(\boldsymbol{W}_i^{-1}\boldsymbol{S}_i\boldsymbol{W}_i^{-1}\boldsymbol{U}_i - \boldsymbol{W}_i^{-1}\boldsymbol{U}_i) + \frac{\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_i)}{\sigma^2}, \\
\nabla_{\boldsymbol{V}} f_i^{\text{pca}}(\boldsymbol{U}_i, \boldsymbol{V}) &= -\frac{\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_i)(\boldsymbol{U}_i^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{U}_i)}{2\sigma^2}, \\
\frac{\partial}{\partial\sigma} f_i^{\text{pca}}(\boldsymbol{U}_i, \boldsymbol{V}, \sigma) &= \frac{d}{\sigma} - \frac{2\xi + d^2(\boldsymbol{V}, \boldsymbol{U}_i)}{\sigma^3}, \\
\nabla_{\boldsymbol{U}_i}\left(\frac{\partial}{\partial\sigma} f_i^{\text{pca}}(\boldsymbol{U}_i, \boldsymbol{V}, \sigma)\right) &= -\frac{2\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_i)}{\sigma^3}, \\
\nabla_{\boldsymbol{V}}\left(\frac{\partial}{\partial\sigma} f_i^{\text{pca}}(\boldsymbol{U}_i, \boldsymbol{V}, \sigma)\right) &= \frac{\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_i)(\boldsymbol{U}_i^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{U}_i)}{\sigma^3}.
\end{aligned}$$

Fact C.11. For two matrices $\boldsymbol{A} \in \mathbb{R}^{a\times b}$ and $\boldsymbol{B} \in \mathbb{R}^{b\times c}$, we have

$$\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_{op}\|\boldsymbol{B}\|_F \text{ and } \|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_{op}.$$

Fact C.12. For matrix to matrix functions, $\{g_i\}_{i=1}^k$, with bounded output operator norms, $\max_{\boldsymbol{X}}\|g_i(\boldsymbol{X})\|_{op} \leq M_i$, we have

$$\|\prod_{i=1}^k g_i(\boldsymbol{X}) - \prod_{i=1}^k g_i(\boldsymbol{Y})\|_F \leq \prod_{j=1}^k M_j\left(\sum_{i=1}^k\|g_i(\boldsymbol{X}) - g_i(\boldsymbol{Y})\|_F\right)$$

C.3.1 Proof of Lemma 3.2

Proof. We use mathematical induction to prove the lemma. For the base case, it is given that $\sigma_0 \geq \omega\sqrt{\frac{2\xi}{d}}$. Assume that for all $\tau \in \{0, 1, \ldots, t\}$,

$$\sigma_\tau \geq \omega\sqrt{\frac{2\xi}{d}}.$$

Then, we consider the following two cases. First, if $\sigma_t \in \left[\omega\sqrt{\frac{2\xi}{d}}, \sqrt{\frac{2\xi}{d}}\right]$, we have

$$\begin{aligned}
&\sigma_t \leq \sqrt{\frac{2\xi}{d}} \quad \Rightarrow \quad d \leq \frac{2\xi}{\sigma_t^2} \quad \Rightarrow \quad \frac{d}{\sigma_t} - \frac{2\xi}{\sigma_t^3} \leq 0 \\
\Rightarrow \quad &\forall i \in [m]: \quad \frac{\partial}{\partial \sigma_t} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t}, \boldsymbol{V}_t, \sigma_t) = \frac{d}{\sigma_t} - \frac{2\xi + d^2(\boldsymbol{V}_t, \boldsymbol{U}_{i,t})}{\sigma_t^3} \leq \frac{d}{\sigma_t} - \frac{2\xi}{\sigma_t^3} \leq 0 \\
\Rightarrow \quad &\forall i \in [m]: \quad \sigma_{i,t+1} = \sigma_t - \eta_3 \frac{\partial}{\partial \sigma_t} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t}, \boldsymbol{V}_t, \sigma_t) \geq \sigma_t \geq \omega\sqrt{\frac{2\xi}{d}} \\
\Rightarrow \quad &\sigma_{t+1} = \frac{1}{m}\sum_{i=1}^m \sigma_{i,t+1} \geq \omega\sqrt{\frac{2\xi}{d}}.
\end{aligned}$$

Otherwise, if we have $\sigma_t > \sqrt{\frac{2\xi}{d}}$, we have

$$\begin{aligned}
\sigma_{i,t+1} &= \sigma_t - \eta_3 \frac{\partial}{\partial \sigma_t} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t}, \boldsymbol{V}_t, \sigma_t) = \sigma_t - \eta_3 \left(\frac{d}{\sigma_t} - \frac{2\xi + d^2(\boldsymbol{V}_t, \boldsymbol{U}_{i,t})}{\sigma_t^3}\right) \\
&\geq \sigma_t - \eta_3 \frac{d}{\sigma_t} \geq \sqrt{\frac{2\xi}{d}} - (1-\omega)\frac{2\xi}{d^2} \cdot \frac{d}{\sqrt{2\xi/d}} = \omega\sqrt{\frac{2\xi}{d}}
\end{aligned}$$

and thus

$$\sigma_{t+1} = \frac{1}{m}\sum_{i=1}^m \sigma_{i,t+1} \geq \omega\sqrt{\frac{2\xi}{d}}.$$

Thus, by mathematical induction, we have

$$\forall t \in \mathbb{N} \ \ \forall i \in [m]: \quad \sigma_t \geq \omega\sqrt{\frac{2\xi}{d}} \quad \text{and} \quad \sigma_{i,t} \geq \omega\sqrt{\frac{2\xi}{d}}.$$

□

C.3.2 Proof of Lemma C.4

Proof. For the Lipschitz smoothness, we have

$$\begin{aligned}
\left|\frac{\partial^2}{\partial \sigma^2} f_i^{\mathrm{pca}}(\boldsymbol{U}_i, \boldsymbol{V}, \sigma)\right| &= \left|\frac{6\xi + 3d^2(\boldsymbol{V}, \boldsymbol{U}_i)}{\sigma^4} - \frac{d}{\sigma^2}\right| \\
&\leq \left|\frac{6\xi + 3d^2(\boldsymbol{V}, \boldsymbol{U}_i)}{\sigma^4}\right| + \left|\frac{d}{\sigma^2}\right| \\
&\leq \frac{6\xi + 12}{4\xi^2\omega^4/d^2} + \frac{d}{2\xi\omega^2/d} \\
&= \frac{3d^2}{2\xi\omega^4} + \frac{3d^2}{\xi^2\omega^4} + \frac{d^2}{2\xi\omega^2} \\
&= L_\sigma
\end{aligned}$$

for any $\boldsymbol{V}$, $\boldsymbol{U}_i$, and $\sigma \geq \omega\sqrt{\frac{2\xi}{d}}$. □

C.3.3 Proof of Lemma C.5

Proof. For the bound on the gradient,

$$\begin{aligned}
&\left\| -\frac{n}{2}(\boldsymbol{W}_i^{-1}S_i\boldsymbol{W}_i^{-1}\boldsymbol{U}_i - \boldsymbol{W}_i^{-1}\boldsymbol{U}_i) + \frac{\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_i)}{\sigma^2} \right\|_{op} \\
\leq & \| -\frac{n}{2}(\boldsymbol{W}_i^{-1}S_i\boldsymbol{W}_i^{-1}\boldsymbol{U}_i - \boldsymbol{W}_i^{-1}\boldsymbol{U}_i)\|_{op} + \frac{2}{\sigma^2} \\
\leq & \frac{n}{2}(\|\boldsymbol{W}_i^{-1}S_i\boldsymbol{W}_i^{-1}\boldsymbol{U}_i\|_{op} + \|\boldsymbol{W}_i^{-1}\boldsymbol{U}_i\|_{op}) + \frac{2}{\sigma^2} \\
\leq & \frac{n}{2}\left(\frac{G_{max,op}}{\sigma_\epsilon^4} + \frac{1}{\sigma_\epsilon^2}\right) + \frac{d}{\xi\omega^2},
\end{aligned}$$

where in the last inequality we use $\|\boldsymbol{W}_i^{-1}\|_{op} \leq \frac{1}{\sigma_\epsilon^2}$. Therefore, we find that the norm of the gradient is bounded by $G_U := \frac{n}{2}(\frac{G_{max,op}}{\sigma_\epsilon^4} + \frac{1}{\sigma_\epsilon^2}) + \frac{d}{\xi\omega^2}$. For the Lipschitz continuity of the gradient, we omit the client index i and use $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to denote two arbitrary points on $St(d,r)$ for simplicity. For any client i, we focus on the first term of the gradient,

