Factor Augmented High-Dimensional SGD

Shubo Li
Department of Statistics
The Pennsylvania State University
University Park, PA 16802, USA
skl6034@psu.edu

Yuefeng Han **Xiufan Yu**
Department of Applied and Computational
Mathematics and Statistics
University of Notre Dame
Notre Dame, IN 46556, USA
{yuefeng.han,xiufan.yu}@nd.edu

Abstract

Stochastic gradient descent (SGD) is a fundamental optimization algorithm widely used in modern machine learning. In this paper, we propose *Factor-Augmented SGD* (FSGD), a new optimization method that leverages latent factor representations in high-dimensional learning tasks. Unlike standard two-stage dimension reduction approaches that rely on offline representation learning and full data storage, a key novelty of FSGD is that it operates in a single pass over streaming data, with an online Oja warm-up before downstream SGD begins, making it scalable to large-scale and high-dimensional problems. Furthermore, we establish a new theoretical framework that explicitly incorporates latent factor estimation error into the analysis of SGD, and provide moment convergence in ℓ^s norm under decaying step sizes and mini-batch updates. Our results provide a new foundation for employing SGD reliably and scalably in high-dimensional machine learning systems.

1 Introduction

Stochastic gradient descent (SGD) has been a cornerstone of machine learning since the pioneering work of Robbins & Monro (1951). Beyond its algorithmic simplicity and scalability, SGD has also become a central object of theoretical study, with refined analyses linking its dynamics to implicit regularization, generalization performance, and algorithmic stability. For decades, theoretical analyses of SGD have largely resided within the realm of classical stochastic approximation (Polyak & Juditsky, 1992; Lai, 2003; Bottou et al., 2018), where the data dimension is considered fixed while the sample size tends to infinity. While this regime has yielded foundational insights, it no longer fully reflects the characteristics of modern learning systems. Contemporary applications often operate in regimes where data dimension, sample size, and model complexity grow together, calling for new theoretical tools and perspectives that go beyond traditional asymptotic analyses.

In this study, we focus on the learning tasks involving high-dimensional predictors. When SGD is applied directly to such data, the dimensionality of the feature space propagates into the optimization process, resulting in a high-dimensional (HD) parameter space. Algorithmically, one trending strategy is to approximate the gradient updates using a low-rank representation to reduce memory costs and accelerate computation (Wang et al., 2018; Vogels et al., 2019; Kozak et al., 2019; Kasiviswanathan, 2021; Zhao et al., 2024). Theoretically, despite the vast literature on SGD, convergence guarantees of HD-SGD remain limited (Garrigos & Gower, 2023; Li et al., 2025).

Despite recent progress, the behavior of SGD in HD regimes remains only partially understood, and its direct application to high-dimensional datasets can be empirically fragile. Constructing models directly in high-dimensional feature spaces therefore leads to excessive parameter complexity and exacerbates the curse of dimensionality (Ben Arous et al., 2021; Schmidt-Hieber, 2020). Moreover, models defined on high-dimensional feature spaces often involve a large number of parameters.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Applying SGD in such high-dimensional parameter regimes is computationally costly and is known to exhibit unstable or oscillatory behavior around flat minima (Hoffer et al., 2017; Wu et al., 2018; Cohen et al., 2021).

To address the challenges induced by HD predictors, instead of directly optimizing HD-SGD in the high-dimensional feature space, we propose a new strategy, named **Factor-Augmented Stochastic Gradient Descent (FSGD)**, that augments SGD with a factor model. The key insight is that the effective information content of HD predictors often lies in a much lower-dimensional latent space, which can be captured by the "factors". Ignoring this structure leads to inefficient updates that waste computation on noisy directions, slow convergence along dominant subspaces, and obscure the true geometry of the optimization landscape. By modeling predictors dynamically through a factor decomposition, we can extract useful representations from predictors and design SGD updates that adapt to the intrinsic dimensionality of the problem.

Integrating factor models into SGD is far from a straightforward adaptation, but involves several non-trivial challenges to be carefully addressed.

First, existing factor-augmented learning approaches (Fan & Gu, 2024; Zhou et al., 2025a,b; Luo et al., 2025, 2026) in other applications typically adopt a two-stage paradigm, where a low-dimensional representation is first learned using an offline procedure, and then treated as fixed during downstream learning tasks. Such two-stage approaches rely on full access to the dataset and require storing or repeatedly scanning all samples, which is inefficient for iterative SGD updates and may even be infeasible in large-scale learning settings where data arrive sequentially and memory resources are limited. To tackle this challenge, we propose an online factor-learning procedure designed for a one-pass streaming setting and seamlessly integrate it with SGD updates, enabling memory-efficient and scalable learning.

Second, a key advantage of factor models is that they explicitly allow for the presence of idiosyncratic errors, capturing variation in predictors that is not explained by a small number of common factors. Most existing works in representation learning literature focus exclusively on learning low-dimensional embeddings and implicitly absorb idiosyncratic components into the representation itself, ignoring both the error structure and the associated estimation uncertainty. There is limited understanding of how idiosyncratic noise and representation estimation error interact with stochastic optimization procedures, particularly when SGD is employed. Our work bridges this gap by rigorously quantifying how representation estimation error and idiosyncratic noise propagate into the convergence rate.

These limitations raise a fundamental question:

Can the latent factor structure of high-dimensional predictors be explicitly utilized to improve the efficiency and theoretical understanding of SGD in high dimensions?

In this work, we provide a positive answer to the above question. We summarize our main contributions as follows.

1. **Factor-Augmented SGD (FSGD).** We propose a new optimization method that utilizes the latent structure of high-dimensional predictors. Specifically, our method learns a low-dimensional latent representation via online PCA and simultaneously updates the model parameters using SGD on the induced factor space. While our theory is developed for strongly convex population objectives, the neural network experiments illustrate the empirical robustness of the FSGD principle beyond the convex setting.
2. **Streaming unsupervised factor learning coupled with SGD.** We develop a streaming learning framework in which both unsupervised latent factors and model parameters are updated using SGD-type procedures after an online Oja warm-up. Unlike classical two-stage representation learning approaches that require an offline PCA step with full data access, FSGD operates on streaming data and does not store the entire dataset, crucial for practical large-scale learning scenarios. As a result, factor estimates can be continuously refined during model training rather than being fixed a priori, causing the population minimizer to shift over time. In Proposition 3.5, we establish a novel bound for minimizer shift to address the nontrivial statistical dependence between factor estimation and gradient updates.
3. **New theoretical guarantee for SGD with latent factor estimation.** Theorem 3.6 provides a new q-th moment convergence analysis of SGD that accounts for both latent factor estimation error

and idiosyncratic noise in high-dimensional settings. Unlike existing approaches that treat the learned representation as fixed or known and ignore the idiosyncratic component in the factor structure, we rigorously quantify how the representation estimation error and idiosyncratic noise propagate into the convergence rate of SGD.

1.1 Related Works

HD-SGD. While the theory of SGD has been extensively studied in traditional low-dimensional settings (Bottou et al., 2018; Garrigos & Gower, 2023), theoretical understanding of HD-SGD remains underdeveloped, with only a handful of recent works. Paquette et al. (2021, 2022, 2025) study HD-SGD under quadratic loss functions. Li et al. (2025) studies HD-SGD and its Ruppert-Polyak averaged variant, and provides guarantees on moment convergence and high-probability concentration bounds in HD regimes with constant learning rates. Our work provides a rigorous characterization of convergence by explicitly accounting for data dimensionality via factor models, a completely different analytical route from existing studies, offering a fresh theoretical perspective on SGD with HD predictors. Methodologically, we also emphasize that the proposed FSGD fundamentally differs from the SGD with low-rank gradient approximation (Vogels et al., 2019; Kozak et al., 2019; Kasiviswanathan, 2021): FSGD is a data-side approach that utilizes the latent structure of high-dimensional predictors, whereas the latter is an optimization-side method that enforces low-rank structure on gradient updates.

Factor Models. Factor models (Lawley & Maxwell, 1962; Bai, 2003; Bai & Ng, 2008; Han et al., 2024a,b,c; Han & Zhang, 2023; Han et al., 2022; Yu et al., 2024a,b; Bolivar et al., 2025; Chen et al., 2025a, 2026b) have long served as a foundational tool to capture the underlying structure of high-dimensional data through a smaller set of latent variables (a.k.a., factors). In time series forecasting literature, they are often employed as a first step to extract low-dimensional representations from multivariate inputs, which are then fed into subsequent learning algorithms for prediction (Stock & Watson, 2002a,b; Bair et al., 2006; Fan et al., 2017; Luo et al., 2022; Yu et al., 2022; Chen et al., 2026a). We emphasize that FSGD fundamentally differs from these approaches: existing factor-augmented learning approaches adopt a straightforward two-stage paradigm, while FSGD involves an online factor-learning step that better aligns with the data-adaptive and stochastic nature of SGD updates.

Notation For $v \in \mathbb{R}^d$, we denote the l^s-norm of v by $|v|_s := (\sum_{i=1}^d |v_i|^s)^{1/s}$. Write $v^{\odot s} = (v_1^s, v_2^s, \cdots, v_d^s)^\top$. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\|A\|_F$ denote its Frobenius norm, and $\|A\|_{\text{op}}$ denote its operator norm. We write $\mathbb{O}_{d,k}$ to be all $d \times k$ orthonormal matrix and denote $\mathbb{O}_k = \mathbb{O}_{k,k}$. For any matrices with orthonormal columns $V, W \in \mathbb{R}^{d\times k}$, we define the subspace distance

$$\text{dist}(W, V) := \|VV^\top - WW^\top\|_{\text{op}}. \tag{1}$$

The expectation and covariance of random vectors are respectively denoted by $\mathbb{E}[\cdot]$ and $\text{Cov}(\cdot)$. For $q > 0$ and a random variable X, we write L^q-norm of X by $\|X\|_q := (\mathbb{E}[|X|^q])^{1/q}$, and $X \in L^q$ iff $\|X\|_q < \infty$. For random variable $X \in \mathbb{R}$, we define sub-Gaussian norm as $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \le 2\}$. For random vector $Z \in \mathbb{R}^d$, we define $\|Z\|_{\psi_2} := \sup_{|u|_2=1} \|\langle u, Z\rangle\|_{\psi_2}$.

For two positive number sequences (a_n) and (b_n), we say $a_n = \mathcal{O}(b_n)$ (resp. $a_n = \Theta(b_n)$) if there exists $c > 0$, such that $a_n/b_n \le c$ (resp. $1/c \le a_n/b_n \le c$) for all large n. We call c an absolute constant if it is independent of all problem parameters, including n, d, t, and other model-specific quantities. Throughout the paper, C denotes an absolute constant, unless stated otherwise. A summary of the main notation is provided in Appendix A.

2 Factor Augmented SGD

2.1 Factor Augmented Regression Model

High-dimensional learning poses both statistical and computational challenges. Statistically, when learning a (β, C)-smooth function, deep neural networks achieve the optimal nonparametric rate of $n^{-\frac{2\beta}{2\beta+d}}$ with sample size n, which deteriorates rapidly as the ambient dimension d grows. To address this challenge, we adopt a factor model for the covariate process that exploits the latent low-dimensional structure underlying the observed data $\{x_i, y_i\}$. Specifically, we assume

$$y_i = \mathcal{M}(f_i) + \varepsilon_i, \text{ and } x_i = Bf_i + u_i, \tag{2}$$

where f_i, u_i, ε_i are mutually independent random vectors, $B \in \mathbb{R}^{d\times k}$ is an unknown factor loading matrix, $f_i \in \mathbb{R}^k$ is a vector of latent factors, $u_i \in \mathbb{R}^d$ is the idiosyncratic component, and $\mathcal{M}$ is the regression function. Under this model, the regression function $\mathcal{M}$ is defined on the k-dimensional latent factor space, with a corresponding oracle nonparametric rate of $n^{-\frac{2\beta}{2\beta+k}}$ when the factors are observed. Algorithmically, building the downstream model on this low-dimensional space allows SGD to optimize a model with fewer parameters, while learning and applying the projection requires a cost linear in d for fixed m and k. We denote $\Sigma := \mathrm{Cov}(X),\ \Sigma_f := \mathrm{Cov}(f)$, and let $V \in \mathbb{R}^{d\times k}$ be the orthogonal matrix whose columns are the leading k eigenvectors of Σ. The following regularity conditions on the factor model in (2) are standard in the literature (Bai, 2003; Bai & Li, 2012).

Assumption 2.1. The idiosyncratic component $u_i \in \mathbb{R}^d$ is a zero-mean sub-Gaussian random vector with ψ_2-Orlicz norm $\|u_i\|_{\psi_2} \le K_u$. The loading matrix satisfies $d^{-1}B^\top B = I_k$, and $\mathbb{E}[f_i] = 0$. Moreover, the minimum eigenvalue of Σ_f is strictly positive, i.e $\lambda_k(\Sigma_f) > 0$.

The following proposition establishes the convergence rate of the estimated factors.

Proposition 2.2 (Factor Estimation)**.** *Suppose the factor model satisfies* (2) *and Assumption 2.1. Given $Q \in \mathbb{R}^{d\times k}$ with orthonormal columns, define the factor estimator for a new observation x as $\hat{f} := d^{-1/2}Q^\top x$. Then there exists a rotation matrix $R \in \mathbb{R}^{k\times k}$, for $q \ge 2$, such that*

$$\big\| \, |\hat{f} - Rf|_s \, \big\|_q \le \sqrt{2}k^{\frac{1}{2}} \, \|\operatorname{dist}(Q,V)\|_q \big\| \, |f|_2 \, \big\|_q + C\Big(\frac{K_u^2}{\lambda_k(\Sigma_f)} \big\| |f|_2 \, \big\|_q + K_u(k^{\frac{1}{2}} + q^{\frac{1}{2}})\Big) d^{-\frac{1}{2}}. \tag{3}$$

The first term in (3) captures the effect of the subspace distance between Q and V, while the second term accounts for both the subspace distance between V and $d^{-1/2}B$, and the contribution of the idiosyncratic components. The sub-Gaussian condition on u_i bounds the noise covariance and the projected idiosyncratic noise, thereby controlling the second term. A smaller $\lambda_k(\Sigma_f)$ enlarges the bound on the subspace distance between V and $d^{-1/2}B$, while heavier-tailed u_i may worsen both bounds in the second term. When Q is obtained via PCA and $\mathrm{Cov}(f)$ has distinct eigenvalues, the factors f are identifiable up to sign flips; equivalently, the indeterminacy reduces to $R \in \mathrm{diag}(\{1,-1\}^k)$.

2.2 Online PCA

Principal component analysis (PCA) is widely used to estimate latent factors in high-dimensional models. However, classical PCA becomes computationally infeasible when both the sample size n and the ambient dimension d are large, or when data arrive in a streaming fashion. To address these challenges, we adopt a streaming PCA approach for online factor estimation.