$$\begin{aligned}
&\|\boldsymbol{W}_1^{-1}S_i\boldsymbol{W}_1^{-1}\boldsymbol{U}_1 - \boldsymbol{W}_1^{-1}\boldsymbol{U}_1 - \boldsymbol{W}_2^{-1}S_i\boldsymbol{W}_2^{-1}\boldsymbol{U}_2 + \boldsymbol{W}_2^{-1}\boldsymbol{U}_2\|_F \\
&\leq \left(\frac{1}{\sigma_\epsilon^2} + \frac{G_{max,op}}{\sigma_\epsilon^4} + \left(1 + \frac{2G_{max,op}}{\sigma_\epsilon^2}\right)\frac{2}{\sigma_\epsilon^4}\right)\|\boldsymbol{U}_2 - \boldsymbol{U}_1\|_F,
\end{aligned} \tag{16}$$

where we defer the algebra to the Appendix. For the second part of the gradient we have

$$\begin{aligned}
&\frac{1}{\sigma^2}\|\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_1) - \mathcal{P}_{\mathcal{T}_{\boldsymbol{V}}}(\boldsymbol{U}_2)\|_F \\
&= \frac{1}{\sigma^2}\|\boldsymbol{U}_1 - \boldsymbol{U}_2 - \frac{1}{2}\boldsymbol{V}(\boldsymbol{V}^\top(\boldsymbol{U}_1 - \boldsymbol{U}_2) + (\boldsymbol{U}_1^\top - \boldsymbol{U}_2^\top)\boldsymbol{V})\|_F \\
&\leq \frac{2}{\sigma^2}\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_F, \\
&\leq \frac{d}{\xi\omega^2}\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_F,
\end{aligned}$$

where in the last inequality we use Fact C.11. As a result, we find that the gradient is Lipschitz continuous with $L_U := \frac{n}{2}\left(\frac{1}{\sigma_\epsilon^2} + \frac{G_{max,op}}{\sigma_\epsilon^4} + \left(1 + \frac{2G_{max,op}}{\sigma_\epsilon^2}\right)\frac{2}{\sigma_\epsilon^4}\right) + \frac{d}{\xi\omega^2}$. □

C.3.4 Proof of Lemma C.6

Proof. For the Lipschitz constant

$$\begin{aligned}
&\frac{2}{\sigma^2}\|\mathcal{P}_{\mathcal{T}_{\boldsymbol{V}_1}}(\boldsymbol{U}_i)\mathrm{sym}(\boldsymbol{U}_i^\top\boldsymbol{V}_1) - \mathcal{P}_{\mathcal{T}_{\boldsymbol{V}_2}}(\boldsymbol{U}_i)\mathrm{sym}(\boldsymbol{U}_i^\top\boldsymbol{V}_2)\|_F \\
&= \frac{2}{\sigma^2}\Big\|\boldsymbol{U}_i\boldsymbol{U}_i^\top(\boldsymbol{V}_1 - \boldsymbol{V}_2) + \boldsymbol{U}_i(\boldsymbol{V}_1 - \boldsymbol{V}_2)\boldsymbol{U}_i^\top \\
&\qquad - \frac{1}{2}(\boldsymbol{V}_1(\boldsymbol{V}_1^\top\boldsymbol{U}_i + \boldsymbol{U}_i^\top\boldsymbol{V}_1) - \boldsymbol{V}_2(\boldsymbol{V}_2^\top\boldsymbol{U}_i + \boldsymbol{U}_i^\top\boldsymbol{V}_2))\Big\|_F \\
&\leq \frac{24}{\sigma^2}\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_F \\
&\leq \frac{12d}{\xi\omega^2},
\end{aligned}$$

where $\mathrm{sym}(\boldsymbol{U}_i^\top\boldsymbol{V}) = \boldsymbol{U}_i^\top\boldsymbol{V} + \boldsymbol{V}^\top\boldsymbol{U}_i$ and we used Fact C.12, hence $L_V = \frac{12d}{\xi\omega^2}$. For the gradient bound, is it straightforward to see that

$$\|\nabla_{\boldsymbol{V}} f_i^{\mathrm{pca}}(\boldsymbol{U}, \boldsymbol{V}, \sigma)\|_2 \leq \frac{4}{\sigma^2} \leq \frac{2d}{\xi\omega^2}.$$

□

C.3.5 Proof of Lemma C.7

Proof. Using Fact C.10, we have

$$\|\nabla_{\boldsymbol{U}_i}\left(\frac{\partial}{\partial\sigma}f_i^{\text{pca}}(\boldsymbol{U}_i,\boldsymbol{V},\sigma)\right)\|_2 \leq \frac{4}{\sigma^3} \leq \frac{\sqrt{2d^3}}{\omega^3\sqrt{\xi^3}}$$

and

$$\|\nabla_{\boldsymbol{V}}\left(\frac{\partial}{\partial\sigma}f_i^{\text{pca}}(\boldsymbol{U}_i,\boldsymbol{V},\sigma)\right)\|_2 \leq \frac{8}{\sigma^3} \leq \frac{2\sqrt{2d^3}}{\omega^3\sqrt{\xi^3}}$$

□

C.3.6 Proof of Lemma C.8

Proof. We have

$$\begin{aligned}
&f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_t) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\\
&= [f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_t) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1})]\\
&+ [f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1}) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})]\\
&+ [f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1}) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})]
\end{aligned}$$

With a similar proof as Lemma 5 in Ozkara et al. [2023b], we have

$$\begin{aligned}
&f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1}) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\\
&\qquad\leq \frac{(-\eta_1+C_{\eta_1}\eta_1^2)}{m}\sum_{i=1}^{m}\|P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}}(\nabla_{\boldsymbol{U}_{i,t-1}}f_i^{\text{pca}}(\boldsymbol{U}_{i,t-1},\boldsymbol{V}_{t-1},\sigma_{i,t}))\|_F^2,\\
&f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t},\sigma_{t-1}) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\\
&\qquad\leq (-\eta_2 + C\eta_2\eta_2^2)\|P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}}(\nabla_{\boldsymbol{V}_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1}))\|_F^2
\end{aligned}$$

where

$$\begin{aligned}
C_{\eta_1} &= C_1G_1 + \frac{L_{gu}(C_1^2G_1^2+1)}{2},\\
C\eta_2 &= C_2G_2 + \frac{L_{gv}(C_2^2G_2^2+1)}{2},\\
G_1 &= 2G_U\sqrt{d},\\
G_2 &= 2G_V\sqrt{d}
\end{aligned}$$

with some constants C_1, C_2 given by Lemma C.1 and G_U, G_V given in Lemma C.5, C.6. For the sufficient decrease with respect to σ, we have

$$\begin{aligned}
&f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_t) - f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1})\\
&\leq \frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1})(\sigma_t-\sigma_{t-1}) + \frac{L_\sigma}{2}(\sigma_t-\sigma_{t-1})^2\\
&= \left[\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1})\right]\left[-\eta_3\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\right]\\
&\quad+\frac{\eta_3^2L_\sigma}{2}\left[\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\right]^2 \qquad \text{(by the update rule of } \sigma_t)\\
&= (-\eta_3)\left[\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t}\}_i,\boldsymbol{V}_t,\sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\right.\\
&\quad\left.+\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\right]\left[\frac{\partial}{\partial\sigma_{t-1}}f^{\text{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i,\boldsymbol{V}_{t-1},\sigma_{t-1})\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{\eta_3^2 L_\sigma}{2}\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&= (-\eta_3)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right] \\
&\quad \times \left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right] \\
&\quad - \left(\eta_3 - \frac{\eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&\le \frac{\eta_3}{2}\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&\quad + \frac{\eta_3}{2}\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&\quad - \left(\eta_3 - \frac{\eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \qquad \text{(since } 2ab \le a^2 + b^2 \text{ for any } a, b \in \mathbb{R}\text{)} \\
&= \frac{\eta_3}{2}\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&\quad - \left(\frac{\eta_3}{2} - \frac{\eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&= \frac{\eta_3}{2}\Bigg[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f_i^{\mathrm{pca}}(\boldsymbol{V}_t, \{\boldsymbol{U}_{i,t-1}\}_i, \sigma_{t-1}) \\
&\quad + \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\Bigg]^2 \\
&\quad - \left(\frac{\eta_3}{2} - \frac{\eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&= \frac{\eta_3}{2}\Bigg[\left(\frac{1}{m}\sum_{i=1}^m \frac{\partial}{\partial\sigma_{t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t}, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t}, \sigma_{t-1})\right) \\
&\quad + \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t}, \sigma_{t-1}) - \frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\Bigg]^2 \\
&\quad - \left(\frac{\eta_3}{2} - \frac{\eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \\
&\le \frac{\eta_3}{2}\left[\left(\frac{1}{m}\sum_{i=1}^m L_U^{(\sigma)} \|\boldsymbol{U}_{i,t} - \boldsymbol{U}_{i,t-1}\|_F\right) + L_V^{(\sigma)} \|\boldsymbol{V}_{i,t} - \boldsymbol{V}_{i,t-1}\|_F\right]^2 \\
&\quad - \left(\frac{\eta_3 - \eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \qquad \text{(from Lemma C.7)} \\
&\le \frac{\eta_3}{2}\left[2\left(\frac{1}{m}\sum_{i=1}^m L_U^{(\sigma)} \|\boldsymbol{U}_{i,t} - \boldsymbol{U}_{i,t-1}\|_F\right)^2 + 2\left(L_V^{(\sigma)} \|\boldsymbol{V}_{i,t} - \boldsymbol{V}_{i,t-1}\|_F\right)^2\right] \\
&\quad - \left(\frac{\eta_3 - \eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \qquad \text{(since } (a+b)^2 \le 2a^2 + 2b^2\text{)} \\
&\le \eta_3\left[\frac{1}{m}\sum_{i=1}^m \left(L_U^{(\sigma)} \|\boldsymbol{U}_{i,t} - \boldsymbol{U}_{i,t-1}\|_F\right)^2 + \left(L_V^{(\sigma)} \|\boldsymbol{V}_{i,t} - \boldsymbol{V}_{i,t-1}\|_F\right)^2\right] \\
&\quad - \left(\frac{\eta_3 - \eta_3^2 L_\sigma}{2}\right)\left[\frac{\partial}{\partial\sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})\right]^2 \qquad \text{(Cauchy–Schwarz inequality)}
\end{aligned}
$$

$$
\begin{aligned}
&= \eta_3 (L_U^{(\sigma)})^2 \frac{1}{m} \left(\sum_{i=1}^{m} \| P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}} \left(\nabla_{\boldsymbol{U}_{i,t-1}} f^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right) \|_F^2 \right) \\
&\quad + \eta_3 (L_V^{(\sigma)})^2 \| P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}} \left(\nabla_{\boldsymbol{V}_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right) \|_F^2 \\
&\quad - \left(\frac{\eta_3 - \eta_3^2 L_\sigma}{2} \right) \left[\frac{\partial}{\partial \sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right]^2 .
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
& f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_t) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \\
= \; & [f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})] \\
& + [f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_{t-1}) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1})] \\
& + [f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_t) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_{t-1})] \\
\leq \; & \left(-\eta_1 + C_{\eta_1} \eta_1^2 + \eta_3 (L_U^{(\sigma)})^2 \right) \frac{1}{m} \sum_{i=1}^{m} \| P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}} \left(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, , \sigma_{t-1}) \right) \|_F^2 \\
& + \left(-\eta_2 + C\eta_2 \eta_2^2 + \eta_3 (L_V^{(\sigma)})^2 \right) \| P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}} \left(\nabla_{\boldsymbol{V}_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right) \|_F^2 \\
& + \left(\frac{-\eta_3 + \eta_3^2 L_\sigma}{2} \right) \left[\frac{\partial}{\partial \sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right]^2 .
\end{aligned}
$$

By choosing $\eta_1 = \min\{\frac{1}{3C_{\eta_1}}, 1\}$, $\eta_2 = \min\{\frac{1}{3C\eta_2}, 1\}$, and $\eta_3 = \min\left\{ \frac{\eta_1}{3(L_U^{(\sigma)})^2}, \frac{\eta_2}{3(L_V^{(\sigma)})^2}, \frac{1}{6L_\sigma} \right\}$, we have

$$
\begin{aligned}
-\eta_1 + C_{\eta_1} \eta_1^2 + \eta_3 (L_U^{(\sigma)})^2 &\leq \eta_1 \left(C_{\eta_1} \eta_1 - \frac{1}{3} \right) - \frac{2\eta_1}{3} + \frac{\eta_1}{3} = -\frac{\eta_1}{3}, \\
-\eta_2 + C\eta_2 \eta_2^2 + \eta_3 (L_V^{(\sigma)})^2 &\leq \eta_2 \left(C\eta_2 \eta_2 - \frac{1}{3} \right) - \frac{2\eta_2}{3} + \frac{\eta_2}{3} = -\frac{\eta_2}{3}, \\
\frac{-\eta_3 + \eta_3^2 L_\sigma}{2} &= \eta_3 \left(L_\sigma \eta_3 - \frac{1}{6} \right) - \frac{\eta_3}{3} \leq -\frac{\eta_3}{3}.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
& f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, \sigma_t) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \\
\leq \; & -\frac{\eta_1}{3} \left(\frac{1}{m} \sum_{i=1}^{m} \| P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}} \left(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, , \sigma_{t-1}) \right) \|_F^2 \right) \\
& - \frac{\eta_2}{3} \| P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}} \left(\nabla_{\boldsymbol{V}_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right) \|_F^2 - \frac{\eta_3}{3} \left[\frac{\partial}{\partial \sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right]^2 .
\end{aligned}
$$

□

C.3.7 Proof of Theorem 3.3

Proof. Following Lemma C.8, by telescoping across the iterations, we have

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=1}^{T} \Bigg[\| P_{\mathcal{T}_{\boldsymbol{V}_{t-1}}} \left(\nabla_{\boldsymbol{V}_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right) \|_F^2 \\
& + \left(\frac{1}{m} \sum_{i=1}^{m} \| P_{\mathcal{T}_{\boldsymbol{U}_{i,t-1}}} \left(\nabla_{\boldsymbol{U}_{i,t-1}} f_i^{\mathrm{pca}}(\boldsymbol{U}_{i,t-1}, \boldsymbol{V}_{t-1}, , \sigma_{t-1}) \right) \|_F^2 \right) \\
& + \left[\frac{\partial}{\partial \sigma_{t-1}} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) \right]^2 \Bigg] \\
\leq \; & \frac{1}{T \min\{\frac{\eta_1}{3}, \frac{\eta_2}{3}, \frac{\eta_3}{3}\}} \sum_{t=1}^{T} f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t-1}\}_i, \boldsymbol{V}_{t-1}, \sigma_{t-1}) - f^{\mathrm{pca}}(\{\boldsymbol{U}_{i,t}\}_i, \boldsymbol{V}_t, , \sigma_t)
\end{aligned}
$$

$$= \frac{3\left(f^{\text{pca}}(\{\boldsymbol{U}_{i,0}\}_i, \boldsymbol{V}_0, \sigma_0) - f^{\text{pca}}(\{\boldsymbol{U}_{i,T}\}_i, \boldsymbol{V}_T, \sigma_T)\right)}{T\min\{\eta_1, \eta_2, \eta_3\}}.$$

□

D Proofs for Adaptive AEs: ADEPT-AE

D.1 Proof Summary

Since the form of the adaptation in the loss function is similar to the PCA loss function, the lower bound on sigma (Lemma 3.2) holds for ADEPT-AE as well. We utilize the lower bound to derive the Lipschitz smoothness constants with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ (Lemma D.1), and the Lipschitz smoothness constant with respect to $\boldsymbol{\theta}$ is stated in Assumption 3.5. Then, we derive the sufficient decrease with respect to $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, and σ with the Lipschitz constants. However, we have multiple local iterations for one communication round in ADEPT-AE. Thus, we have to deal with the sufficient decrease separately depending on whether the round is a communication round. With careful derivation under the two cases, we can combine them in the end and get to our Theorem 3.6.

D.2 Lemmas

In the following proof, $d = d_\theta$ is the dimension of the models.

Lemma D.1. *The loss function $f_i^{\text{ae}}(\boldsymbol{\theta}_i, \boldsymbol{\mu}, \sigma)$ is $L_{\boldsymbol{\mu}}$-smooth w.r.t. $\boldsymbol{\mu}$ and L_σ-smooth w.r.t. σ with*

$$L_{\boldsymbol{\mu}} = \frac{d}{2\xi\omega^2}$$
$$L_\sigma = \frac{3\xi d^2}{2\xi^2\omega^4} + \frac{3d^2B^2}{\xi^2\omega^4} + \frac{d^2}{2\xi\omega^2}.$$

Also, we have

$$\|\nabla_{\boldsymbol{\mu}} f_i^{\text{ae}}(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma_1) - \nabla_{\boldsymbol{\mu}} f_i^{\text{ae}}(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma_2)\| \leq L_\sigma^{(\boldsymbol{\mu})}|\sigma_1 - \sigma_2|$$

with

$$L_\sigma^{(\boldsymbol{\mu})} = \frac{B\sqrt{d^3}}{\omega^3\sqrt{2\xi^3}}.$$

Assumption D.2. Assume that there exists some $B > 0$ such that for the weights of the autoencoders, we have $\|\boldsymbol{\mu}\| \leq B$ and $\|\boldsymbol{\theta}_i\| \leq B$ for all $i \in [m]$.

D.3 Proofs

D.3.1 Proof of Lemma D.1

Proof. Following the same proof in Lemma 3.2, we have the same lower bound on σ_t if we initialized it in the same way.