Let $\{A_t\}_{t\ge1} \subset \mathbb{R}^{d\times d}$ be a sequence of i.i.d. symmetric random matrices with $\mathbb{E}[A_t] = \Sigma$, and let $\{\eta_t^{\mathrm{oja}}\}_{t\ge1}$ denote the step sizes. Initialize with any $Z_0 \in \mathbb{R}^{d\times k}$ of full column rank and set $Q_0 = \mathrm{QR}[Z_0]$, where $\mathrm{QR}[\cdot]$ denotes the Q-factor of the thin QR decomposition. For $t \ge 1$, Oja's algorithm (Oja, 1982; Oja & Karhunen, 1985) updates

$$Q_t := \mathrm{QR}[(I + \eta_t^{\mathrm{oja}} A_t)\, Q_{t-1}]. \tag{4}$$

Let $V \in \mathbb{R}^{d\times k}$ denote the matrix whose columns are the k leading eigenvectors of Σ. We impose the following conditions on Oja's updates throughout this paper.

Assumption 2.3. For Oja's update, we assume

1. There exists a constant $M > 0$ such that,
$$\sup_{P\in\mathbb{O}_{d,k}} \|P^\top(A_t - \Sigma)\|_F = \Big(\sum_{i=1}^k \sigma_i^2(A_t - \Sigma)\Big)^{\frac{1}{2}} \le M$$
holds almost surely, where $\sigma_1(X) \ge \cdots \ge \sigma_d(X)$ denote the singular values of X.
2. (Warm start) The initialization Q_0 in stage-II satisfies $\mathrm{dist}(Q_0, V) \le \frac{1}{2}$.

The condition $\sup_{P\in\mathbb{O}_{d,k}} \|P^\top(A_t - \Sigma)\|_F \le M$ a.s. is standard in the analysis of online PCA; see, e.g., Jain et al. (2016); Allen-Zhu & Li (2017); Huang et al. (2021); Liu et al. (2022). Under the factor model in (2) and Assumption 2.1, with $A_t = m^{-1}\sum_{i=1}^m x_i x_i^\top$ and sub-Gaussian (f_i, u_i), one may take $M = \mathcal{O}(m^{-1/2} d \log d)$ a.s.; see Remark 3.4 of Liu et al. (2022) for further discussion. The

warm-start condition in Assumption 2.3(ii) is also standard and requires the initial estimate to be sufficiently close to the population principal subspace. In particular, it is guaranteed by Theorem 3 of Huang et al. (2021) after a burn-in period. This condition is used later in Proposition 3.5 to control how subspace updates affect the population minimizer.

Adapted from Huang et al. (2021), the following lemma provides an estimation error bound for Oja's algorithm in terms of the subspace distance between the estimated and true principal subspaces.

Lemma 2.4 (L^q-Moment Bound for Oja Update)**.** *Suppose Assumption 2.3 holds. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of Σ and define the eigengap $\rho_k = \lambda_k - \lambda_{k+1}$. Set $\eta_t^{\mathrm{oja}} = \frac{\alpha}{(\beta+t)\rho_k}$ and define the normalized gap $\bar{\rho}_k := \min\left\{\frac{\rho_k}{M}, \frac{\rho_k}{\|\Sigma\|_{\mathrm{op}}}, 1\right\}$. If*

$$\alpha \geq 8, \qquad \beta \geq \max\Big\{\frac{4(1+\sqrt{2e})\alpha}{\bar{\rho}_k},\, C\cdot\frac{\alpha^2}{\bar{\rho}_k^2}\Big\},$$

where C is an absolute constant. Then for every $q \geq 2$ and $t \geq 1$, it holds that

$$\|\mathrm{dist}(Q_t, V)\|_q \leq k^{1/q}\left[\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} + C\cdot\frac{\alpha q^{1/2}}{\bar{\rho}_k}\cdot\frac{t^{1/2}}{\beta+t} + C^{2/q}\cdot\alpha^{2/q}\bar{\rho}_k^{-2/q}\exp\left\{-\frac{\beta\bar{\rho}_k^2}{\alpha^2 C^2 q}\right\}\right].$$

2.3 Factor Augmented SGD

When the latent factor structure in (2) is present, it is natural to consider a factor-augmented stochastic gradient descent (FSGD) framework in which the loss function is defined on the latent factors rather than the high-dimensional covariates. Specifically, given factor representations $f_{t,i}$ and responses $y_{t,i}$, we consider a loss function of the form $l(\theta; f_{t,i}, y_{t,i})$. The optimization problem of interest is

$$\theta^* = \arg\min_{\theta} L(\theta; I), \text{ where } L(\theta; R) = \mathbb{E}l(\theta; Rf_{t,i}, y_{t,i}), \tag{5}$$

$\theta \in \mathbb{R}^p$, $p \ll d$, and $R \in \mathbb{O}_k$ denoting a rotation matrix. The rotation matrix R accounts for the rotational ambiguity inherent in factor models; setting $R = I$ corresponds to the ideal case where the factors are perfectly identified. At iteration t, given a mini-batch $B_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^m$ of size m, the FSGD update takes the form

$$\theta_{t+1} = \theta_t - \eta_t \cdot \frac{1}{m}\sum_{i=1}^m \nabla_\theta l(\theta_t; \hat{f}_{t,i}, y_{t,i}), \tag{6}$$

where $\hat{f}_{t,i}$ denotes an estimated latent factor obtained from the observation $x_{t,i}$. Since the true factors $f_{t,i}$ are not directly observed, they must be estimated online from the streaming data via a dimension reduction procedure.

The FSGD proceeds in two stages, detailed in Algorithm 1.

Stage I: Initialization. The goal of the first stage is to obtain a good initialization satisfying the warm-start condition in Assumption 2.3. One approach is to initialize Z_0 with i.i.d. Gaussian entries and run Oja's algorithm for a sufficient burn-in period. During this burn-in, mini-batches are processed sequentially without storing or replaying past data. When offline initialization is permitted, one may perform offline PCA on a preliminary dataset to estimate the loading matrix, as studied in Yang et al. (2025), and use the resulting subspace estimator to initialize Q_0 in the second stage.

Stage II: Joint Optimization. In this stage, at each iteration t, given the subspace estimate $Q_{t-1} \in \mathbb{R}^{d\times k}$ obtained from the previous $t-1$ mini-batches, we perform linear dimension reduction to extract $k \ll d$ latent factors by projecting each observation onto the estimated factor space:

$$\hat{f}_{t,i} = d^{-\frac{1}{2}} Q_{t-1}^\top x_{t,i} \in \mathbb{R}^k,$$

where $x_{t,i}$ denotes the i-th sample in the mini-batch at iteration t. The stochastic gradient is then computed as $\nabla_\theta l(\theta; \hat{f}_{t,i}, y_{t,i})$. To fully utilize the data, the same mini-batch B_t is reused to perform the Oja update in Step 15 of Algorithm 1, yielding an updated subspace estimate Q_t. We note that each Oja update costs $O(mdk + dk^2)$ time. Specifically, computing $X_t^\top Q_{t-1}$ and then $X_t(X_t^\top Q_{t-1})$ costs $O(mdk)$ in total, while the thin QR decomposition costs $O(dk^2)$. Storing the subspace estimate Q_t requires $O(dk)$ memory.

Algorithm 1 Factor Augmented SGD (FSGD)

Require: Step sizes $\{\eta_t, \eta^{\mathrm{oja},(0)}, \eta_t^{\mathrm{oja}}\}$, factor dimension k, mini-batch size m

Stage I: Online PCA Warm-up

Initialize $Z_0 \in \mathbb{R}^{d\times k}$ of full column rank and set $Q_0^{(0)} = QR[Z_0]$.
for $s = 1, 2, \ldots, T_0$ **do**
  Sample an independent mini-batch $B_s^{(0)} = \{x_{s,i}^{(0)}\}_{i=1}^m$ from model (2).
  Form the mini-batch matrix $X_s^{(0)} = [x_{s,1}^{(0)}, \ldots, x_{s,m}^{(0)}] \in \mathbb{R}^{d\times m}$.
  Update subspace $Q_s^{(0)} \leftarrow QR\Big[Q_{s-1}^{(0)} + \frac{\eta^{\mathrm{oja},(0)}}{m} X_s^{(0)} \Big((X_s^{(0)})^\top Q_{s-1}^{(0)}\Big)\Big]$.
end for
Set $Q_0 \leftarrow Q_{T_0}^{(0)}$.

Stage II: Joint Online PCA and Mini-batch SGD

Initialize θ_0
for $t = 1, 2, \ldots, T$ **do**
  Mini-batch SGD:
  Sample an independent mini-batch $B_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^m$ from model (2).
  Estimate factors: $\hat{f}_{t,i} \leftarrow d^{-1/2} Q_{t-1}^\top x_{t,i}$ for all $(x_{t,i}, y_{t,i}) \in B_t$.
  Compute stochastic gradient: $g_t \leftarrow \frac{1}{m}\sum_{i=1}^m \nabla_\theta l\big(\theta_t; \hat{f}_{t,i}, y_{t,i}\big)$.
  Update parameters: $\theta_{t+1} \leftarrow \theta_t - \eta_t\, g_t$.
  Online PCA:
  Form the mini-batch matrix $X_t = [x_{t,1}, \ldots, x_{t,m}] \in \mathbb{R}^{d\times m}$.
  Update subspace $Q_t \leftarrow QR\Big[Q_{t-1} + \frac{\eta_{t-1}^{\mathrm{oja}}}{m} X_t \big(X_t^\top Q_{t-1}\big)\Big]$.
  (Optional) Procrustes alignment $S_t = Q_t^\top Q_{t-1}$, compute SVD $S_t = U_{S_t}\Lambda_{S_t} V_{S_t}^\top$, and then set $Q_t \leftarrow Q_t U_{S_t} V_{S_t}^\top$.
end for

It is important to note that recovery of the factor loading matrix is subject to the intrinsic rotational ambiguity of factor models. Specifically, given the orthogonal projection matrix Q_{t-1} at iteration t, there exists an orthogonal rotation matrix $R_{t-1} \in \mathbb{O}^k$ (depending on Q_{t-1}) such that the estimated factor $\hat{f}_{t,i} = d^{-\frac{1}{2}} Q_{t-1}^\top x_{t,i}$ is close to the rotated true factor $R_{t-1} f_{t,i}$; see Proposition 2.2 for a formal statement. In offline PCA, if the leading k eigenvalues of the covariance matrix are distinct, then the factor loading matrix is uniquely determined up to sign changes. In our online PCA setting, we could perform a Procrustes alignment of Q_t to Q_{t-1} at each iteration. When the leading k eigenvalues of Σ are distinct and the subspace error of Q_{t-1} is sufficiently small, the associated Procrustes rotation R_t varies smoothly across iterations, i.e., remaining close to R_{t-1}. In Appendix D.3, we show that this additional alignment step does not affect the theoretical bounds.

3 Convergence of FSGD

In this section, we establish the moment convergence property of FSGD. Recall the optimization problem defined in (5) and the data (x, f, y) from model (2). We begin by imposing the following assumptions on the objective function and the stochastic gradients.

Assumption 3.1 (ℓ^s norm - Strong Convexity)**.** Let $s \geq 2$ be an even integer. For any $R \in \mathbb{O}_k$, assume there exists a constant $\mu > 0$ such that

$$\Big\langle (\theta_1 - \theta_2)^{\odot(s-1)}, \nabla L(\theta_1; R) - \nabla L(\theta_2; R) \Big\rangle \geq \mu |\theta_1 - \theta_2|_s^s,$$

holds for all $\theta_1, \theta_2 \in \mathbb{R}^p$.

A similar condition is adopted in Li et al. (2025). When $s = 2$, this reduces to standard strong convexity. For example, consider the least squares loss under the linear model $y_i = f_i^\top \theta + \varepsilon_i$. The Hessian of the population loss is given by

$$\nabla_\theta^2 L(\theta; R) = R\Sigma_f R^\top, \quad \text{where } \Sigma_f := \mathbb{E}[f_i f_i^\top].$$

Thus, a sufficient condition for strong convexity is $\lambda_{\min}(\Sigma_f) > 0$.

Assumption 3.2 (ℓ^s norm - Lipschitz Continuity for θ)**.** Let $s \geq 2$ be an even integer and $q \geq 2$. For (x, f, y) from model (2), assume there exists a constant $L_{s,\theta} > 0$ such that

$$\left\| |\nabla_\theta \, l(\theta_1; Rf, y) - \nabla_\theta \, l(\theta_2; Rf, y)|_s \right\|_q \leq L_{s,\theta} \cdot \left\| |\theta_1 - \theta_2|_s \right\|_q,$$
$$\left\| |\nabla_\theta \, l(\theta_1; d^{-1/2} Q^\top x, y) - \nabla_\theta \, l(\theta_2; d^{-1/2} Q^\top x, y)|_s \right\|_q \leq L_{s,\theta} \cdot \left\| |\theta_1 - \theta_2|_s \right\|_q,$$

hold for all $R \in \mathbb{O}_k, Q \in \mathbb{O}_{d,k}$.

Assumption 3.3 (Lipschitz Continuity for f)**.** Let $s \geq 2$ be an even integer and $q \geq 2$. For random vectors y, f_1, f_2 from model (2), there exists a constant $L_{s,f} > 0$ such that

$$\left\| |\nabla_\theta \, l(\theta; f_1, y) - \nabla_\theta \, l(\theta; f_2, y)|_s \right\|_q \leq L_{s,f} \cdot \left\| |f_1 - f_2|_s \right\|_q,$$

holds uniformly for all θ.

Assumptions 3.1 and 3.2 are standard in the analysis of SGD and have appeared extensively in the literature; see, e.g., Ruppert (1988) and Polyak & Juditsky (1992).

As discussed earlier, factor loading estimation is subject to rotational ambiguity. Accordingly, given Q_{t-1} obtained from the first $t-1$ Oja's updates, we define θ^*_{t-1} as the minimizer of the rotated population risk

$$\theta^*_{t-1} = \arg\min_\theta L(\theta; R_{t-1}), \tag{7}$$
$$L(\theta; R_{t-1}) = \mathbb{E}_z \left[l(\theta; R_{t-1} f_{t,i}, y_{t,i}) \mid R_{t-1} \right], \tag{8}$$

where $R_{t-1} \in \mathbb{O}_k$ is the Procrustes rotation aligning Q_{t-1} with the oracle loading matrix V, and $\mathbb{E}_z[\cdot]$ denotes the expectation taken with respect to $(f_{t,i}, y_{t,i})$ only. Our goal is to characterize how close the FSGD iterate θ_t is to θ^*_{t-1}, which represents a rotated version of θ^* corresponding to the rotation R_{t-1}.

A key challenge arising from Oja's updates is that the rotation matrix R_t varies across iterations, causing the population minimizer θ^*_t to shift over time. To quantify this effect, the warm-start condition and Assumption 3.4 allow us to bound the change in R_t induced by each update of Q_t. Proposition 3.5 then uses this bound to control the *minimizer shift*, defined as the difference between successive minimizers θ^*_t and θ^*_{t-1}. Specifically, the following result shows that $\| |\theta^*_t - \theta^*_{t-1}|_s \|_q$ is of the same order as the Oja's step size η^{oja}_t multiplied by the perturbation bound M_A. Consequently, the drift of the target minimizer remains sufficiently controlled, allowing the algorithm to retain its convergence guarantees.