The gradient w.r.t. $\boldsymbol{\mu}$ is

$$\nabla_{\boldsymbol{\mu}} f_i^{\text{ae}}(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma) = \frac{\boldsymbol{\mu} - \boldsymbol{\theta}}{2\sigma^2}.$$

Thus, we have

$$\left\|\frac{\boldsymbol{\mu}_1 - \boldsymbol{\theta}}{2\sigma^2} - \frac{\boldsymbol{\mu}_2 - \boldsymbol{\theta}}{2\sigma^2}\right\| = \left\|\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2\sigma^2}\right\| \leq \frac{d}{2\xi\omega^2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \leq L_{\boldsymbol{\mu}}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|.$$

For L_σ, we have

$$\left|\frac{\partial^2}{\partial\sigma^2} f_i^{\text{ae}}(\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma)\right| = \left|\frac{6\xi + 3\|\boldsymbol{\mu} - \boldsymbol{\theta}\|^2}{\sigma^4} - \frac{d}{\sigma^2}\right|$$

$$
\begin{aligned}
&\leq \left|\frac{6\xi + 3\|\boldsymbol{\mu}-\boldsymbol{\theta}\|^2}{\sigma^4}\right| + \left|\frac{d}{\sigma^2}\right| \\
&\leq \frac{6\xi + 3(2B)^2}{4\xi^2\omega^4/d^2} + \frac{d^2}{2\xi\omega^2} \\
&= \frac{3\xi d^2}{2\xi^2\omega^4} + \frac{3d^2B^2}{\xi^2\omega^4} + \frac{d^2}{2\xi\omega^2} \\
&= L_\sigma.
\end{aligned}
$$

For $L_\sigma^{(\boldsymbol{\mu})}$, we have

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\mu}} f_i^{\mathrm{ae}}(\boldsymbol{\theta},\boldsymbol{\mu},\sigma_1) - \nabla_{\boldsymbol{\mu}} f_i^{\mathrm{ae}}(\boldsymbol{\theta},\boldsymbol{\mu},\sigma_2)\| &= \left\|\frac{\boldsymbol{\mu}-\boldsymbol{\theta}}{2\sigma_1^2} - \frac{\boldsymbol{\mu}-\boldsymbol{\theta}}{2\sigma_2^2}\right\| \\
&= \left|\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right| \frac{\|\boldsymbol{\mu}-\boldsymbol{\theta}\|}{2} \\
&= |\sigma_1-\sigma_2| \left|\frac{1}{\sigma_1^2\sigma_2} + \frac{1}{\sigma_1\sigma_2^2}\right| \frac{\|\boldsymbol{\mu}-\boldsymbol{\theta}\|}{2} \\
&\leq 2\left(\omega\sqrt{\frac{2\xi}{d}}\right)^{-3} B|\sigma_1-\sigma_2| \\
&= \frac{B\sqrt{d^3}}{\omega^3\sqrt{2\xi^3}}|\sigma_1-\sigma_2| \\
&= L_\sigma^{(\boldsymbol{\mu})}|\sigma_1-\sigma_2|.
\end{aligned}
$$

□

D.3.2 Proof of Theorem 3.6

Proof. Since $\boldsymbol{\mu}_t$ and σ_t are updated only when τ divides t, we consider the two cases separately.

When τ divides t First, for the sufficient decrease of $\boldsymbol{\theta}_{t,i}$, we have

$$
\begin{aligned}
&f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t},\boldsymbol{\mu}_{t-1},\sigma_{t-1}) - f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1}) \\
&\leq \left\langle \nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1}), \boldsymbol{\theta}_{i,t}-\boldsymbol{\theta}_{i,t-1}\right\rangle + \frac{L_{\boldsymbol{\theta}}}{2}\|\boldsymbol{\theta}_{i,t}-\boldsymbol{\theta}_{i,t-1}\|^2 \\
&= \left\langle \nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1}), -\eta_1\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1})\right\rangle \\
&\quad + \frac{L_{\boldsymbol{\theta}}}{2}\left\|-\eta_1\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{i,t-1},\sigma_{t-1})\right\|^2 \\
&\leq \left(-\eta_1 + \eta_1^2\frac{L_{\boldsymbol{\theta}}}{2}\right)\left\|\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1})\right\|^2.
\end{aligned}
$$

Sum over the clients and we have

$$
\begin{aligned}
&f^{\mathrm{ae}}(\{\boldsymbol{\theta}_{i,t}\},\boldsymbol{\mu}_{t-1},\sigma_{t-1}) - f^{\mathrm{ae}}(\{\boldsymbol{\theta}_{i,t-1}\},\boldsymbol{\mu}_{t-1},\sigma_{t-1}) \\
&\leq \left(-\eta_1 + \eta_1^2\frac{L_{\boldsymbol{\theta}}}{2}\right)\left(\frac{1}{m}\sum_{i=1}^m \left\|\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t-1},\boldsymbol{\mu}_{t-1},\sigma_{t-1})\right\|^2\right).
\end{aligned} \tag{17}
$$

Second, for the sufficient decrease of σ_t, define

$$
g_t^\sigma = \frac{1}{m}\sum_{i=1}^m \frac{\partial}{\partial\sigma_{t-1}} f_i^{\mathrm{ae}}(\boldsymbol{\theta}_{i,t},\boldsymbol{\mu}_{t-1},\sigma_{t-1}).
$$

Thus, $\sigma_t = \sigma_{t-1} - \eta_2 g_t^\sigma$ and

$$\begin{aligned}
&f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_t) - f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) \\
&\leq \left(\frac{\partial}{\partial \sigma_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})\right)(\sigma_t - \sigma_{t-1}) + \frac{L_\sigma}{2}(\sigma_t - \sigma_{t-1})^2 \\
&= \left(\frac{\partial}{\partial \sigma_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})\right)(-\eta_2 g_t^\sigma) + \frac{L_\sigma}{2}(-\eta_2 g_t^\sigma)^2.
\end{aligned}$$

Sum over the clients and we have

$$\begin{aligned}
f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_{t-1}, \sigma_t) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) &\leq -\eta_2 (g_t^\sigma)^2 + \eta_2^2 \frac{L_\sigma}{2} (g_t^\sigma)^2 \\
&= \left(-\eta_2 + \eta_2^2 \frac{L_\sigma}{2}\right)(g_t^\sigma)^2. \qquad (18)
\end{aligned}$$

Then, for the sufficient decrease of $\boldsymbol{\mu}_t$, define

$$\begin{aligned}
\boldsymbol{g}_{i,t}^{\boldsymbol{\mu}} &= \nabla_{\boldsymbol{\mu}_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}), \qquad & \boldsymbol{g}_t^{\boldsymbol{\mu}} &= \frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_{i,t}^{\boldsymbol{\mu}}, \\
\tilde{\boldsymbol{g}}_{i,t}^{\boldsymbol{\mu}} &= \nabla_{\boldsymbol{\mu}_{t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_t), & \tilde{\boldsymbol{g}}_t^{\boldsymbol{\mu}} &= \frac{1}{m}\sum_{i=1}^m \tilde{\boldsymbol{g}}_{i,t}^{\boldsymbol{\mu}}.
\end{aligned}$$

We have

$$\begin{aligned}
&f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_t, \sigma_t) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_{t-1}, \sigma_t) \\
&\leq \left\langle \tilde{\boldsymbol{g}}_t^{\boldsymbol{\mu}}, \boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1} \right\rangle + \frac{L_{\boldsymbol{\mu}}}{2}\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}\|^2 \\
&= -\eta_3 \left\langle \tilde{\boldsymbol{g}}_t^{\boldsymbol{\mu}} - \boldsymbol{g}_t^{\boldsymbol{\mu}} + \boldsymbol{g}_t^{\boldsymbol{\mu}}, \boldsymbol{g}_t^{\boldsymbol{\mu}} \right\rangle + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 \\
&= \left(-\eta_3 + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \eta_3 \left\langle \boldsymbol{g}_t^{\boldsymbol{\mu}} - \tilde{\boldsymbol{g}}_t^{\boldsymbol{\mu}}, \boldsymbol{g}_t^{\boldsymbol{\mu}} \right\rangle \\
&\leq \left(-\eta_3 + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{\eta_3}{2}\|\boldsymbol{g}_t^{\boldsymbol{\mu}} - \tilde{\boldsymbol{g}}_t^{\boldsymbol{\mu}}\|^2 + \frac{\eta_3}{2}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 \\
&\leq \left(-\frac{\eta_3}{2} + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{\eta_3 L_\sigma^{(\boldsymbol{\mu})^2}}{2}(\sigma_t - \sigma_{t-1})^2 \\
&\leq \left(-\frac{\eta_3}{2} + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{\eta_3 \eta_2^2 L_\sigma^{(\boldsymbol{\mu})^2}}{2}(g_t^\sigma)^2 \\
&\leq \left(-\frac{\eta_3}{2} + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{\eta_2^2 L_\sigma^{(\boldsymbol{\mu})^2}}{2}(g_t^\sigma)^2. \qquad (19)
\end{aligned}$$

Finally, we have the overall decrease when τ divides t by summing equation (18), (17), and (19),

$$\begin{aligned}
&f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_t, \sigma_t) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) \\
&\leq \left(-\eta_1 + \eta_1^2 \frac{L_{\boldsymbol{\theta}}}{2}\right)\left(\frac{1}{m}\sum_{i=1}^m \left\|\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})\right\|^2\right) \\
&\quad + \left(-\eta_2 + \eta_2^2 \frac{L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}}{2}\right)(g_t^\sigma)^2 + \left(-\frac{\eta_3}{2} + \frac{L_{\boldsymbol{\mu}}\eta_3^2}{2}\right)\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2. \qquad (20)
\end{aligned}$$

When τ does not divide t At time steps that are not communication rounds we simply have a decrease due to the updates of $\{\boldsymbol{\theta}_i\}$, that is,

$$f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_t, \sigma_t) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t-1}\}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$$

$$
\begin{aligned}
&= f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t}\}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,t-1}\}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) \\
&= \frac{1}{m}\sum_{i=1}^{m}(f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1}) - f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})) \\
&\leq \left(-\eta_1 + \eta_1^2 \frac{L_{\boldsymbol{\theta}}}{2}\right)\left(\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})\right\|^2\right).
\end{aligned}
$$

The final bound By choosing, $\eta_1 = \frac{1}{L_{\boldsymbol{\theta}}}, \eta_2 = \frac{1}{L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}}, \eta_3 = \min\{1, \frac{1}{L_{\boldsymbol{\mu}}}\}$, and by averaging over time steps while combining two type of decrease, we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{m}\sum_{i=1}^{m}\left\|\nabla_{\boldsymbol{\theta}_{i,t-1}} f_i^{\text{ae}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})\right\|^2\right) + \frac{1}{T}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{1}{T}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}(g_t^{\sigma})^2 \\
&\qquad \leq \frac{\max\{L_{\boldsymbol{\theta}}, L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}, L_{\boldsymbol{\mu}}, 1\}\Delta_T}{T},
\end{aligned}
$$

where $\Delta_T = f^{\text{ae}}(\{\boldsymbol{\theta}_{i,0}\}, \boldsymbol{\mu}_0, \sigma_0) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,T}\}, \boldsymbol{\mu}_T, \sigma_T)$. Given that τ is a finite constant let us denote $R = T/\tau$ as the number of communication rounds. Then we have,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\right\|^2\right) + \frac{1}{T}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{1}{T}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}(g_t^{\sigma})^2 \\
&\quad = \frac{R}{T}\left(\frac{1}{R}\sum_{t=1}^{T}\left(\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\right\|^2\right) + \frac{1}{R}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{1}{R}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}(g_t^{\sigma})^2\right) \\
&\quad \geq \frac{R}{T}\left(\frac{1}{R}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}\left(\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\right\|^2\right) + \frac{1}{R}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}\|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + \frac{1}{R}\sum_{\substack{t=1 \\ t\%\tau=0}}^{T}(g_t^{\sigma})^2\right) \\
&\quad \geq \frac{R}{T}\min_{t\in[T],\tau|t}\left\{\left\|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\right\|^2 + \|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + (g_t^{\sigma})^2\right\}.
\end{aligned}
$$

Finally this yields,

$$
\min_{t\in[T],\tau|t}\left\{\left\|\boldsymbol{g}_{i,t}^{\boldsymbol{\theta}}\right\|^2 + \|\boldsymbol{g}_t^{\boldsymbol{\mu}}\|^2 + (g_t^{\sigma})^2\right\} \leq \frac{\max\{L_{\boldsymbol{\theta}}, L_\sigma + L_\sigma^{(\boldsymbol{\mu})^2}, L_{\boldsymbol{\mu}}, 1\}\Delta_T^{\text{ae}}}{R},
$$

where $\Delta_T^{\text{ae}} = f^{\text{ae}}(\{\boldsymbol{\theta}_{i,0}\}_i, \boldsymbol{\mu}_0, \sigma_0) - f^{\text{ae}}(\{\boldsymbol{\theta}_{i,T}\}_i, \boldsymbol{\mu}_T, \sigma_T)$.

□

Remark D.3. In the experiments for `ADEPT-AE`, we treat sigma as a vector $\boldsymbol{\sigma} \in \mathbb{R}^d$ so that each weight in the models can learn its own σ_j instead of sharing one σ across all the weights. The convergence for the modified algorithm is almost identical to the proof here. Moreover, it can be shown that for the Lipschitz smoothness constant L_σ, the dependence on the dimension in the numerators becomes d instead of d^2. This is because our lower bound in Lemma 3.2 will not depend on d in this case.

E Details and Proofs for Adaptive Diffusion Model: `ADEPT-DGM`

The main change compared to `ADEPT-AE` is that we input a corrupted sample to the network during the forward pass (**line 11**) to train it as a denoiser. Accordingly, $f_{i,\boldsymbol{\alpha}}^{\text{df}}(\boldsymbol{\theta}_i, \boldsymbol{\mu}, \sigma) := \|\boldsymbol{\phi}(\boldsymbol{X}(1-\boldsymbol{\alpha}) + \boldsymbol{Z}\boldsymbol{\alpha}; \boldsymbol{\theta}_i) - \boldsymbol{X}\|^2 + \frac{2\xi + \|\boldsymbol{\mu}-\boldsymbol{\theta}_i\|^2}{2\sigma^2} + d\log\sigma$, α is the random noise amount.

Algorithm 3 Personalized Adaptive Diffusion Model: `ADEPT-DGM`

Input: Number of iterations T, learning rates (η_2, η_1, η_3), number of local iterations τ, sample corruption range $\gamma \in \mathbb{Z}^+$

Init local models $\{\boldsymbol{\theta}_{i,0}\}_{i=1}^m$, global model $\boldsymbol{\mu}_0$, and σ_0.
On server:
Broadcast $\boldsymbol{\mu}_0$, σ_0 to all clients
for $t = 1$ **to** T **do**
  On Clients:
  for $i = 1$ **to** m **do**
    if τ divides $t-1$ **then**
      Receive $\boldsymbol{\mu}_{t-1}$, σ_{t-1}
    end if
    Sample noise amount $\alpha_i \in \text{Uniform}[1, \dots, \gamma]$ independently for each sample and construct $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]$
    $\boldsymbol{\theta}_{i,t} = \boldsymbol{\theta}_{i,t-1} - \eta_1 \nabla_{\boldsymbol{\theta}_{i,t-1}} f^{\text{df}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\theta}_{i,t-1}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
    if τ divides t **then**
      $\boldsymbol{\mu}_{i,t} = \boldsymbol{\mu}_{t-1} - \eta_2 \nabla_{\boldsymbol{\mu}_{t-1}} f^{\text{df}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
      $\sigma_{i,t} = \sigma_{t-1} - \eta_3 \frac{\partial}{\partial \sigma_{t-1}} f^{\text{df}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\theta}_{i,t}, \boldsymbol{\mu}_{t-1}, \sigma_{t-1})$
      Send $\boldsymbol{\mu}_{i,t}, \sigma_{i,t}$ to server
    else
      $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1}$, $\sigma_t = \sigma_{t-1}$
    end if
  end for
  At the Server:
  if τ divides t **then**
    Receive $\{\boldsymbol{\mu}_{i,t}\}_{i=1}^m$ and $\{\sigma_{i,t}\}_{i=1}^m$
    $\boldsymbol{\mu}_t = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\mu}_{i,t}$, $\sigma_t = \frac{1}{m}\sum_{i=1}^m \sigma_{i,t}$
    Broadcast $\boldsymbol{\mu}_t$, σ_t to all clients
  end if
end for

Output: Personalized autoencoders $\{\boldsymbol{\theta}_{1,T}, \dots, \boldsymbol{\theta}_{m,T}\}$.