Assumption 3.4. Let $M_A < \infty$ be such that $\|A_t\|_{\mathrm{op}} \leq M_A$ a.s. for all t. Assume the Oja's step size in Stage II is non-increasing and satisfies

$$\eta^{\mathrm{oja}}_0 \leq \frac{1}{4(\sqrt{2}+1)M_A} k^{-2}.$$

The upper bound on $\eta^{\mathrm{oja}}_0 \leq \frac{1}{4(\sqrt{2}+1)M_A} k^{-2}$ depends on the uniform bound for A_t, rather than for $A_t - \Sigma$. This condition is required to control the perturbation error introduced by the QR normalization step.

Proposition 3.5 (Stability of θ^* under Oja's update)**.** *Suppose Assumptions 2.3, 3.1, 3.3 and 3.4 hold. Set* $\eta^{\mathrm{oja}}_t = \frac{\alpha}{(\beta+t)\rho_k}$ *with*

$$\beta \geq C \bar{\rho}_k^{-2} \left[q \log d + \log\left(\frac{k}{\bar{\rho}_k^2} \right) \right], \qquad \alpha \geq 8.$$

Then, for every $1 \leq t \leq T$ *with* $\log T = \mathcal{O}(\log d)$*,*

$$\left\| \left| \theta^*_t - \theta^*_{t-1} \right|_s \right\|_q \leq C \frac{L_{s,f}}{\mu} M_f \left(M_A k^{1/2} \eta^{\mathrm{oja}}_t + d^{-1/2} T^{-1} \right),$$

where $M_f := \| |f|_2 \|_q$*. In particular, if* $\eta^{\mathrm{oja}}_t \leq c_k / t$ *for some* c_k *independent of* t*, then*

$$\left\| \left| \theta^*_t - \theta^*_{t-1} \right|_s \right\|_q \leq C \frac{L_{s,f}}{\mu} M_f (M_A k^{1/2} c_k + d^{-1/2}) t^{-1}.$$

We now state the first main result of this paper. Theorem 3.6 establishes the moment convergence of FSGD iterates θ_t to the rotated oracle θ^*_{t-1}.

Theorem 3.6 (Moment Convergence of FSGD)**.** *Suppose Assumptions 2.1, 2.3, 3.1, 3.2, 3.3 and 3.4 hold. Let the step sizes $\{\eta_t\}_{t=1}^{\infty}$ be non-negative and monotonically non-increasing, satisfying $\sum_{i=0}^{\infty}\eta_i=\infty, t\eta_{t-1}\to\infty$, $\eta_0\le\frac{\mu}{9(\max\{q,s\}-1)L_{s,\theta}^2}$. Assume $\mathbb{E}|\nabla_\theta l(\theta;f_i,y_i)|_q^q<\infty$ for any θ. Set $\eta_t^{\mathrm{oja}}=\frac{\alpha}{(\beta+t)\rho_k}$ with*

$$\beta\ge\max\Big\{C\bar\rho_k^{-2}\big[q\log d+\log(\frac{k}{\bar\rho_k^2})\big],\frac{4(1+\sqrt{2e})\alpha}{\bar\rho_k},\,C\frac{\alpha^2}{\bar\rho_k^2}\Big\},$$

where C,α are absolute constant with $\alpha\ge 8$. Under polynomial step-size $\eta_t=\tilde{\mathcal{O}}((t+1)^{-\gamma})$ for some $\gamma\in(0,1)$ and $\log T=\mathcal{O}(\log d)$, we have the convergence rate for any $1\le t\le T$,

$$\begin{aligned}\Big\||\theta_t-\theta^*_{t-1}|_s\Big\|_q\le{}&A_1d^{-1/2}+A_2(m^{-1/2}+d^{-1/2})t^{-\gamma/2}\\&+A_3(C_{\alpha,\beta}+\frac{q^{1/2}}{\bar\rho_k})\,t^{-1/2}+A_4M_A\rho_k^{-1}t^{-(1-\gamma)}.\end{aligned}\tag{9}$$

where the constants A_i collect the remaining quantities that are independent of d, m, and t when k,q,s and the model constants are fixed. Their explicit definitions are given in (108) *in Appendix D.6. $C_{\alpha,\beta}=\mathcal{O}(\beta^{1/2})$ defined in* (64)*.*

Remark 3.7. Suppose the conditions of Theorem 3.6 hold. Consider sub-Gaussian (f_i,u_i) in (2), and assume $k=\Theta(1)$ and $m\ll d$. Then $M=\mathcal{O}(m^{-1/2}d\log d)$ a.s., $M_A=\mathcal{O}(d\log d)$, $\rho_k=\Theta(d),\bar\rho_k^{-1}=\Theta(\max\{m^{-1/2}\log d,1\})$. Thus, $q\log^3(d)$ is a sufficient upper order of β. Therefore, the rate in (9) becomes

$$\Big\||\theta_t-\theta^*_{t-1}|_s\Big\|_q\lesssim d^{-\frac12}+m^{-\frac12}t^{-\frac\gamma2}+\log^{3/2}(d)\,t^{-\frac12}+\log(d)\,t^{-(1-\gamma)}.\tag{10}$$

The above bound groups the errors according to their final rates rather than by mutually exclusive sources of error. The first term contains the idiosyncratic component in factor estimation; see Proposition 2.2, and also absorbs the bad-event contribution from the Oja's random updates by choosing β sufficiently large. The second term represents the standard convergence rate of mini-batch SGD (Tang et al., 2023), which scales with the batch size and learning rate. The third term captures the error propagation from Oja's online PCA into the FSGD updates. By Lemma 2.4, this term consists of the warm start transient term, summarized by $C_{\alpha,\beta}t^{-1/2}$, and the stochastic Oja term $\bar\rho_k^{-1}t^{-1/2}$. The fourth term accounts for the drift in the time-varying optimal rotation matrix R_t relative to the population minimizer θ^*. By balancing second and fourth error terms, the optimal rate is $d^{-\frac12}+t^{-\frac13}$ achieved at $\gamma=\frac23$.

Remark 3.8. The ℓ^s-norm provides a smooth surrogate for the ℓ^∞-norm, which is widely used in high-dimensional statistics; see, e.g., Fan & Li (2001); Wainwright (2019); Chen et al. (2025b); Li et al. (2025). Specifically, let $s=s_p$, where $s_p:=2\min\{\ell\in\mathbb{N}:2\ell>\log p\}$. Then the ℓ^{s_p}-norm and the ℓ^∞-norm are equivalent in the sense that $\|x\|_\infty\le\|x\|_{s_p}\le e\,\|x\|_\infty$. Consequently, by setting $s=s_p$, the moment bounds established in Theorem 3.6 immediately yield ℓ^∞-norm guarantees.

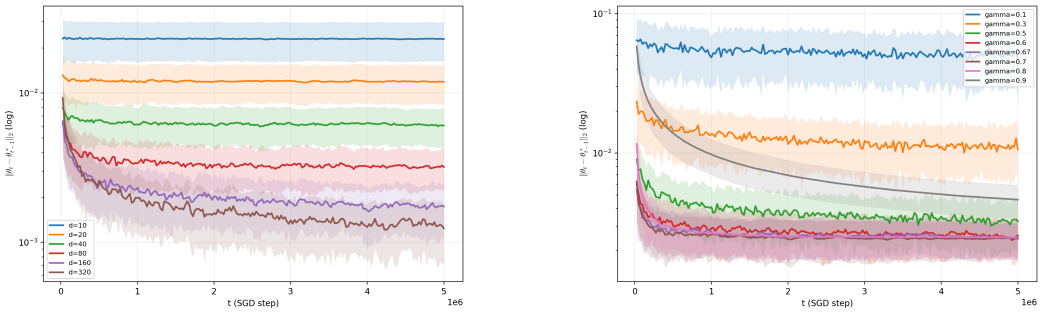
The fourth term in the convergence bounds (9) and (10) arises from the time-varying target induced by online factor loading updates. This term is undesirable as it does not benefit from larger learning rates. A practical remedy is to allocate the first half of the iterations to latent subspace learning and freeze the projection matrix after iteration $t_1=\lfloor t/2\rfloor$. As demonstrated in Corollary 3.9 below, this strategy ensures exponential decay of the instability induced by the time-varying minimizer θ_t^*.

Corollary 3.9 (Moment Convergence of FSGD (Frozen))**.** *Suppose the assumptions of Theorem 3.6 hold. Consider polynomial step size $\eta_t=\tilde{\mathcal{O}}((t+1)^{-\gamma})$ for some $\gamma\in(0,1)$, and suppose Oja's updates are frozen after iteration t_1. Define*

$$F_{t_1}:=C_2M_ft_1^{-\frac12}+C_3d^{-1/2},\qquad \Xi_{t_1}^2:=C_{s,q}^2m^{-1}+L_{s,f}^2F_{t_1}^2.$$

Then for the $\lfloor ta\rfloor$-th iterate with $t_1/t<a\le 1$, we have the following convergence rate

$$\Big\||\theta_{\lfloor ta\rfloor}-\theta^*_{t_1-1}|_s\Big\|_q\le\sqrt{\frac4\mu(\frac8\mu L_{s,f}^2+1)}F_{t_1}+2C_1\Xi_{t_1}\cdot t^{-\frac\gamma2},\tag{11}$$



(a) Varying d (b) Varying γ

Figure 1: Numerical validation of theoretical properties for FSGD. The results exhibit dimension and step-size scaling that aligns with the derived theoretical rates in Section 3.

where constants $C_1, C_2, C_3, C_{s,q}$ do not depend on t and are collected in Remark D.25. In particular, if $t_1 = \lfloor t/c_0 \rfloor$ for $c_0 > 1$, $m \ll d$ and $m \lesssim t^{1-\gamma}$, then (11) *simplifies to*

$$\left\| |\theta_{\lfloor ta \rfloor} - \theta^*_{t_1-1}|_s \right\|_q \lesssim d^{-\frac{1}{2}} + m^{-\frac{1}{2}} t^{-\frac{\gamma}{2}}.$$

Remark 3.10. The term F_{t_1} arises from factor estimation error, which decreases as the ambient dimension d grows and the subspace estimation improves, see the bound on $\| |\hat{f} - Rf|_s \|_q$ in Proposition 2.2 for details. Our theoretical results also apply to the case of offline PCA, where F_{t_1} can be replaced by the corresponding factor estimation bound and depends on the sample size of the offline PCA rather than t_1. As noted in Remark 3.7, the error propagation from Oja's online PCA decays faster than the standard SGD rate. Consequently, one may choose t_1 of smaller order than t while still matching or improving upon the mini-batch SGD rate $m^{-1/2} t^{-\gamma/2}$.

Compared with the rate in Theorem 3.6, the projection-frozen regime mitigates the effect of minimizer shift. As a result, it may exhibit improved convergence behavior when γ is large and the dimension d is high, at the cost of a larger constant in the factor estimation error term F_{t_1}.

4 Experiments

We evaluate FSGD on both synthetic and real data. Complete experimental details, baseline configurations, sensitivity analyses, and additional neural network experiments are deferred to Appendix B and Appendix C.

4.1 Synthetic Experiments

We first test FSGD under a linear factor model that matches our theoretical assumptions. The goal is to verify the scaling behavior stated by Theorem 3.6. We generate covariates from the factor model (2) and generate responses from a linear model. Unless otherwise specified, we use $k = 3$, mini-batch size $m = 5$, and average results over 100 repetitions. We also evaluate predictive performance and computational efficiency in a nonlinear neural network framework; details are in the Appendix B.2.

Figure 1 summarizes two key phenomena. First, when fixing $\gamma = 0.6$ and increasing d, the rotated parameter error decreases with the ambient dimension, consistent with the $d^{-1/2}$ term in Theorem 3.6. Second, when fixing $d = 100$ and varying the SGD decay exponent γ, the best performance is attained near $\gamma \approx 2/3$, matching the trade-off described in Remark 3.7: small γ leads to slow SGD convergence, while large γ amplifies minimizer drift induced by online factor learning.

4.2 Real-Data Experiment: MNIST

We evaluate FSGD on MNIST using one-pass image classification. For each of 30 random splits, the official training set is partitioned into 320 unlabeled warm-up images, 54,680 supervised training images and 5,000 validation images. Each split is evaluated using five training seeds, with the official 10,000-image test set fixed. Images are flattened into $d = 784$ dimensions. Projected methods use $\hat{k} = 64$ and a $64 \to 128 \to 10$ ReLU network, while Vanilla-NN uses $784 \to 128 \to 10$. All

methods minimize cross-entropy loss using plain mini-batch SGD with batch size 32 and learning rate $\eta_t = t^{-0.3}$.

Table 1: MNIST test accuracy (%) after one pass. Accuracy is mean (SD) across 30 split means, each averaging five training seeds. Gains are Factor-NN minus each baseline, in percentage points.

Method	Test accuracy	Gain [95% CI]
Factor-NN	**96.24 (0.07)**	—
Factor-NN-F0	95.71 (0.10)	0.535 [0.509, 0.561]
NN-Stiefel	95.96 (0.09)	0.284 [0.264, 0.303]
Vanilla-NN	95.82 (0.05)	0.418 [0.397, 0.440]
Random Projection	93.44 (0.10)	2.801 [2.763, 2.840]

Table 1 compares Factor-NN, which continues Oja updates throughout training, with four baselines: Factor-NN-F0 freezes the projection after warm-up; NN-Stiefel jointly learns the projection and classifier from the supervised loss while maintaining orthonormal projection columns; Vanilla-NN trains directly on the raw pixels; and Random Projection uses a fixed random orthogonal projection. In particular, the superior performance of Factor-NN supports the benefit of continued covariance-based representation learning for classification. These results provide empirical evidence for the effectiveness of FSGD on real image data beyond the strongly convex setting covered by our theory. Complete experimental settings, learning-rate and freezing-time comparisons, and additional CIFAR-10 and NCEP results are provided in Appendix C.

5 Conclusion

We develop FSGD, a new optimization framework that integrates online latent factor learning with stochastic gradient descent. Our analysis explicitly characterizes how representation estimation error and idiosyncratic noise propagate through the optimization dynamics, revealing a fundamental trade-off between factor estimation accuracy, optimization error and minimizer shift induced by online factor learning, which is a phenomenon that does not arise in classical SGD theory with fixed covariates. Empirical studies further confirm the strong empirical performance of FSGD. Overall, this work advances the theoretical understanding of high-dimensional SGD and opens new directions for integrating representation learning with scalable optimization.

6 Limitations

Despite its convergence guarantees and promising empirical results, FSGD has limitations in its modeling assumptions and theoretical scope.

First, the linear factor representation in (2) may fail to capture predictive structure on nonlinear manifolds, despite allowing a nonlinear regression function and full-dimensional idiosyncratic noise u_i. Nevertheless, linear PCA remains effective in many scientific applications (Greenacre et al., 2022; Zhang et al., 2024). Extending FSGD to nonlinear representations is an important future direction.

Second, our theoretical guarantees assume that the true factor dimension k is known and used by the algorithm. Extending the analysis to an estimated or misspecified dimension would require reformulating both the factor-estimation comparison (Proposition 2.2) and the Oja subspace analysis (Lemma 2.4). This entails substantial additional arguments beyond substituting the selected dimension for k in the existing bound. A rigorous treatment of these effects, together with adaptive rank selection for streaming data, remains an important direction for future research.

References

Allen-Zhu, Z. and Li, Y. First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–492. IEEE, 2017.

Bai, J. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

Bai, J. and Li, K. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40 (1):436–465, 2012.

Bai, J. and Ng, S. Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, 3 (2):89–163, 2008.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Ben Arous, G., Gheissari, R., and Jagannath, A. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

Bolivar, S., Chen, R., and Han, Y. Threshold tensor factor model in CP form. *arXiv preprint arXiv:2511.19796*, 2025.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Cai, T. T. and Zhang, A. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

Chen, B., Han, Y., and Yu, Q. Diffusion index forecasting with tensor data. *Journal of Econometrics*, 254:106204, 2026a.

Chen, B., Han, Y., and Yu, Q. Estimation and inference for CP tensor factor models. *Journal of Econometrics*, 253:106167, 2026b.

Chen, E., Han, Y., Li, J., and Xu, K. Modewise additive factor model for matrix time series. *arXiv preprint arXiv:2512.25025*, 2025a.

Chen, Z., Maguluri, S. T., and Zubeldia, M. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *The Annals of Applied Probability*, 35(2):1298–1352, 2025b.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Fan, J. and Gu, Y. Factor augmented sparse throughput deep ReLU neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548):2680–2694, 2024.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Fan, J., Xue, L., and Yao, J. Sufficient forecasting using factor models. *Journal of Econometrics*, 201 (2):292–306, 2017.

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'Enza, A., Markos, A., and Tuzhilina, E. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

Han, Y. and Zhang, C.-H. Tensor principal component analysis in high dimensional CP models. *IEEE Transactions on Information Theory*, 69(2):1147–1167, 2023.

Han, Y., Chen, R., and Zhang, C.-H. Rank determination in tensor factor model. *Electronic Journal of Statistics*, 16(1):1726–1803, 2022.

Han, Y., Chen, R., Yang, D., and Zhang, C.-H. Tensor factor model estimation by iterative projection. *The Annals of Statistics*, 52(6):2641–2667, 2024a.

Han, Y., Chen, R., Zhang, C.-H., and Yao, Q. Simultaneous decorrelation of matrix time series. *Journal of the American Statistical Association*, 119(546):957–969, 2024b.

Han, Y., Yang, D., Zhang, C.-H., and Chen, R. CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1383–1413, 2024c.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Huang, D., Niles-Weed, J., and Ward, R. Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pp. 2463–2498. PMLR, 2021.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P., and Sidford, A. Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Conference on Learning Theory*, pp. 1147–1164. PMLR, 2016.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.

Kasiviswanathan, S. P. SGD with low-dimensional gradients with applications to private and distributed learning. In *Uncertainty in Artificial Intelligence*, pp. 1905–1915. PMLR, 2021.

Kozak, D., Becker, S., Doostan, A., and Tenorio, L. Stochastic subspace descent. *arXiv preprint arXiv:1904.01145*, 2019.

Lai, T. L. Stochastic approximation. *The Annals of Statistics*, 31(2):391–406, 2003.

Lawley, D. N. and Maxwell, A. E. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

Li, J., Lou, Z., Schmidt-Hieber, J., and Wu, W. B. Statistical guarantees for high-dimensional stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Li, R.-C. New perturbation bounds for the unitary polar factor. *SIAM Journal on Matrix Analysis and Applications*, 16(1):327–332, 1995.

Liu, X., Kong, W., Jain, P., and Oh, S. DP-PCA: Statistically optimal and differentially private PCA. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29929–29943, 2022.

Luo, W., Xue, L., Yao, J., and Yu, X. Inverse moment methods for sufficient forecasting using high-dimensional predictors. *Biometrika*, 109(2):473–487, 2022.

Luo, Z., Han, Y., and Yu, X. Factor augmented supervised learning with text embeddings. *arXiv preprint arXiv:2508.06548*, 2025.

Luo, Z., Han, Y., and Yu, X. Supervised dynamic dimension reduction with deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24169–24177, 2026.

Oja, E. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

Oja, E. and Karhunen, J. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

Paquette, C., Lee, K., Pedregosa, F., and Paquette, E. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pp. 3548–3626. PMLR, 2021.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Implicit regularization or implicit conditioning? Exact risk trajectories of SGD in high dimensions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35984–35999, 2022.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214(1–2):1–90, 2025.

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

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Ruppert, D. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Stewart, G. W. Perturbation bounds for the QR factorization of a matrix. *SIAM Journal on Numerical Analysis*, 14(3):509–518, 1977.

Stock, J. H. and Watson, M. W. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002a.

Stock, J. H. and Watson, M. W. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002b.

Tang, K., Liu, W., Zhang, Y., and Chen, X. Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*, 2023.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Vogels, T., Karimireddy, S. P., and Jaggi, M. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, H., Sievert, S., Liu, S., Charles, Z., Papailiopoulos, D., and Wright, S. ATOMO: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Wu, L., Ma, C., and E, W. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yang, H., Zhang, T., and Xue, L. Theoretical guarantees for sparse principal component analysis based on the elastic net. *IEEE Transactions on Information Theory*, 71(9):7149–7175, 2025.

Yu, R., Chen, R., Xiao, H., and Han, Y. Dynamic matrix factor models for high dimensional time series. *arXiv preprint arXiv:2407.05624*, 2024a.

Yu, X., Yao, J., and Xue, L. Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors. *Journal of Business & Economic Statistics*, 40(1): 342–354, 2022.

Yu, X., Yao, J., and Xue, L. Power enhancement for testing multi-factor asset pricing models via Fisher's method. *Journal of Econometrics*, 239(2):105458, 2024b.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Zhang, K., Zemke, N. R., Armand, E. J., and Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature Methods*, 21(2):217–227, 2024.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024.

Zhou, G., Han, Y., and Yu, X. Covariate-adjusted deep causal learning for heterogeneous panel data models. *arXiv preprint arXiv:2505.20536*, 2025a.

Zhou, G., Han, Y., and Yu, X. Factor augmented tensor-on-tensor neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22928–22936, 2025b.

Appendices

This supplement consists of four sections. Appendix A summarizes the main notation. Appendix B reports detailed simulation studies, supplementing Section 4.1. Appendix C presents the evaluation on real-world datasets, supplementing Section 4.2. Appendix D includes all technical lemmas and complete proofs of the propositions and theorems in Sections 2 and 3.

A Summary of Notation

Table 2 summarizes the main notation. Data-defined quantities describe the data-generating model, the observed samples, or properties of the population objective for a fixed loss. Algorithmic quantities specify or are computed by FSGD. Analysis-only quantities are introduced to state and prove the theoretical results and need not be computed by the algorithm. In our theoretical results, the factor dimension used by FSGD equals the true factor dimension k.

Table 2: Summary of the main notation.

Symbol	Definition	Role
d	Ambient dimension of the observed covariates.	Data-defined
k	True factor dimension; also the factor dimension used by FSGD in the theoretical results.	Data-defined
$x_{t,i}$, $y_{t,i}$	Observed covariate and response for sample i in mini-batch t.	Data-defined
$f_{t,i}$, $u_{t,i}$, $\varepsilon_{t,i}$	Latent factor, idiosyncratic component, and response noise in the factor model.	Data-defined
B	Factor loading matrix in $\mathbb{R}^{d\times k}$, normalized by $d^{-1}B^\top B = I_k$.	Data-defined
Σ, Σ_f	Covariance matrices of the observed covariate and latent factor, respectively.	Data-defined
V	Matrix of the leading k eigenvectors of Σ.	Data-defined
λ_i	i-th largest eigenvalue of Σ.	Data-defined
ρ_k	Population eigengap $\lambda_k - \lambda_{k+1}$.	Data-defined
K_u	Upper bound on the sub-Gaussian norm of the idiosyncratic component.	Data-defined
$\lambda_k(\Sigma_f)$	Minimum eigenvalue of the factor covariance matrix.	Data-defined
M	Covariance-fluctuation bound in Assumption 2.3(i).	Data-defined
M_A	Almost-sure upper bound on $\|A_t\|_{\mathrm{op}}$ in Assumption 3.4.	Data-defined
M_f	Factor moment $\big\|\,\|f\|_2\,\big\|_q$.	Data-defined
μ	Strong-convexity constant in Assumption 3.1.	Data-defined
$L_{s,\theta}$, $L_{s,f}$	Gradient Lipschitz constants for the parameter and factor inputs, respectively, in Assumptions 3.2 and 3.3.	Data-defined
B_t	Observed mini-batch $\{(x_{t,i}, y_{t,i})\}_{i=1}^m$.	Data-defined
X_t	Mini-batch covariate matrix $[x_{t,1}, \ldots, x_{t,m}] \in \mathbb{R}^{d\times m}$.	Data-defined
A_t	Mini-batch matrix $m^{-1}X_tX_t^\top$. Its product with the subspace estimate can be evaluated without explicitly forming A_t.	Data-defined
p	Dimension of the model parameter θ.	Algorithmic
m	Mini-batch size.	Algorithmic
T_0, T	Number of warm-up updates and number of Stage-II iterations, respectively.	Algorithmic
Q_t	Subspace estimate after t Stage-II Oja updates; Q_0 is the estimate obtained after warm-up.	Algorithmic
$\widehat{f}_{t,i}$	Estimated factor $d^{-1/2}Q_{t-1}^\top x_{t,i}$.	Algorithmic
$l(\theta; f, y)$	Downstream loss function.	Algorithmic
θ_t	FSGD parameter iterate.	Algorithmic
g_t	Mini-batch stochastic gradient computed using the estimated factors.	Algorithmic
η_t	SGD step size.	Algorithmic

Continued on the next page

Table 2 continued

Symbol	Definition	Role
$\eta^{\mathrm{oja},(0)}$	Constant Oja step size during Stage-I warm-up.	Algorithmic
η_t^{oja}	Oja step size during Stage II.	Algorithmic
$\alpha,\ \beta$	Parameters in the theoretical Oja schedule $\eta_t^{\mathrm{oja}} = \alpha/[(\beta+t)\rho_k]$.	Algorithmic
γ	Decay exponent of the polynomial SGD step-size schedule.	Algorithmic
t_1	Iteration after which the subspace estimate is frozen in the frozen variant.	Algorithmic
$s,\ q$	Exponent of the parameter-space ℓ^s norm and order of the L^q moment, respectively.	Analysis-only
$\bar{\rho}_k$	Normalized eigengap $\min\{\rho_k/M, \rho_k/\|\Sigma\|_{\mathrm{op}}, 1\}$.	Analysis-only
R_t	Population-based orthogonal rotation used to compare estimated factors with rotated latent factors.	Analysis-only
$L(\theta; R)$	Rotated population risk $\mathbb{E}[l(\theta; Rf, y)]$.	Analysis-only
θ^*	Oracle minimizer $\arg\min_\theta L(\theta; I)$.	Analysis-only
θ_t^*	Moving population minimizer $\arg\min_\theta L(\theta; R_t)$.	Analysis-only
C	Absolute constant, unless stated otherwise.	Analysis-only
$C_{\alpha,\beta}$	Constant summarizing the warm-start transient in the Oja error bound.	Analysis-only
$A_1, \ldots, A_4$	Coefficients collecting the remaining quantities in the convergence bound of Theorem 3.6.	Analysis-only
$F_{t_1},\ \Xi_{t_1}$	Factor-estimation and gradient-error quantities used in the frozen-variant convergence bound.	Analysis-only

B Simulation

The synthetic experiments serve two purposes: (i) under a linear factor model satisfying the assumptions of our theory, they verify the predicted scaling behavior in Theorem 3.6; and (ii) under a nonlinear neural network model, they assess practical predictive performance and computational efficiency. All synthetic experiments, including both the linear and neural network simulations, were run on a computational cluster powered by Intel Xeon Gold 6342 processors. The experiments used up to 100 CPU cores in parallel, with approximately 4GB of memory allocated per core. The total compute budget was on the order of 10^4 CPU core-hours.

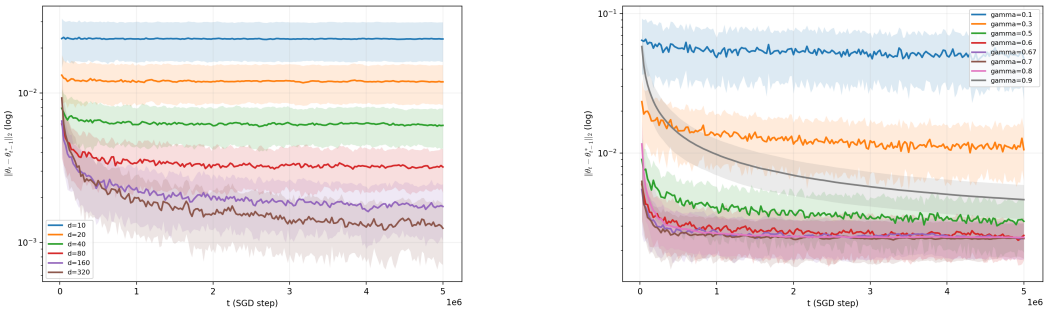
B.1 Synthetic Experiments: Linear Model

We first conduct linear model experiments under a factor model that exactly matches our theoretical assumptions. The goal is to verify the scaling behavior stated by Theorem 3.6.

Data-generating process. We strictly follow the factor model (2). Specifically, let $Z \in \mathbb{R}^{d\times k}$ have i.i.d. standard Gaussian entries, and let Q be the orthonormal factor from the QR decomposition of Z. We set the loading matrix to be $B = \sqrt{d}\,Q$. The latent factor f and idiosyncratic component u are independent and have i.i.d. $\mathrm{Unif}[-0.5, 0.5]$ entries. The response is generated by $y = f^\top\theta^* + \varepsilon$, where $\theta^* \sim \mathrm{Unif}(0,1)$ and $\varepsilon \sim \mathcal{N}(0, 0.3)$ are both independent from anything else. Each mini-batch contains $m = 5$ newly generated samples.

Optimization details. In the Stage I of Algorithm 1, we set constant learning rate 0.01 for $T_0 = 10$ iterations. In Stage II, the SGD step size follows $\eta_t = 0.5\,t^{-\gamma}$, and the Oja step size is $\eta_t^{\mathrm{oja}} = \frac{0.1}{50+t}$. Unless otherwise specified, we fix $k = 3$, run $T = 5\times 10^6$ updates, and average over 100 repetitions.