E.1 Proof of Lemma 4.1

Proof of Lemma 4.1. Let $\beta_t = \frac{1}{T-t+\sigma_0^2}$. For the stochastic differential equation

$$d\boldsymbol{x}_t^{\leftarrow} + \beta_t(\boldsymbol{x}_t^{\leftarrow} - \hat{\boldsymbol{\theta}})dt = d\boldsymbol{w}_t, \quad \boldsymbol{x}_0^{\leftarrow} \sim \mathcal{N}(0, (\sigma_0^2 + T)\boldsymbol{I}_d),$$

we have

$$\begin{aligned} d(e^{\int_0^t \beta_s ds}(\boldsymbol{x}_t^{\leftarrow} - \hat{\boldsymbol{\theta}})) &= e^{\int_0^t \beta_s ds} d\boldsymbol{x}_t^{\leftarrow} + e^{\int_0^t \beta_s ds}\beta_t(\boldsymbol{x}_t^{\leftarrow} - \hat{\boldsymbol{\theta}})dt \\ &= e^{\int_0^t \beta_s ds}\left(d\boldsymbol{x}_t^{\leftarrow} + \beta_t(\boldsymbol{x}_t^{\leftarrow} - \hat{\boldsymbol{\theta}})dt\right) = e^{\int_0^t \beta_s ds} d\boldsymbol{w}_t. \end{aligned}$$

Note that

$$e^{\int_0^t \beta_s ds} = e^{\int_0^t \frac{1}{T-s+\sigma_0^2} ds} = e^{\ln(T+\sigma_0^2) - \ln(T-t+\sigma_0^2)} = \frac{T+\sigma_0^2}{T-t+\sigma_0^2}$$

and

$$\int_0^T e^{2\int_0^t \beta_s ds} dt = \int_0^T \frac{(T+\sigma_0^2)^2}{(T+\sigma_0^2-t)^2} dt = (T+\sigma_0^2)^2\left(\frac{1}{\sigma_0^2} - \frac{1}{T+\sigma_0^2}\right) = \left(1 + \frac{T}{\sigma_0^2}\right)T.$$

It then follows that

$$e^{\int_0^T \beta_s ds}(\boldsymbol{x}_T^{\leftarrow} - \hat{\boldsymbol{\theta}}) - (\boldsymbol{x}_0^{\leftarrow} - \hat{\boldsymbol{\theta}}) \sim \mathcal{N}\left(0, \int_0^T e^{2\int_0^t \beta_s ds} dt\right) = \mathcal{N}\left(0, \left(1 + \frac{T}{\sigma_0^2}\right)T\right),$$

and equivalently

$$\boldsymbol{x}_T^{\leftarrow} = \hat{\boldsymbol{\theta}} + \frac{\sigma_0^2}{\sigma_0^2+T}(\boldsymbol{x}_0^{\leftarrow} - \hat{\boldsymbol{\theta}}) + \sqrt{\frac{\sigma_0^2 T}{\sigma_0^2+T}}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

Since $\boldsymbol{x}_0^{\leftarrow} \sim \mathcal{N}(0, (\sigma_0^2+T)\boldsymbol{I}_d)$, we have $p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}} = \mathcal{N}\left(\hat{\boldsymbol{\theta}} - \frac{\sigma_0^2}{\sigma_0^2+T}\hat{\boldsymbol{\theta}}, \sigma_0^2\boldsymbol{I}_d\right)$. The KL-divergence between the target distribution $p_{\boldsymbol{x}|\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, \sigma_0^2\boldsymbol{I}_d)$ and $p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}}$ can then be calculated as $\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} + \frac{\sigma_0^2}{\sigma_0^2+T}\hat{\boldsymbol{\theta}}\right\|^2$. □

E.2 Proof of Lemma 4.2

Proof of Lemma 4.2. The parameterized score function is $\boldsymbol{\phi}(x; \boldsymbol{\theta}, t) = -\frac{\boldsymbol{x}-\boldsymbol{\theta}}{\sigma_0^2+t}$. Note that $\nabla_{\boldsymbol{x}_t} \ln p_{\boldsymbol{x}_t|\boldsymbol{x}_0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = -\frac{\boldsymbol{x}_t-\boldsymbol{x}_0}{t}$ since $\boldsymbol{x}_t|\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{x}_0, t\boldsymbol{I}_d)$. The training loss of (10) can then be written as

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\int_{t=0}^{T}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\frac{1}{2}\left\|\boldsymbol{\phi}(\boldsymbol{x}_{ij}+\sqrt{t}\boldsymbol{\epsilon}; \boldsymbol{\theta}_j, t) + \boldsymbol{\epsilon}/\sqrt{t}\right\|^2\right]dt + \sum_{j=1}^{m}\frac{2\xi + \|\boldsymbol{\theta}_j - \boldsymbol{\mu}\|^2}{2\sigma^2} + \frac{d}{2}\ln\sigma^2.$$

Note that

$$\begin{aligned}
&\int_{t=0}^{T}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}_d)}[\|\boldsymbol{\phi}(\boldsymbol{x}_{ij}+\sqrt{t}\boldsymbol{\epsilon}; \boldsymbol{\theta}, t) + \boldsymbol{\epsilon}/\sqrt{t}\|^2]dt \\
&= \int_{t=0}^{T}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\left\|-\frac{\boldsymbol{x}_{ij}+\sqrt{t}\boldsymbol{\epsilon}-\boldsymbol{\theta}}{\sigma_0^2+t} + \frac{\boldsymbol{\epsilon}}{\sqrt{t}}\right\|^2\right]dt \\
&= \int_{0}^{T}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\left\|\frac{\boldsymbol{\theta}-\boldsymbol{x}_{ij}}{\sigma_0^2+t} + \frac{\sigma_0^2}{\sqrt{t}(\sigma_0^2+t)}\boldsymbol{\epsilon}\right\|^2\right]dt \\
&= \left(\int_0^T \frac{1}{(\sigma_0^2+t)^2}dt\right)\|\boldsymbol{\theta}-\boldsymbol{x}_{ij}\|^2 + \text{const.}
\end{aligned}$$

Since $\int_0^T \frac{1}{(\sigma_0^2+t)^2}dt = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_0^2+T}$, the optimization of minimizing the training loss is equivalent to

$$\min_{\boldsymbol{\theta}_{1:m},\boldsymbol{\theta},\sigma^2} \quad \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\frac{\alpha}{2}\|\boldsymbol{\theta}_j - \boldsymbol{x}_{ij}\|^2 + \frac{2\xi + \|\boldsymbol{\theta}_j-\boldsymbol{\mu}\|^2}{2\sigma^2}\right) + \frac{d}{2}\ln\sigma^2,$$

where $\alpha = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_0^2+T}$. By the KKT condition that

$$\begin{aligned}
&\sum_{j=1}^{n}\alpha(\boldsymbol{\theta}_i - \boldsymbol{x}_{ij}) + \frac{\boldsymbol{\theta}_i - \boldsymbol{\mu}}{\sigma^2} = 0, \quad \forall i = 1, 2, \ldots, m, \\
&\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}_i, \\
&-\frac{1}{m}\sum_{i=1}^{m}\frac{2\xi + \|\boldsymbol{\theta}_j-\boldsymbol{\mu}\|^2}{2\sigma^4} + \frac{d}{2\sigma^2} = 0,
\end{aligned}$$

We thus have

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{x}_{ij}}{mn}$$

$$\hat{\boldsymbol{\theta}}_i = \frac{n\alpha\hat{\sigma}^2}{n\alpha\hat{\sigma}^2+1}\frac{\sum_{j=1}^n \boldsymbol{x}_{ij}}{n} + \frac{\hat{\boldsymbol{\mu}}}{n\alpha\hat{\sigma}^2+1},$$

where $\hat{\sigma}^2$ satisfies $\hat{\sigma}^2 = \frac{2\xi}{d} + s^2(\frac{n\alpha\hat{\sigma}^2}{n\alpha\hat{\sigma}^2+1})^2$ with $s^2 = \frac{\sum_{i=1}^m \left\|\hat{\boldsymbol{\mu}} - \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_{ij}\right\|^2}{md}$. □

E.3 Proof of Theorem 4.4 and Corollary 4.5

Proof of Theorem 4.4. When $m \to \infty$, $s^2 \to \frac{\sigma_0^2}{n} + \sigma_*^2$ and $\hat{\boldsymbol{\mu}} \to \boldsymbol{\mu}_*$, a.s.. $\hat{\sigma}^2$ satisfies

$$\hat{\sigma}^2 = \frac{2\xi}{d} + (\frac{\sigma_0^2}{n} + \sigma_*^2)(\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + 1/(n\alpha)})^2.$$

Since $\boldsymbol{\theta}_i$ are sampled i.i.d. from a population distribution $\mathcal{N}(\boldsymbol{\mu}_*, \sigma_*^2\boldsymbol{I}_d)$. $\alpha = 1/\sigma_0^2$ since $T \to \infty$. Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\theta}, \frac{\sigma_0^2}{n}\boldsymbol{I}_d)$, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_d)$, by Lemma 4.1, we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m D_{KL}(p_{\boldsymbol{x}|\boldsymbol{\theta}_i}||p_{\boldsymbol{x}_T^{\leftarrow}|\hat{\boldsymbol{\theta}}_i}) &= \frac{1}{m}\sum_{i=1}^m \left\|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i + \frac{\sigma_0^2}{\sigma_0^2+T}\hat{\boldsymbol{\theta}}_i\right\|^2 \\
&= \mathbb{E}\left[\left\|\boldsymbol{\theta} - \boldsymbol{x} - \frac{1}{n\alpha\hat{\sigma}^2+1}(\boldsymbol{\mu}-\boldsymbol{x})\right\|^2\right] \\
&= (\frac{\sigma_0^2/n}{\hat{\sigma}^2+\sigma_0^2/n})^2\mathbb{E}\left[\|\boldsymbol{\theta}-\boldsymbol{\mu}\|^2\right] + (\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\sigma_0^2/n})^2\mathbb{E}\left[\|\boldsymbol{\theta}-\boldsymbol{x}\|^2\right] \\
&= (\frac{\sigma_0^2/n}{\hat{\sigma}^2+\sigma_0^2/n})^2\sigma_*^2 + (\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\sigma_0^2/n})^2\sigma_0^2/n \\
&= \frac{\sigma_0^2}{n} + \frac{(\sigma_*^2+\sigma_0^2/n)\sigma_0^2/n}{(\hat{\sigma}^2+\sigma_0^2/n)^2}\frac{\sigma_0^2}{n} - 2\frac{\sigma_0^2/n}{\hat{\sigma}^2+\sigma_0^2/n}\frac{\sigma_0^2}{n} \\
&= \frac{\sigma_0^2}{n} - \left(\frac{2\hat{\sigma}^2+\sigma_0^2/n-\sigma_*^2}{\hat{\sigma}^2+\sigma_0^2/n}\right)\frac{\sigma_0^2/n}{\hat{\sigma}^2+\sigma_0^2/n}\frac{\sigma_0^2}{n}
\end{aligned}$$

where the expectation is taken w.r.t. $\boldsymbol{x}, \boldsymbol{\theta}$.

Since without collaboration, the training of maximizing $ELBO_i$ for client-i leads to parameter $\hat{\boldsymbol{\theta}}_i = \frac{\sum_{j=1}^n \boldsymbol{x}_{ij}}{n}$ and the KL-divergence between the target distribution and the output distribution is $\frac{\sigma_0^2}{n}$, it follows that collaboration improves the performance as long as $\hat{\sigma}^2 > \frac{\sigma_*^2}{2} - \frac{\sigma_0^2}{2n}$.

The improvement is $\left(\frac{2\hat{\sigma}^2+\sigma_0^2/n-\sigma_*^2}{\hat{\sigma}^2+\sigma_0^2/n}\right)\frac{\sigma_0^2/n}{\hat{\sigma}^2+\sigma_0^2/n}\frac{\sigma_0^2}{n}$ and achieves the maximum when $\hat{\sigma}^2 = \sigma_*^2$, i.e., the learned $\hat{\sigma}^2 = \sigma_*^2$ □

Proof of Corollary 4.5. Under the same setting as in Theorem 4.4, by Lemma 4.2, we have $\hat{\sigma}^2 = \frac{2\xi}{d} + (\frac{\sigma_0^2}{n} + \sigma_*^2)(\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\sigma_0^2/n})^2$. Taking $\xi > \frac{2\xi}{d} = \frac{3d\sigma_0^2}{2n}$ gives that $\hat{\sigma}^2 \geq \frac{3\sigma_0^2}{n}$, and thus $\hat{\sigma}^2 > (\frac{\sigma_0^2}{n} + \sigma_*^2)(\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\sigma_0^2/n})^2 \geq \frac{9}{16}(\frac{\sigma_0^2}{n} + \sigma_*^2) > \frac{\sigma_*^2}{2} - \frac{\sigma_0^2}{2n}$, which guarantees strictly improvement by Theorem 4.4. □

F Additional Experiments and Experimental Details

F.1 Experiments for ADEPT-PCA on synthetic data

We use synthetic datasets for the experiments of ADEPT-PCA. In the dataset, we first sample a global PC, $\boldsymbol{V}^* \in St(d,r)$, uniformly on the Steifel manifold. We sample $\{\hat{\boldsymbol{U}}_i^*\}_{i=1}^m$ where the entries of each $\hat{\boldsymbol{U}}_i^*$ follows Gaussian distribution with mean being $\boldsymbol{V}^*$ and variance σ^*. Then, we let $\boldsymbol{U}_i^* = \mathcal{R}_{\boldsymbol{V}^*}(P_{\mathcal{T}_{\boldsymbol{V}^*}}(\hat{\boldsymbol{U}}_i^*))$ so that it is in the Steifel manifold. Data on each client are then generated by $\boldsymbol{x} = \boldsymbol{U}_i^*\boldsymbol{z} + \boldsymbol{\epsilon}$.

We compare the reconstruction error between Algorithm 2, local training, and global training. In the global training setting, we train a single global model with the average of local gradients in each iteration. In Figure 3,

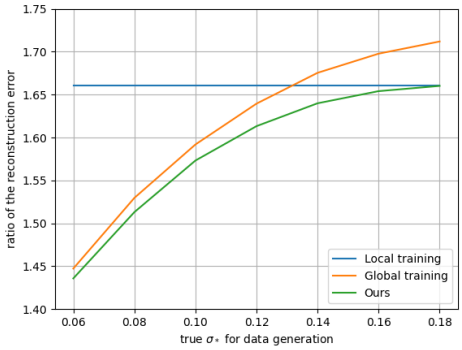

Figure 3: Ratio of the reconstruction error to the optimal error for different values of σ^*. We have $d = 100$, $r = 20$, $m = 10$, and $n = 20$.

the value in the y-axis is the ratio of the reconstruction error of each training method to the true error, which is evaluated by $\{\boldsymbol{U}_i^*\}_{i=1}^m$. When σ^* is small, the heterogeneity of the data among the clients is low, and thus global training benefits from the sample size and performs better than pure local training. Our algorithm also makes use of the sample size and achieves an even smaller reconstruction error with the personalized models. When σ^* is large, the heterogeneity of the data among the clients is high and thus training a single global model for each client does not work well. In this case, our algorithm learns a larger σ and performs more like local training. In the scenario between the two cases, our algorithm also outperforms both global and local training.

F.2 Training and hyper-parameter Details

ADEPT-AE. We set $\xi = 1e{-}6$ for all experiments. For the synthetic experiments, we don't apply local iterations and just do distributed training. For the other experiments, we do 20 local iterations per communication round, and each communication round corresponds to an epoch. We use 300 global batch size for MNIST and 750 for CIFAR-10. For all datasets and methods, we use SGD with a constant learning rate of 0.01 after individually tuning in the set $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$ and momentum coefficient of 0.9. For our method, we choose $\eta_2 = 0.01, \eta_3 = 0.001$ and use SGD without momentum. For MNIST and Fashion MNIST datasets, we train for 150 epochs/comm. rounds and for CIFAR-10 for 250 epochs. We initialize $\sigma = 1$ in MNIST and Fashion MNIST experiments, and $\sigma = 0.4$ in synthetic experiments, and do not update σ for the first two epochs. For CIFAR-10 we initialize $\sigma = 0.2$ and we do lazy updates that is we start updates after 200 epochs. We observed lazy updates with relatively small initial σ works better for deeper models, whereas simpler models do not require it. For pFedMe, we set the number of local optimization steps to 3 and use a $\lambda =$

For AE experiments we use simpler models (compared to Diffusion Model experiments) as AEs tend to overfit the data and deeper models do not necessarily provide a better performance in general. For synthetic experiments, we use a two-layer fully connected AE, and for MNIST and Fashion MNIST we also use a two-layer AE with 784 input dimension with 10 and 20 latent dimensions depending on the experiment. For CIFAR-10 we use a symmetric convolutional AE whose input and output layers are convolutional layers with 16 channels, 3 kernel size, 2 strides, and no padding; the intermediate layers are fully connected layers that map 3600 dimensions to latent dimensions. We use 10 and 50 latent dimensions depending on the experiment. We use ReLU activation function after the first layer and sigmoid after the last layer. To improve the empirical performance and have a more stable training we make a few changes to Algorithm 1. Namely, we keep individual σ for each scalar weight, for personalized and global models we clip the ℓ_∞ norm of the gradients by 1, and for σ by 10. We update σ at the first iteration instead of the last one. We also update the global model locally in each local iteration.

ADEPT-DGM. On MNIST and Fashion MNIST, we use 20 local iterations per communication round and epoch during training. We use Adam optimizer with $1e{-}3$ for all methods. For our method, we use Adam with 0.01 learning rate for the updates of the global model and SGD with 0.001 lr for σ. We do 100 epochs/comm. rounds in total. We initialize $\sigma = 0.8$ and do not update it for the first 2 epochs. We multiply the learning rates η_1, η_2 by 0.1 at the 75th epoch. We employ the same changes in Algorithm 1 and Algorithm 3 as well. For demonstration, we use a variance preserving SDE (as in Algorithm 3) instead of variance exploding (as in Section 4). On MNIST and Fashion MNIST, the models are trained to predict images instead of noise. On MNIST we use a cosine noise

scheduler and on CIFAR-10 we use a linear noise scheduler. On CIFAR-10, for the results in Table 3, we train the model using 50 local iterations per communication round. We use Adam optimizer with $1e{-}4$ for all methods, and we use Adam with $1e{-}4$ learning rate for the updates of the global model and SGD with $1e{-}4$ lr for σ and we initialize $\sigma = 0.35$. We train for 200 comm. rounds. For the results in Table 3 we train for 100 comm. rounds in Table 6 and let each client access samples from 3 classes.