Scaling with ambient dimension. In the first experiment, we fix $\gamma = 0.6$ and vary $d \in \{10, 20, 40, 80, 160, 320\}$. Figure 2a shows that the rotated parameter error $|\theta_T - \theta^*_{T-1}|_s$ decreases as d grows, matching the $d^{-1/2}$ scaling predicted by the first term in Theorem 3.6.



(a) Performance of FSGD over different d

(b) Performance of FSGD over different γ

Figure 2: Performance of FSGD in linear SGD experiments.

Effect of the SGD decay exponent. In the second experiment, we fix $d = 100$ and vary $\gamma \in \{0.1, 0.3, 0.5, 0.6, 0.67, 0.7, 0.8, 0.9\}$. Figure 2b shows that the best performance is attained near $\gamma \approx 2/3$, which agrees with Remark 3.7. Small values of γ lead to slower SGD convergence, whereas overly large values increase the effect of minimizer drift.

B.2 Neural Network

We next evaluate FSGD in a nonlinear setting using two-layer ReLU neural networks. These experiments are adapted from Section 5.1 of Fan & Gu (2024) and are intended to assess the empirical behavior of FSGD beyond the linear theory.

Data-generating process. Under the factor model (2), the loading matrix B has i.i.d. $\mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ entries. The latent factor f and idiosyncratic component u are independent and have i.i.d. $\mathrm{Unif}[-1, 1]$ entries, respectively. The response variable is generated as $y = \mathcal{M}(f) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.3)$ is independent of (f, u), and $\mathcal{M}(f) = \sum_{j=1}^{k} \mathcal{M}_j(f_j)$. In each repetition, the component functions $\mathcal{M}_j$ are randomly selected from $\left\{\cos(\pi x),\ \sin(x),\ (1-|x|)^2,\ \frac{1}{1+e^{-x}},\ 2\sqrt{|x|} - 1\right\}$. Unless otherwise specified, the true latent dimension is fixed at $k = 5$. Depending on the experiment, we vary the ambient dimension d and the estimated factor dimension $\hat{r}$.

For Stage II of Algorithm 1, we generate $n_{\mathrm{II}} = 500$ labeled samples. For factor-based methods, we additionally use $n_{\mathrm{I}} = 50$ unlabeled samples for online PCA warm-up. We also generate an independent validation set of size $n_{\mathrm{valid}} = 150$. Model performance is evaluated by the test L_2 loss averaged over 15,000 independent test samples.

Implementation. We use fully connected ReLU neural networks with depth $L = 2$ and width $N = 100$ for all estimators. All models are trained using mini-batch SGD with batch size $m = 32$, initial learning rate 0.05, and 500 epochs. The SGD learning rate follows a polynomial decay schedule with exponent $\gamma = 0.3$.

For FSGD, the projection matrix is estimated using Oja's algorithm. In stage I, we initialize the factor estimator with a constant learning rate of 0.005 for 200 Oja iterations using the warm-up samples. In Stage II, we use $\eta_t^{\mathrm{oja}} = 0.05(50 + t)^{-1}$.

Methods. We compare FSGD with several baselines under different NN structures and projection updating strategies:

- **Factor-NN**: the proposed FSGD method, which updates the projection matrix online via Oja's algorithm throughout training;
- **Factor-NN-F0**: the same architecture as Factor-NN, but the projection matrix is frozen immediately after warm-up;
- **Factor-NN-F300**: the same architecture as Factor-NN, but the projection matrix is frozen after 300 epochs;

- **Oracle-NN**: uses the true latent factors as inputs and serves as an ideal benchmark without factor estimation error;
- **Vanilla-NN**: trains the neural network directly on the raw covariates x without representation learning;
- **PCA-A-NN**: augments the raw input with the online factor projection, i.e.,

$$z = \left[d^{-1/2} Q_{t-1}^\top x,\ x\right];$$

- **NN-Joint**: jointly learns the projection matrix and network weights from data, initialized from the warm-up phase;
- **NN-Stiefel**: jointly learns the projection matrix and network weights from the supervised loss, with the projection initialized by the Oja warm-up and constrained to have orthonormal columns;
- **Random Proj**: replaces Oja's online update with a random projection matrix sampled once at initialization and kept fixed thereafter;
- **PPCA**: periodically recomputes the projection matrix via offline PCA using the most recent W mini-batches, and performs this refresh every M mini-batches.

NN-Stiefel implementation. NN-Stiefel jointly optimizes the network parameters θ and the projection matrix Q, subject to $Q^\top Q = I_k$. For mini-batch B_t, define the empirical loss

$$\widehat{L}_t(\theta, Q) = \frac{1}{m} \sum_{i=1}^{m} \ell\big(\theta; d^{-1/2} Q^\top x_{t,i}, y_{t,i}\big),$$

where ℓ is the squared loss used in these experiments. Starting from the Euclidean gradient

$$G_t = \nabla_Q \widehat{L}_t(\theta_{t-1}, Q_{t-1}),$$

we project it onto the tangent space of the Stiefel manifold:

$$G_{t,\tan} = G_t - Q_{t-1}\,\mathrm{sym}(Q_{t-1}^\top G_t), \qquad \mathrm{sym}(A) = \frac{A + A^\top}{2}.$$

The projection is then updated by a tangent step followed by QR retraction,

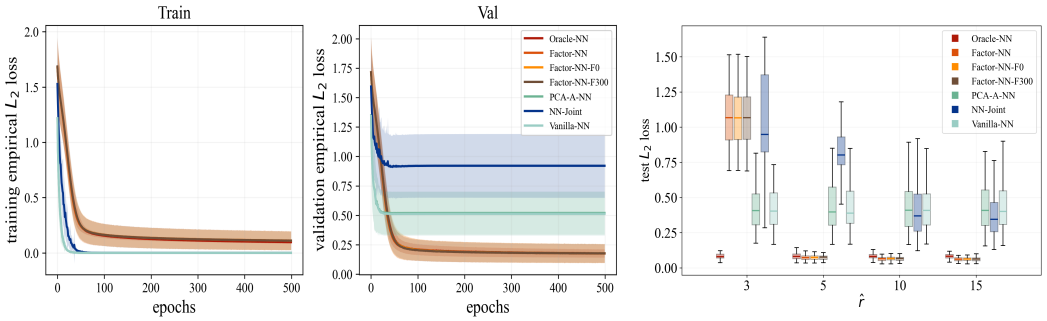
$$Q_t = \mathrm{qf}\big(Q_{t-1} - \eta_Q G_{t,\tan}\big),$$

where $\mathrm{qf}(A)$ denotes the orthonormal factor in the thin QR decomposition of A, with diagonal signs chosen so that the triangular factor has nonnegative diagonal entries. The network parameters are updated by ordinary SGD on the same mini-batch loss, with both gradients evaluated at (θ_{t-1}, Q_{t-1}). The tangent projection and QR retraction cost $O(dk^2)$ per update.

The projection learning rate η_Q is selected using validation loss from $\{0.001, 0.01, 0.1, 1.0\}$, yielding $\eta_Q = 0.1$. The network architecture and downstream SGD settings are the same as those used for Factor-NN.

Main comparison. Figure 3a plots the training and validation curves when $\hat{r} = 5$. FSGD-based methods consistently outperform all practical baselines and achieve performance close to the oracle benchmark. Figure 3b further reports the test loss under different estimated factor dimensions $\hat{r}$. The results show that Factor-NN is robust to moderate overestimation of the factor dimension, whereas underestimation leads to clear degradation. The "NN-Joint" baseline is unstable when $\hat{r}$ is small, but improves when the representation is over-specified. Moreover, the differences among Factor-NN, Factor-NN-F0, and Factor-NN-F300 are small when $\hat{r}$ is sufficiently large, suggesting that minimizer drift has only a limited effect in this regime.

Effect of ambient dimension. We also evaluate the effect of increasing ambient dimension $d \in \{100, 200, 400, 600, 800, 1000, 1500, 2000, 3000\}$, while fixing the true factor dimension $k = 5$ and the SGD decay exponent $\gamma = 0.3$. The results are summarized in Table 3. Across all settings, training for 500 epochs yields stable optimization, so the remaining error is primarily driven by factor estimation. Both PCA-A-NN and Vanilla-NN deteriorate as d increases, whereas Factor-NN



(a) Train and Validation Curves (b) Performance over Different Methods

Figure 3: Evaluation of Empirical Performance of FSGD

and its frozen-projection variants generally improve and approach the oracle benchmark. NN-Stiefel also outperforms PCA-A-NN and Vanilla-NN across all evaluated dimensions, supporting the benefit of learning a low-dimensional representation in this setting. However, its test loss remains consistently higher than that of Factor-NN, indicating that the advantage of Factor-NN is not solely due to correcting the scaling and conditioning issues of unconstrained joint optimization. The comparison suggests an advantage of covariance-based representation learning in this small-sample, high-dimensional setting. The overall improvement of Factor-NN with increasing d is qualitatively consistent with our factor-estimation analysis, in which a larger ambient dimension facilitates recovery of the latent factor structure.

Table 3: Mean empirical test L_2 loss under varying ambient dimension d. Numbers in parentheses are standard deviations over repetitions.

Method	d=100	d=200	d=400
Oracle-NN	0.087 (0.021)	0.083 (0.022)	0.082 (0.019)
Factor-NN	0.150 (0.047)	0.105 (0.021)	0.086 (0.018)
Factor-NN-F0	0.147 (0.045)	0.105 (0.022)	0.085 (0.019)
Factor-NN-F300	0.148 (0.045)	0.105 (0.021)	0.086 (0.019)
NN-Stiefel	0.248 (0.040)	0.196 (0.033)	0.143 (0.045)
PCA-A-NN	0.342 (0.060)	0.350 (0.092)	0.394 (0.131)
Vanilla-NN	0.347 (0.064)	0.360 (0.106)	0.387 (0.129)
Method	d=600	d=800	d=1000
Oracle-NN	0.084 (0.023)	0.085 (0.023)	0.085 (0.022)
Factor-NN	0.081 (0.019)	0.077 (0.017)	0.086 (0.090)
Factor-NN-F0	0.080 (0.019)	0.078 (0.017)	0.087 (0.092)
Factor-NN-F300	0.080 (0.017)	0.077 (0.017)	0.086 (0.091)
NN-Stiefel	0.137 (0.060)	0.111 (0.027)	0.111 (0.035)
PCA-A-NN	0.406 (0.149)	0.420 (0.171)	0.430 (0.168)
Vanilla-NN	0.403 (0.147)	0.417 (0.162)	0.433 (0.171)
Method	d=1500	d=2000	d=3000
Oracle-NN	0.085 (0.023)	0.085 (0.024)	0.084 (0.024)
Factor-NN	0.072 (0.019)	0.071 (0.018)	0.073 (0.019)
Factor-NN-F0	0.073 (0.018)	0.071 (0.018)	0.070 (0.015)
Factor-NN-F300	0.074 (0.018)	0.072 (0.017)	0.069 (0.015)
NN-Stiefel	0.104 (0.036)	0.105 (0.047)	0.097 (0.037)
PCA-A-NN	0.443 (0.179)	0.447 (0.184)	0.453 (0.193)
Vanilla-NN	0.444 (0.178)	0.447 (0.182)	0.454 (0.190)

Interestingly, FSGD-based methods can outperform Oracle-NN when d is large. Our diagnostic experiments attribute this observed advantage to the interaction between factor rotation and the neural network's default first-layer initialization. PyTorch initializes the first-layer weights independently from $\mathrm{Unif}(-1/\sqrt{k}, 1/\sqrt{k})$, which is not rotationally invariant. We compare this initialization with the rotation-invariant Gaussian distribution $\mathcal{N}(0, 1/(3k))$, which has the same entrywise variance.

We conducted 50 paired runs at $d = 1000$. Let Q_0 denote the projection obtained after warm-up, R the corresponding orthogonal alignment such that $d^{-1/2}Q_0^\top x \approx R^\top f$, and W_0 the initial first-layer weight matrix. We compare four configurations, specified by their inputs and initial first-layer weights in Table 4. The rotation-only configuration changes the factor coordinates while keeping W_0 fixed. The rotation-with-matched-weights configuration also replaces W_0 by W_0R, preserving the initial predictions because $(W_0R)(R^\top f) = W_0 f$. The estimated-factor configuration additionally includes factor-estimation error. Within each run, all configurations share the data, mini-batch order, remaining initialization, and other training settings.

Table 4: Test MSE at $d = 1000$ under uniform and Gaussian first-layer initialization, reported as mean (SD) over 50 paired runs.

Configuration	Input	Weights	Uniform	Gaussian
Oracle	f	W_0	0.0850 (0.0220)	0.0662 (0.0146)
Rotation only	$R^\top f$	W_0	0.0670 (0.0156)	0.0684 (0.0154)
Rotation + matched weights	$R^\top f$	W_0R	0.0850 (0.0221)	0.0662 (0.0146)
Estimated factors	$d^{-1/2}Q_0^\top x$	W_0	0.0742 (0.0174)	0.0754 (0.0166)

Matching the first-layer weights to the factor rotation reproduces the oracle performance under both initializations. Under Gaussian initialization, the difference between the rotation-only and oracle configurations is not statistically significant (paired 95% confidence interval: $[-0.00018, 0.00447]$), while the estimated-factor configuration performs worse. These comparisons support an explanation based on the interaction between factor coordinates and initialization, rather than a benefit from factor-estimation error.

Time-memory trade-off. To further quantify the computational cost of different online learning schemes, we compare the per-step runtime, memory usage, and test loss at $d = 4000$, averaged over 50 repetitions. Both the per-step runtime and memory usage include the additional overhead from Oja updates.

We summarize the main findings from Table 5. First, FSGD-based methods achieve the best test loss among the neural network-based methods, including PCA-A-NN and NN-Joint, and substantially outperform Random Projection and the periodic online projection update scheme PPCA. In terms of efficiency, FSGD is considerably cheaper than Vanilla SGD in both per-step runtime and memory usage (see comparison between Factor-NN and Vanilla-NN), although it is more expensive than NN-Joint and Oracle-NN due to the additional Oja update. However, after freezing the learned subspace, the Factor-NN-F300 achieves a computational cost comparable to the NN-Joint and Oracle-NN baselines while maintaining the lowest test loss.

In contrast, PPCA can be computationally expensive when the projection is recomputed frequently or when the PCA window size W is large, yet it still suffers from high prediction error. One possible explanation is that each periodic SVD recomputation induces an abrupt change in the learned representation, making it difficult for the downstream neural network to readapt, especially under a decaying learning rate schedule. PPCA also incurs a larger memory cost because it needs to store the sliding window of mini-batch samples for PCA recomputation.

Warm-up sensitivity. We consider a deliberately challenging setting with $d = 1000$ and $n_{\mathrm{warm}} = 5$, averaged over 50 repetitions. The oracle test MSE is 0.087. Table 6 reports the results for Factor-NN-F0 where the projection is frozen immediately after warm-up, while Table 7 reports the results for Factor-NN where Oja updates continue throughout Stage II.

Overall, FSGD (Factor-NN) is insensitive to the warm-up hyperparameters as long as Oja updates are continued in Stage II. The reason is that online representation learning can correct an inaccurate subspace estimate obtained from a poor warm-up phase. In Table 6 where the projection is frozen

Table 5: Neural-network time-memory trade-off at $d = 4000$, averaged over 50 repetitions. Per-step time includes Oja/SVD overhead.

Method	SGD Per-step (ms)	SGD Memory (MB)	Test Loss
Factor-NN	1.08	1.15	0.0687
Factor-NN-F0	0.48	0.62	0.0694
Factor-NN-F300	0.50	1.15	**0.0677**
PCA-A-NN	2.61	4.21	0.4814
NN-Joint	0.57	0.23	0.6233
Oracle-NN	**0.39**	**0.08**	0.0876
Vanilla-NN	1.79	3.13	0.4746
Random Proj	0.48	0.62	1.6147
PPCA $(M = 20, W = 2)$	1.17	4.33	1.0936
PPCA $(M = 20, W = 5)$	3.98	10.35	1.0363
PPCA $(M = 20, W = 10)$	8.77	20.70	1.0588
PPCA $(M = 100, W = 2)$	0.62	4.33	1.2450
PPCA $(M = 100, W = 5)$	1.18	10.35	1.2811
PPCA $(M = 100, W = 10)$	2.14	20.70	1.1816
PPCA $(M = 500, W = 2)$	0.51	4.33	1.5810
PPCA $(M = 500, W = 5)$	0.62	10.35	1.6352
PPCA $(M = 500, W = 10)$	0.81	20.70	1.6684

immediately after warm-up (Factor-NN-F0), performance degrades severely with MSE ranging from 0.53 to 1.49. In contrast, Factor-NN with continued Oja updates achieves MSE 0.074-0.097 even at $T_0 = 1$, demonstrating that Stage II online updates can correct a poor initialization without relying on warm-up. These results indicate that a long or carefully tuned warm-up phase is not necessary in practice.

Table 6: Warm-up sensitivity for Factor-NN-F0 at $d = 1000$ with $n_{\text{warm}} = 5$ (50 repetitions). Rows correspond to warm-up length T_0 and columns correspond to the warm-up learning rate. Entries are test MSE.

T_0	0.0005	0.001	0.005	0.01	0.05
1	1.487	1.455	1.423	1.426	1.435
5	1.362	1.318	1.243	1.233	1.222
10	1.231	1.140	1.027	1.014	0.999
20	1.055	0.952	0.831	0.812	0.800
50	0.791	0.686	0.568	0.548	0.532

Table 7: Warm-up sensitivity for Factor-NN at $d = 1000$ with $n_{\text{warm}} = 5$ (50 repetitions). Rows correspond to warm-up length T_0 and columns correspond to the warm-up learning rate. Entries are test MSE.

T_0	0.0005	0.001	0.005	0.01	0.05
1	0.083	0.084	0.097	0.074	0.076
5	0.077	0.076	0.075	0.079	0.075
10	0.078	0.075	0.077	0.077	0.083
20	0.085	0.076	0.075	0.075	0.075
50	0.088	0.076	0.076	0.081	0.075

Sensitivity to Stage-II Oja hyperparameters. Finally, we study the sensitivity to the Stage-II Oja step-size schedule $\eta_t^{\text{oja,II}} = c_{\text{oja}}(C_{\text{oja}} + t)^{-1}$. We vary $c_{\text{oja}} \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ and $C_{\text{oja}} \in \{500, 100, 50, 10, 5\}$ at $d = 1000$, using 50 repetitions and no alignment.

Tables 8 and 9 report the results for Factor-NN and Factor-NN-F300, respectively. Across all configurations, the performance remains remarkably stable. For Factor-NN, the test loss ranges roughly from 0.0719 to 0.0954, and for Factor-NN-F300 from 0.0706 to 0.0844, both close to the oracle benchmark. Thus, the method is not particularly sensitive to the choice of Stage-II Oja hyperparameters, which further supports the practical robustness of the online subspace update.

Table 8: Sensitivity to Stage-II Oja hyperparameters for Factor-NN at $d = 1000$ (50 repetitions). Entries are test MSE under $\eta_t^{\mathrm{oja,II}} = c_{\mathrm{oja}}(C_{\mathrm{oja}} + t)^{-1}$.

$C_{\mathrm{oja}} \backslash c_{\mathrm{oja}}$	0.005	0.01	0.05	0.1	0.5
500	0.0799	0.0902	0.0836	0.0771	0.0804
100	0.0858	0.0954	0.0752	0.0768	0.0801
50	0.0914	0.0910	0.0735	0.0739	0.0803
10	0.0910	0.0820	0.0722	0.0722	0.0806
5	0.0778	0.0759	0.0734	0.0719	0.0806

Table 9: Sensitivity to Stage-II Oja hyperparameters for Factor-NN-F300 at $d = 1000$ (50 repetitions). Entries are test MSE under $\eta_t^{\mathrm{oja,II}} = c_{\mathrm{oja}}(C_{\mathrm{oja}} + t)^{-1}$.

$C_{\mathrm{oja}} \backslash c_{\mathrm{oja}}$	0.005	0.01	0.05	0.1	0.5
500	0.0844	0.0843	0.0715	0.0716	0.0718
100	0.0844	0.0730	0.0714	0.0724	0.0718
50	0.0770	0.0716	0.0720	0.0718	0.0710
10	0.0722	0.0710	0.0721	0.0709	0.0714
5	0.0715	0.0706	0.0716	0.0710	0.0714

Benefits of Stage-II representation updates. The similar performance of Factor-NN and Factor-NN-F0 in Table 3 can be explained by the informative standard warm-up, which uses 50 samples and 200 Oja iterations. The benefit of continued representation learning becomes clearer when the initial projection is less accurate. With only five warm-up samples, Factor-NN-F0 retains the inaccurate projection and incurs test MSEs of 0.532–1.487 (Table 6), whereas Factor-NN achieves 0.074–0.097 by continuing to refine it (Table 7). Thus, freezing can be effective when warm-up already provides an accurate projection, while Stage-II updates reduce reliance on the quality of the initial representation.

Summary. Taken together, these experiments support three conclusions. First, factor-based learning substantially improves predictive performance over training directly in the ambient space. Second, the online update in FSGD achieves an attractive statistical-computational trade-off, combining low test loss with moderate runtime and memory. Third, the method is robust to both warm-up design and Stage-II Oja hyperparameters, indicating that its empirical performance does not rely on delicate tuning.

C Real-Data Experiment

C.1 MNIST

We evaluate FSGD on MNIST to assess its empirical performance in image classification.

Data and preprocessing. For each of 30 random splits, the official training set is partitioned into 320 unlabeled warm-up images, 5,000 validation images, and 54,680 supervised training images. The official test set of 10,000 images remains fixed. Each split is evaluated using five training seeds. Images are flattened into $d = 784$ dimensions, divided by 255, and centered using the pixelwise mean estimated from the warm-up images.

Models and training. The selected projection dimension is $\hat{k} = 64$. Projected methods use a fully connected ReLU network with architecture $64 \to 128 \to 10$, while Vanilla-NN uses $784 \to 128 \to$

10. All methods minimize cross-entropy loss using plain mini-batch SGD without momentum, with batch size 32 and weight decay 10^{-4}. Training makes one pass over the supervised samples, giving 1,709 updates. The classifier learning rate is

$$\eta_t = \eta_0 t^{-0.3}, \qquad \eta_0 \in \{0.75, 1.00, 1.25\},$$

where t starts at 1.

Stage I makes one pass over the warm-up images using Oja's algorithm with batch size 8 and constant step size 0.1, giving 40 updates. During Stage II, Factor-NN updates the projection after each supervised mini-batch using

$$\eta_u^{\mathrm{oja}} = \frac{5}{500 + u}, \qquad u = 0, 1, \ldots.$$

After each Stage-I or Stage-II Oja update, the basis obtained by thin QR is aligned to the preceding basis using Procrustes alignment.

Comparison methods. We compare five methods: Factor-NN continues the Oja updates throughout supervised training; Factor-NN-F0 freezes the projection immediately after Stage I; NN-Stiefel jointly learns the projection and classifier from the supervised loss; Vanilla-NN uses the raw covariates; and Random Projection uses a fixed random orthogonal projection. In these image experiments, NN-Stiefel updates the projection using a Euclidean gradient step with learning rate 0.01, followed by thin QR retraction to maintain orthonormal columns. Factor-NN, Factor-NN-F0, and NN-Stiefel are initialized with the same projection matrix obtained from the Oja warm-up. Within each split and training seed, projected methods share the downstream network initialization, and all methods use the same supervised mini-batch order.

Evaluation and statistical analysis. For both image datasets, we report test accuracy at the end of the single supervised pass. We first average the five training seeds within each split and then report the mean and sample standard deviation across the 30 split means. The full learning-rate comparisons are reported below; $\eta_0 = 1$ serves as a common reference setting for the two datasets. At this setting, we compare Factor-NN with each of the four baselines using paired differences between split means. We use two-sided t-tests with 29 degrees of freedom and apply Holm correction jointly to the eight comparisons across the two datasets. The reported 95% confidence intervals are unadjusted paired t intervals. These comparisons quantify variability under the repeated-split evaluation with the official test set fixed.

Predictive performance. Table 10 reports the complete learning-rate comparison. Factor-NN achieves the highest mean test accuracy at every evaluated learning rate. At $\eta_0 = 1$, its mean accuracy is 96.2426%, compared with 95.7078% for Factor-NN-F0, 95.9590% for NN-Stiefel, and 95.8245% for Vanilla-NN. The paired gains and confidence intervals are reported in Table 13. All four comparisons remain significant after Holm correction. These results provide empirical evidence that FSGD is effective on real image data, without imposing an explicit linear factor model on the data-generating process.

Table 10: MNIST test accuracy (%) after one pass, reported as mean (SD) across 30 split means.

Method	$\eta_0 = 0.75$	$\eta_0 = 1.00$	$\eta_0 = 1.25$
Factor-NN	**95.8541 (0.0593)**	**96.2426 (0.0710)**	**96.4695 (0.0780)**
Factor-NN-F0	95.3087 (0.0954)	95.7078 (0.0998)	95.9494 (0.0929)
NN-Stiefel	95.5961 (0.0840)	95.9590 (0.0880)	96.1737 (0.0952)
Vanilla-NN	95.2636 (0.0493)	95.8245 (0.0533)	96.1847 (0.0628)
Random Projection	92.9389 (0.0842)	93.4412 (0.0992)	93.7388 (0.0960)

Effect of freezing time. To examine the benefit of continued representation learning, we additionally freeze the projection after the first $\lceil \alpha T \rceil$ supervised updates, where $\alpha \in \{1/4, 1/2, 3/4\}$ and $T = 1709$ is the total number of supervised updates. The corresponding freezing times are 428, 855, and 1,282 updates. The downstream classifier continues training after the projection is frozen, and all other settings remain unchanged.

Table 11 shows that mean test accuracy increases with the coupling duration at all three learning rates. Freezing immediately after warm-up incurs the largest loss, while freezing after $3T/4$ gives accuracy very close to full coupling. This pattern illustrates the trade-off reflected in Corollary 3.9: freezing eliminates subsequent minimizer drift but retains the representation error at the freezing time. The empirical results show that freezing too early can reduce predictive performance. Full coupling also avoids the need to choose a freezing time in advance.

Table 11: MNIST test accuracy (%) under different freezing times, reported as mean (SD) across 30 split means.

Method	$\eta_0 = 0.75$	$\eta_0 = 1.00$	$\eta_0 = 1.25$
Factor-NN-F0	95.3087 (0.0954)	95.7078 (0.0998)	95.9494 (0.0929)
Factor-NN-$T/4$	95.7549 (0.0672)	96.1515 (0.0766)	96.3741 (0.0891)
Factor-NN-$T/2$	95.8251 (0.0667)	96.2125 (0.0749)	96.4371 (0.0831)
Factor-NN-$3T/4$	95.8461 (0.0609)	96.2365 (0.0760)	96.4581 (0.0803)
Factor-NN	**95.8541 (0.0593)**	**96.2426 (0.0710)**	**96.4695 (0.0780)**

C.2 CIFAR-10

We further evaluate the same five methods on CIFAR-10 using a one-pass classification protocol.

Data and preprocessing. For each of 30 random splits, the official training set is partitioned into 1,000 unlabeled warm-up images, 5,000 validation images, and 44,000 supervised training images. Each split is evaluated using five training seeds, and the official test set of 10,000 images remains fixed. Images are flattened into $d = 3072$ dimensions. All pixels are divided by 255 and centered using pixelwise means estimated from the warm-up images.

Models and training. The selected projection dimension is $\hat{k} = 128$. Projected methods use a fully connected ReLU network with architecture $128 \to 256 \to 256 \to 10$, while Vanilla-NN uses $3072 \to 256 \to 256 \to 10$. All methods minimize cross-entropy loss using plain SGD without momentum, with batch size 128 and weight decay 10^{-4}. One pass over the supervised samples gives 344 updates. The classifier learning rate is

$$\eta_t = \eta_0 t^{-0.3}, \qquad \eta_0 \in \{0.3, 0.5, 1.0, 1.5\},$$

where t starts at 1.

Stage I makes one pass over the warm-up images using Oja's algorithm with batch size 32 and constant step size 0.1, giving 32 updates. During Stage II, Factor-NN updates the projection after each supervised mini-batch using

$$\eta_u^{\mathrm{oja}} = \frac{5}{500 + u}, \qquad u = 0, 1, \ldots.$$

The Oja alignment procedure and comparison methods follow Section C.1. NN-Stiefel again uses a projection learning rate of 0.01 and thin QR retraction. Factor-NN, Factor-NN-F0, and NN-Stiefel are initialized with the same projection matrix obtained from the Oja warm-up. Evaluation and statistical analysis also follow Section C.1.

Predictive performance. Table 12 reports the complete learning-rate comparison. Factor-NN achieves the highest mean test accuracy at every evaluated learning rate. At the common setting $\eta_0 = 1$, it improves over Factor-NN-F0, NN-Stiefel, Vanilla-NN, and Random Projection by 0.600, 0.471, 0.547, and 2.532 percentage points, respectively. All four comparisons remain significant after Holm correction; the paired confidence intervals are reported in Table 13. Together with the MNIST results, these comparisons provide empirical evidence for the predictive benefit of continued representation learning in one-pass classification.

C.3 NCEP/NCAR Reanalysis

We further evaluate FSGD on the NCEP/NCAR Reanalysis 1 dataset (Kalnay et al., 1996), a global atmospheric reanalysis dataset containing monthly Z500 anomalies on a $2.5^\circ \times 2.5^\circ$ grid, publicly

Table 12: CIFAR-10 test accuracy (%) after one pass, reported as mean (SD) across 30 split means.