F.3 Additional Experiments

Convergence plots In both Figures 4 and 5 we observe that our theoretical findings from Section 3 hold; that is, high variance imply faster convergence albeit it can result in inferior testing performance.

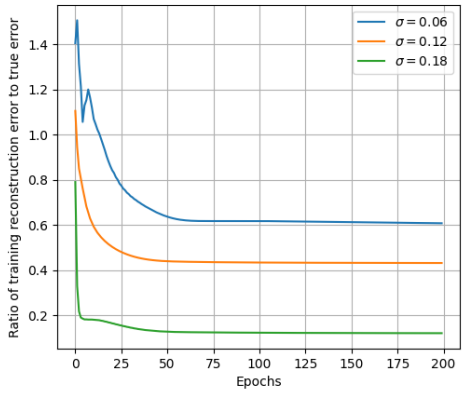

Figure 4: Training loss vs epochs for `ADEPT-PCA` with a different true standard deviation of data.

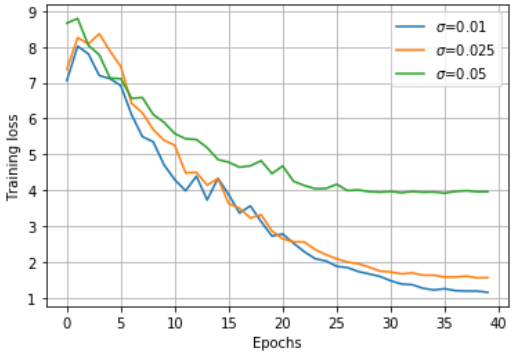

Figure 5: Training loss vs epochs for `ADEPT-AE` with a different true standard deviation of data.

Sensitivity to ξ As seen in Table 4, the sensitivity to ξ is quite low in our experiments.

Additional ADEPT-DGM results In Table 5, compared to the results in Table 3, we observe similar trends across the performance comparison of methods.

For CIFAR-10 when we have 50 comm. rounds and samples from 3 classes per client; we observe a significant performance difference between our method and competition; indicating faster convergence of our method in terms of image quality.

For F. MNIST, we look at a setting where each client has 200 samples from 2 classes; despite the low number of samples per client, our method is able to outperform other methods. We note that on F. MNIST, it is particularly hard to outperform local training due to less diverse training and test samples.

Table 4: Energy captured (%) versus hyper-prior parameter for the setting of F. MNIST experiments with high latent dimensionality.

Value of ξ	KID
$1e{-}5$	85.6
$1e{-}6$	85.6
$1e{-}7$	85.5
$1e{-}8$	85.7
$1e{-}9$	85.8
$1e{-}10$	85.9

Table 5: Diffusion model generation quality for generating MNIST samples when 1200 samples per client are available.

Method	KID
Baseline	0.062 ± 0.003
`ADEPT-DGM`	$\mathbf{0.067} \pm 0.003$
FedAvg+fine-tuning	0.082 ± 0.004
Local Training	0.075 ± 0.001

Table 6: Diffusion model generation quality for generating CIFAR-10 samples (lower is better).

Method	KID
`ADEPT-DGM`	$\mathbf{0.086} \pm 0.003$
FedAvg+fine-tuning	0.104 ± 0.005
Local Training	0.111 ± 0.006

Table 7: Diffusion model generation quality for generating F. MNIST samples (lower is better). 200 samples and 2 classes accessed per client, $m = 30$.

Method	KID
`ADEPT-DGM`	$\mathbf{0.110} \pm 0.008$
FedAvg+fine-tuning	0.118 ± 0.010
Local Training	0.119 ± 0.005

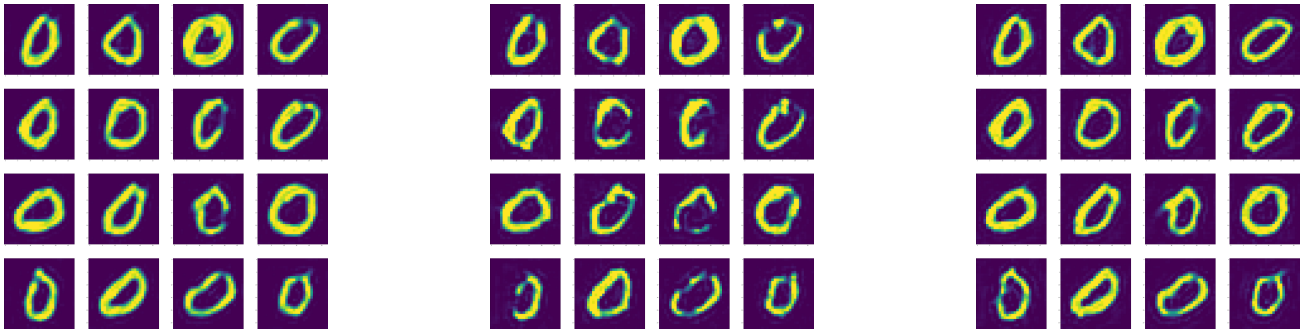

Figure 6: Randomly chosen samples (Left:`ADEPT-DGM`, noise $\sigma = 0.024$; Middle:FedAvg+fine-tuning, noise $\sigma = 0.028$; Right: Local training, noise $\sigma = 0.032$) (models are trained and samples are chosen with the same seed across runs) from generated dataset for a client with data from '0' class. The pictures are also included in the supplementary material.

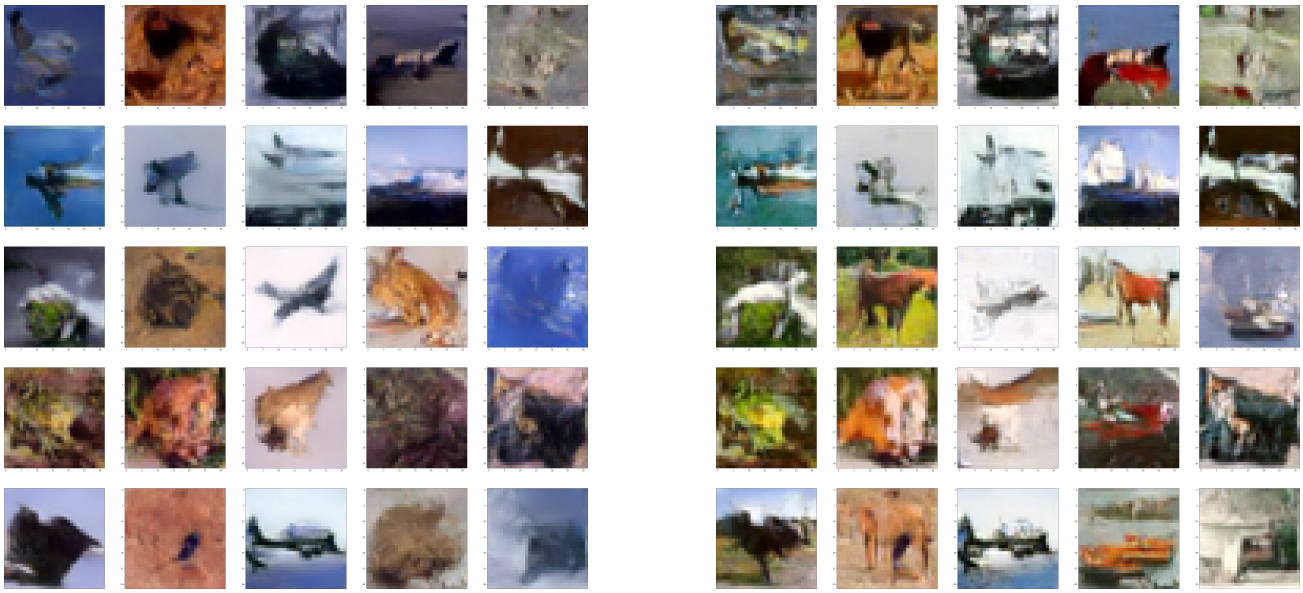

Figure 7: Randomly chosen samples (Left:`ADEPT-DGM`; Right: FedAvg+fine-tuning (models are trained and samples are chosen with the same seed across runs) from the generated dataset for a client with data from 'frog' and 'airplane' class. We observe that FedAvg+fine-tuning hallucinates images from other classes/clients (e.g. horses); which is both a privacy concern and indicates the lack of performance in generating images from the target distribution.