Method	$\eta_0 = 0.3$	$\eta_0 = 0.5$	$\eta_0 = 1.0$	$\eta_0 = 1.5$
Factor-NN	**40.2748 (0.1450)**	**42.9919 (0.2033)**	**45.2353 (0.3625)**	**45.6488 (0.4436)**
Factor-NN-F0	39.8564 (0.1608)	42.4607 (0.1958)	44.6351 (0.3429)	45.0264 (0.4238)
NN-Stiefel	40.0243 (0.1741)	42.6857 (0.2117)	44.7638 (0.3831)	45.1144 (0.4161)
Vanilla-NN	37.9615 (0.1574)	41.9580 (0.2046)	44.6887 (0.5061)	45.2200 (0.5833)
Random Projection	37.6999 (0.1870)	40.4037 (0.1796)	42.7037 (0.2985)	43.2419 (0.3561)

Table 13: Paired test-accuracy gains of Factor-NN at $\eta_0 = 1$, in percentage points.

Dataset	Baseline	Mean gain	95% CI
MNIST	Factor-NN-F0	0.535	[0.509, 0.561]
	NN-Stiefel	0.284	[0.264, 0.303]
	Vanilla-NN	0.418	[0.397, 0.440]
	Random Projection	2.801	[2.763, 2.840]
CIFAR-10	Factor-NN-F0	0.600	[0.562, 0.638]
	NN-Stiefel	0.471	[0.391, 0.552]
	Vanilla-NN	0.547	[0.433, 0.660]
	Random Projection	2.532	[2.457, 2.606]

available from NOAA Physical Sciences Laboratory. The dataset consists of $T = 887$ monthly observations from 1950 to 2023 over $d = 10{,}511$ spatial grid points. Our prediction task is to forecast the next-month anomaly at a target location using the current global field. This setting is well suited to our framework because large-scale climate variability induces a strong low-dimensional factor structure in the high-dimensional covariates. The first 120 months are used for warm-up, the next 240 months for online learning and hyperparameter selection, and the held-out period 1980–2023 (527 months) is used for test evaluation.

Linear prediction model. To isolate the effect of representation learning, all methods use the same linear predictor. Specifically, each method predicts the response using either $y_t = x_t^\top \theta$ or $y_t = f_t^\top \theta, \quad f_t = Q^\top x_t$, depending on whether the method operates in the original ambient space or in a learned low-dimensional factor space.

Methods. We compare the following estimators.

FSGD. FSGD first estimates a projection matrix $Q \in \mathbb{R}^{d\times k}$ during warm-up via streaming Oja updates, and then initializes the regression coefficient by ordinary least squares on the warm-up period:

$$\hat{\theta}_{\text{warm}} = (F_{\text{warm}}^\top F_{\text{warm}})^{-1} F_{\text{warm}}^\top y_{\text{warm}}, \qquad F_{\text{warm}} = X_{\text{warm}} Q.$$

During online learning, FSGD jointly updates the projection matrix Q using Oja's algorithm and the regression coefficient θ using SGD in the k-dimensional factor space. We search over

$$k \in \{5, 10, 20\}, \quad c_{\text{sgd}} \in \{10^{-4}, \dots, 1\}, \quad \gamma_{\text{sgd}} \in \{0.1, 0.3, 0.67, 0.8\},$$

$$c_{\text{oja}} \in \{10^{-5}, \dots, 10^{-2}\}, \quad \gamma_{\text{oja}} = 1, \quad c_{\text{warm}} \in \{0.5, 1.0\}.$$

Its storage cost is $O(dk)$, which is about 0.40 MB when $k = 5$.

Periodic PCA + SGD (PPCA). PPCA first computes an offline PCA projection matrix on the warm-up period and initializes θ by OLS in the projected space. During online learning, it updates θ by SGD, and every M months recomputes the projection matrix by offline PCA on the most recent window of size W. After each PCA refresh, the regression coefficient is rotated accordingly to align with the new subspace. We search over

$$k \in \{5, 10, 20\}, \quad M \in \{3, 6, 12, 24\}, \quad W \in \{30, 60, 120\},$$

together with SGD hyperparameters

$$c_{\text{sgd}} \in \{10^{-5}, \dots, 10^{-3}\}, \qquad \gamma_{\text{sgd}} \in \{0.1, 0.3, 0.67, 0.8\}.$$

Its storage cost is $O(Wd)$, since the window data must be retained for repeated SVD computations; for $W = 120$, this is approximately 9.62 MB when $k = 5$.

Vanilla SGD. Vanilla SGD fits a linear model directly in the original $d = 10{,}511$ dimensional space. It is initialized using the minimum-norm warm-up OLS solution

$$\hat{\theta}_{\text{warm}} = (X_{\text{warm}}^\top X_{\text{warm}})^+ X_{\text{warm}}^\top y_{\text{warm}},$$

and is then updated by SGD online. We search over

$$c_{\text{sgd}} \in \{10^{-7}, \ldots, 10^{-4}\}, \qquad \gamma_{\text{sgd}} \in \{0.1, 0.3, 0.67, 0.8\}.$$

Its storage cost is $O(d)$, about 0.08 MB.

Random Projection + SGD (RP). RP replaces the learned projection matrix with a fixed random orthogonal matrix $R \in \mathbb{R}^{d\times k}$, sampled once at initialization and never updated thereafter. The regression coefficient is initialized by OLS on the projected warm-up data and then updated by SGD online in the k-dimensional projected space. We search over

$$k \in \{5, 10, 20\}, \qquad c_{\text{sgd}} \in \{10^{-5}, \ldots, 10^{-3}\}, \qquad \gamma_{\text{sgd}} \in \{0.1, 0.3, 0.67, 0.8\}.$$

Its storage cost is also $O(dk)$, about 0.40 MB when $k = 5$.

Persistence. The persistence baseline predicts $y_{t+1} = y_t$.

Prevailing Mean. As an additional naive baseline, we also consider the prevailing mean predictor $y_{t+1} = \frac{1}{t}\sum_{s=1}^{t} y_s$.

Hyperparameter selection. All hyperparameters are selected by maximizing R^2 on the validation period 1960–1979. The held-out test period 1980–2023 is never used for model selection.

Predictive performance. Table 14 reports the test R^2 on the held-out period 1980–2023 for four response locations, which were randomly selected from those satisfying a common stability criterion: all compared methods achieved test $R^2 > -1$. This criterion was applied uniformly to exclude severe numerical or predictive failures. FSGD achieves the best performance in three of the four selected locations, outperforming both Vanilla SGD and Random Projection in all four cases. Compared with PPCA, FSGD is slightly better on three locations, while PPCA performs better on one low-persistence location. These results suggest that a low-dimensional factor structure is useful in the NCEP task. FSGD further provides an online strategy to track this structure, achieving performance comparable to or better than PPCA while avoiding repeated offline SVD and large window storage. The prevailing mean baseline performs poorly across all locations.

Table 14: Test R^2 on the held-out period 1980–2023 for four selected response variables.

Response	FSGD	PPCA	Persistence	Vanilla SGD	Random Proj	Prev. Mean
P06_TropSPac4	**0.641**	0.631	0.601	0.549	0.542	-0.965
P05_TropNPac2	**0.568**	0.553	0.500	0.459	0.487	-0.774
P05_TropSPac8	**0.592**	0.576	0.502	0.449	0.375	-0.655
P01_SubNAtl10	0.379	**0.418**	0.104	0.175	0.329	-0.581

Computational cost. We also compare total runtime, per-step runtime, and storage cost on the NCEP task. The results are summarized in Table 15. For FSGD, runtime increases with the projection dimension k due to the additional online Oja updates. Vanilla SGD and Random Projection have lower runtimes, whereas FSGD achieves higher test R^2 at all four selected locations in the reported run. PPCA ($k = 5$) can become significantly more expensive when projection updates are frequent or the window size is large, since it repeatedly performs SVD on high-dimensional data and must store the window observations.

We note that runtime measurements depend on the hardware, BLAS/LAPACK backend, and Python/numpy environment; therefore, the reported computational times should be interpreted as machine-dependent wall-clock benchmarks rather than hardware-independent quantities.

Table 15: Computational cost on the NCEP real-data task.

Method	Total Time (s)	Per-step (ms)	Storage (MB)
FSGD ($k = 5$)	**0.61**	**0.79**	**0.40**
FSGD ($k = 10$)	1.34	1.75	0.80
FSGD ($k = 15$)	2.18	2.84	1.20
PPCA (range over $M \in \{3, 6, 12, 24\}$, $W \in \{30, 60, 120\}$)	0.30–11.11	0.39–14.48	2.41–9.62
PPCA ($M = 3, W = 120$)	11.11	14.48	9.62
PPCA ($M = 12, W = 120$)	2.78	3.63	9.62
PPCA ($M = 24, W = 30$)	0.30	0.39	2.41
Vanilla SGD	0.04	0.05	0.08
Random Proj ($k = 5$)	0.02	0.03	0.40
Random Proj ($k = 10$)	0.04	0.05	0.80
Persistence	0.00	0.00	0.00

Summary. The NCEP results illustrate the usefulness of factor-based representations for the selected forecasting tasks, while the relative predictive performance of FSGD and PPCA varies across locations. The main computational benefit of FSGD relative to periodic PCA is that it avoids repeated offline SVD and the storage of a sliding data window. The four selected locations provide illustrative comparisons rather than evidence of an average predictive advantage across all NCEP targets.

D Proofs

D.1 Auxiliary Lemmas

Lemma D.1 (Lemma 2 of Li et al. (2025))**.** *For any $q \geq 2$, any even integer $s \geq 2$, and any two d-dimensional random vectors w, v, we have*

$$\Big\| |w+v|_s \Big\|_q^2 \leq \Big\| |w|_s \Big\|_q^2 + 2 \Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\left(|w|_s^{q-s} \sum_{j=1}^{d} w_j^{s-1} v_j \right) + (\max\{q, s\} - 1) \Big\| |v|_s \Big\|_q^2.$$

Lemma D.2 (Norm of sub-Gaussian vectors)**.** *If X is a mean-zero, sub-Gaussian random vector in $\mathbb{R}^d$ with $\|X\|_{\psi_2} = K$. Let B be an $d \times k$ matrix. Then for any $t > 0$, we have*

$$\mathbb{P}(|B^\top X|_2 \geq CK\|B\|_F + t) \leq \exp\Big(- \frac{ct^2}{K^2 \|B\|_{\mathrm{op}}^2} \Big),$$

where $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ denote the Frobenius norm, and operator norm respectively, c, C are absolute constants.

Proof. See Exercises 6.2.6 and 6.3.5 in Vershynin (2018). □

Lemma D.3 (Rosenthal Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ independent random variables with $\mathbb{E}X_i = 0$, and $\mathbb{E}|X_i|^q < \infty$ for $q \geq 2$. Then*

$$\mathbb{E}|\sum_{i=1}^{n} X_i|^q \leq C_q \Big(\sum_{i=1}^{n} \mathbb{E}|X_i|^q + (\sum_{i=1}^{n} \mathbb{E}|X_i|^2)^{\frac{q}{2}} \Big),$$

where C_q is constant only depending on q.

Lemma D.4. *Let $Y \geq 0$ be a random variable. Let $B > 0$ and $\beta > 0$. Assume that for every $\delta \in (0, 1)$ satisfying $\beta\delta < 1$,*

$$\mathbb{P}\left(Y > B\sqrt{\log\Big(\frac{1}{\beta\delta}\Big)} \right) \leq \delta. \tag{12}$$

Then for any $q \geq 2$,

$$\|Y\|_q \leq B \left(\sqrt{\log_+\Big(\frac{1}{\beta}\Big)} + \sqrt{\frac{q}{2}}\, \beta^{-1/q} \right). \tag{13}$$

Proof. Define $t_0 := B\sqrt{\log_+\left(\frac{1}{\beta}\right)}$ and

$$\delta(t) := \beta^{-1}\exp\left(-\frac{t^2}{B^2}\right) \in (0,1), \text{ for } t \geq t_0.$$

We also note that $\beta\delta(t) = \exp(-\frac{t^2}{B^2}) < 1$ when $t \geq t_0$. Thus $\delta(t)$ lies in the admissible range. Plugging $\delta(t)$ into (12) yields the tail bound

$$\mathbb{P}(Y > t) \leq \delta(t) = \beta^{-1}\exp\left(-\frac{t^2}{B^2}\right), \qquad t \geq t_0. \tag{14}$$

For $t < t_0$, we trivially have $\mathbb{P}(Y > t) \leq 1$. Now we consider the L_q-norm,

$$\mathbb{E}[Y^q] = \int_0^\infty qt^{q-1}\mathbb{P}(Y > t)\,dt \leq \int_0^{t_0} qt^{q-1}\,dt + \beta^{-1}\int_{t_0}^\infty qt^{q-1}e^{-t^2/B^2}\,dt.$$

The first integral equals t_0^q. For the second one, change variables $u = t^2/B^2$:

$$\begin{aligned}\beta^{-1}\int_{t_0}^\infty qt^{q-1}e^{-t^2/B^2}\,dt &\leq \beta^{-1}\int_0^\infty qt^{q-1}e^{-t^2/B^2}\,dt \\ &= \beta^{-1}\,q\,\frac{B^q}{2}\int_0^\infty u^{\frac{q}{2}-1}e^{-u}\,du \\ &= \beta^{-1}\,\frac{q}{2}\,B^q\,\Gamma\Big(\frac{q}{2}\Big).\end{aligned}$$

Hence,

$$\mathbb{E}[Y^q] \leq t_0^q + \beta^{-1}\,\frac{q}{2}\,B^q\,\Gamma\Big(\frac{q}{2}\Big).$$

Taking q-th roots and using $(x+y)^{1/q} \leq x^{1/q} + y^{1/q}$ for $q \geq 1$ yields

$$\|Y\|_q \leq t_0 + B\left(\beta^{-1}\,\frac{q}{2}\,\Gamma\Big(\frac{q}{2}\Big)\right)^{1/q} \leq t_0 + B\sqrt{\frac{q}{2}}\,\beta^{-1/q} \tag{15}$$

where the second inequality is due to $\Gamma(x) \leq x^{x-1}$ for $x \geq 1/2$. We complete the proof by substituting the definition of t_0 in (15). □

Lemma D.5 (Embedding between l_s and l_2)**.** *Let $k \in \mathbb{N}$ and $z \in \mathbb{R}^k$. For any $s \in [1,\infty]$,*

$$|z|_s \;\leq\; k^{\left(\frac{1}{s}-\frac{1}{2}\right)_+}|z|_2, \qquad x_+ := \max\{x, 0\}. \tag{16}$$

Specifically, when $s \geq 2$, we have $|z|_s \;\leq\; |z|_2$

Let $V, Q \in \mathbb{R}^{d\times k}$ to be matrices with orthogonal columns. Suppose the singular values of $V^\top Q$ are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$. Then we call

$$\Theta(Q,V) = \operatorname{diag}\left(\cos^{-1}(\sigma_1), \cos^{-1}(\sigma_2), \ldots, \cos^{-1}(\sigma_k)\right) \tag{17}$$

as the principal angles. Lemma D.6 provides some convenient expressions of $\sin\Theta$ distance, and shows $\operatorname{dist}(Q,V) := \|VV^\top - QQ^\top\|$ distance is equivalent to $\sin\Theta$ distance.

Lemma D.6 ($\sin\Theta$ Distance (Lemma 1 of Cai & Zhang (2018)))**.** *Let $V, Q \in \mathbb{R}^{d\times k}$ to be matrices with orthogonal columns. Then*

$$\|\sin\Theta(Q,V)\|_{\mathrm{op}} = \sqrt{1 - \sigma_{\min}^2(V^\top Q)},$$

and

$$\big\|\sin\Theta(Q,V)\big\|_{\mathrm{op}} \leq \operatorname{dist}(Q,V) = \big\|VV^\top - QQ^\top\big\|_{\mathrm{op}} \leq 2\big\|\sin\Theta(Q,V)\big\|_{\mathrm{op}}.$$

Let $R = \arg\min_{R\in\mathbb{O}_k}\big\|Q - VR\big\|_F$, and $D_F(Q,V) = \|Q - VR\|_F$, then

$$\big\|\sin\Theta(Q,V)\big\|_F \;\leq\; D_F(Q,V) \;\leq\; \sqrt{2}\,\big\|\sin\Theta(Q,V)\big\|_F.$$

Now, let $A \in \mathbb{R}^{k\times k}$ be a non-singular matrix, then there exist a unique Q with orthonormal column vectors and semi-definite H such that

$$A = QH \tag{18}$$

The decomposition (18) is called the polar decomposition of A. In fact, we have

$$Q = A(A^\top A)^{-\frac{1}{2}}, \quad H = (A^\top A)^{\frac{1}{2}}.$$

Lemma D.7 (Theorem 1 of Li (1995))**.** *Let A_1 and A_2 be two $n \times n$ non-singular complex matrices whose polar decompositions are given by* (18)*. Then*

$$\|Q_1 - Q_2\|_{\mathrm{op}} \le \frac{2}{\sigma_{min}(A_1) + \sigma_{min}(A_2)} \|A_1 - A_2\|_{\mathrm{op}}. \tag{19}$$

D.2 Proofs for Proposition 2.2 in Section 2.1

Suppose the factor model in (2) holds. Define $\Sigma := \mathrm{Cov}(x)$, we write $V \in \mathbb{R}^{d\times k}$ to be the orthogonal matrix whose columns are the leading k eigenvectors of Σ. The following proposition shows that the column space of V is close to column space of B.

Lemma D.8. *Define the normalized loading matrix*

$$B_0 := d^{-1/2}B \in \mathbb{R}^{d\times k}, \quad \textit{such that } B_0^\top B_0 = I_k.$$

Suppose factor model (2) *satisfies Assumption 2.1, then there exists $O_\star \in \mathbb{O}_k$ such that*

$$\|B_0 - VO_\star\|_F \le \sqrt{2}\,\|\sin\Theta(B_0, V)\|_F \le C\,\frac{K_u^2}{\lambda_k(\Sigma_f)} d^{-\frac{1}{2}}, \tag{20}$$

where C is an absolute constant.

Proof. Denote $\Sigma_u := \mathrm{Cov}(u)$ and $\Sigma_f := \mathrm{Cov}(f)$. From $x = Bf+u$, $\mathbb{E}f = 0$, and the independence of f and u, we have

$$\Sigma = B\Sigma_f B^\top + \Sigma_u,$$

We denote the signal covariance as $\Sigma_s := B\Sigma_f B^\top = d\,B_0\Sigma_f B_0^\top$, so that $\Sigma = \Sigma_s + \Sigma_u$.

Since Σ_f is symmetric positive definite with $\lambda_k(\Sigma_f) > 0$, the matrix $B_0\Sigma_f B_0^\top$ has rank k and its nonzero eigenvalues are exactly $\{\lambda_j(\Sigma_f)\}_{j=1}^k$, while the remaining eigenvalues are 0. Consequently,

$$\lambda_k(\Sigma_s) = d\,\lambda_k(\Sigma_f), \qquad \lambda_{k+1}(\Sigma_s) = 0,$$

and the eigengap equals

$$\delta_k := \lambda_k(\Sigma_s) - \lambda_{k+1}(\Sigma_s) = d\,\lambda_k(\Sigma_f) > 0.$$

Let $V \in \mathbb{R}^{d\times k}$ collect the leading k eigenvectors of Σ. By the Davis–Kahan $\sin\Theta$ theorem (Theorem 2 in Yu et al. (2015)), it holds that

$$\|\sin\Theta(B_0, V)\|_F \le \frac{\|\Sigma - \Sigma_s\|_F}{\delta_k} = \frac{\|\Sigma_u\|_F}{d\,\lambda_k(\Sigma_f)}. \tag{21}$$

By definition of the sub-Gaussian norm of a random vector, $\|u\|_{\psi_2} = \sup_{\|v\|_2=1}\|v^\top u\|_{\psi_2} = K_u$. Thus, for any unit vector v, the random variable $v^\top u$ is also sub-Gaussian with $\|v^\top u\|_{\psi_2} \le K_u$. The standard moment bound for sub-Gaussian random variables implies

$$\|\Sigma_u\|_{\mathrm{op}} = \sup_{\|v\|_2=1} v^\top\Sigma_u v \le \sup_{\|v\|_2=1} \mathbb{E}[(v^\top u)^2] \le CK_u^2, \tag{22}$$

It then implies

$$\|\Sigma_u\|_F \le d^{\frac{1}{2}}\|\Sigma_u\|_{\mathrm{op}} \le CK_u^2 d^{\frac{1}{2}} \tag{23}$$

At the same time, Lemma D.6 implies

$$\inf_{O\in\mathbb{O}_k}\|B_0 - VO\|_F \le \sqrt{2}\,\|\sin\Theta(B_0, V)\|_F. \tag{24}$$

We complete the proof by combining (21), (23) and (24). □

Lemma D.9. *For any $Q \in \mathbb{R}^{d\times k}$ consisting of orthonormal columns, suppose random vector $u_i \in \mathbb{R}^d$ is sub-Gaussian with $\|u_i\|_{\psi_2} \le K_u$, and $s \ge 2$, then*

$$\Big\| |d^{-\frac{1}{2}} Q^\top u_i|_s \Big\|_q \le K_u (k^{\frac{1}{2}} + C\, q^{\frac{1}{2}}) \cdot d^{-\frac{1}{2}},$$

where c, C are absolute constants.

Proof. Since $Q \in \mathbb{R}^{d\times k}$ has orthonormal columns, it satisfies $\|Q\|_F = k^{\frac{1}{2}}$ and $\|Q\|_{\mathrm{op}} = 1$. Applying Lemma D.2, it holds that

$$\mathbb{P}\Big(|Q^\top u_i|_2 \ge C\, K_u\, k^{\frac{1}{2}} + t \Big) \le \exp\Big(- \frac{ct^2}{K_u^2} \Big) \tag{25}$$

So that random variable $|Q^\top u_i|_2$ is sub-Gaussian. By the property of sub-Gaussian, the moment satisfies

$$\Big\| |Q^\top u_i|_2 \Big\|_q \le c\, K_u k^{\frac{1}{2}} + C K_u q^{\frac{1}{2}}, \tag{26}$$

where c, C are absolute constants. Next, followed by Lemma D.5 and $s \ge 2$, we have

$$\Big\| |Q^\top u_i|_s \Big\|_q \le c\, K_u k^{\frac{1}{2}} + C K_u\, q^{\frac{1}{2}},$$

The proof can be finished by multiplying factor $d^{-\frac{1}{2}}$ on both sides. □

Proof of Proposition 2.2. Let $O_* := \arg\min_{O\in\mathbb{O}_k} \|B_0 - VO\|_F$ as defined in Lemma D.8, where $B_0 := d^{-\frac{1}{2}} B$. We then define $R := \arg\min_{O\in\mathbb{O}_k} \|VO_* - QR\|_F$.

Using $x = d^{1/2} B_0 f + u$ and $\hat{f} = d^{-1/2} Q^\top x$,

$$\hat{f} = d^{-1/2} Q^\top (d^{1/2} B_0 f + u) = Q^\top B_0 f + d^{-1/2} Q^\top u,$$

Hence

$$\hat{f} - Rf = d^{-1/2} Q^\top u + (Q^\top B_0 - R) f = d^{-1/2} Q^\top u + (Q^\top B_0 - Q^\top V O_\star) f + (Q^\top V O_\star - R) f$$

Applying triangular inequality gives

$$\Big\| |\hat{f} - Rf|_s \Big\|_q \le \Big\| |d^{-1/2} Q^\top u|_s \Big\|_q + \Big\| |(Q^\top B_0 - Q^\top V O_\star) f|_s \Big\|_q + \Big\| |(Q^\top V O_\star - R) f|_s \Big\|_q. \tag{27}$$

Lemma D.9 provides the error bound of the first term of (27), i.e.

$$\Big\| |d^{-\frac{1}{2}} Q^\top u_i|_s \Big\|_q \le K_u d^{-\frac{1}{2}} \cdot (c\, k^{\frac{1}{2}} + C\, q^{\frac{1}{2}}), \tag{28}$$

where c, C are absolute constants.

Note that Q is defined to have orthonormal columns. Applying Lemma D.8, Lemma D.5 and the relationship $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, we have

$$\Big\| |(Q^\top B_0 - Q^\top V O_\star) f|_s \Big\|_q \le \Big\| \|(B_0 - VO_\star)\|_{\mathrm{op}} |f|_2 \Big\|_q \le C \frac{K_u^2}{\lambda_k(\Sigma_f)} \Big\| |f|_2 \Big\|_q d^{-\frac{1}{2}} \tag{29}$$

It remains to bound the second term $\Big\| |(Q^\top V O_\star - R) f|_s \Big\|_q$ on RHS of (27). By Lemma D.5, we only need to focus on $\Big\| |(Q^\top V O_\star - R) f|_2 \Big\|_q$.

Utilizing Lemma D.6, the independence of (Q, R) and f, and the relationship $\|A\|_{\mathrm{op}} \le \|A\|_F \le k^{\frac{1}{2}} \|A\|_{\mathrm{op}}$ for $A \in \mathbb{R}^{d\times k}, k \le d$, we have

$$\begin{aligned}
\Big\| |(Q^\top V O_\star - R) f|_s \Big\|_q &\le \Big\| \|Q^\top V O_\star - R\|_{\mathrm{op}} \Big\|_q \Big\| |f|_2 \Big\|_q \le \Big\| \|Q^\top V O_\star - R\|_F \Big\|_q \Big\| |f|_2 \Big\|_q \\
&\le \Big\| \|VO_\star - QR\|_F \Big\|_q \Big\| |f|_2 \Big\|_q \le \sqrt{2} \Big\| \|\sin\Theta(Q, VO_*)\|_F \Big\|_q \Big\| |f|_2 \Big\|_q \\
&\le \sqrt{2} k^{\frac{1}{2}} \Big\| \|\sin\Theta(Q, VO_*)\|_{\mathrm{op}} \Big\|_q \Big\| |f|_2 \Big\|_q \le \sqrt{2} k^{\frac{1}{2}} \Big\| \mathrm{dist}(Q, VO_\star) \Big\|_q \Big\| |f|_2 \Big\|_q \\
&= \sqrt{2} k^{\frac{1}{2}} \Big\| \mathrm{dist}(Q, V) \Big\|_q \Big\| |f|_2 \Big\|_q
\end{aligned} \tag{30}$$

where $\mathrm{dist}(Q, VO_\star) = \mathrm{dist}(Q, V)$ since $O_k \in \mathbb{R}^{k\times k}$ is orthonormal matrix.

Combining (28), (29), (30) and (27), we obtain

$$\left\| |\hat{f} - Rf|_s \right\|_q \le \left(C \frac{K_u^2}{\lambda_k(\Sigma_f)} \left\| |f|_2 \right\|_q + K_u (c\, k^{\frac{1}{2}} + C\, q^{\frac{1}{2}}) \right) \cdot d^{-\frac{1}{2}} + \sqrt{2} k^{\frac{1}{2}} \left\| \mathrm{dist}(Q, V) \right\|_q \left\| |f|_2 \right\|_q \tag{31}$$

Now we complete the proof. □

D.3 Proofs for Section 2.3

In this section, we show that the additional alignment step introduced in Stage II of Algorithm 1 does not affect the theoretical bounds established in Theorem 3 of Huang et al. (2021).

Now let's recall the Oja's iteration. Let $\{A_t\}_{t\ge 1} \subset \mathbb{R}^{d\times d}$ be a sequence of matrices and $\{\eta_t^o\}_{t\ge 1}$ be step sizes. Initialize with any $Z_0 \in \mathbb{R}^{d\times k}$ of full column rank and set $Q_0 = \mathrm{QR}[Z_0]$, where $\mathrm{QR}[\cdot]$ denotes the Q-factor of the thin QR decomposition. For $t \ge 1$, we set

$$Q_t' := \mathrm{QR}[(I + \eta_t^{\mathrm{oja}} A_t)\, Q_{t-1}]. \tag{32}$$

Once we have Q_t', we perform an orthogonal Procrustes alignment to fix the rotational ambiguity of Q's. Specifically, we compute $S_t := Q_t'^\top Q_{t-1}$, and take SVD $S_t := U_t \Lambda_t V_t^\top$. We then set $Q_t := Q_t' U_t V_t^\top$. Equivalently, Q_t is the unique solution to the orthogonal Procrustes problem

$$Q_t := Q_t' O_t, \quad O_t = \arg\min_{O\in\mathbb{O}_k} \|Q_{t-1} - Q_t' O\|_F, \tag{33}$$

where O_t has closed form solution, $O_t = S_t (S_t^\top S_t)^{-\frac{1}{2}} = U_t V_t^\top$.

Lemma D.10. *If $\sup_{P\in\mathbb{O}_{d,k}} \|P^\top (A_t - \Sigma)\|_{\mathrm{op}} \le M$ almost surely, Let $\lambda_1 \ge \cdots \ge \lambda_d$ be the eigenvalues of Σ, and let $\rho_k = \lambda_k - \lambda_{k+1}$. For any $\delta \in (0,1)$, define*

$$T_0 = \Theta\left(\frac{kM^2}{\delta^2 \rho_k^2} \log^4(\frac{dM}{\rho_k \delta}) \right),\ \beta = \Theta\left(\frac{M^2}{\rho_k^2} \log(\frac{Mk}{\rho_k \delta}) \right),$$

$$\eta_t^{\mathrm{oja}} = \begin{cases} \Theta\left(\frac{\log(d/\delta)}{\rho_k T_0}\right), & t \le T_0 \\ \Theta\left(\frac{1}{\rho_k(\beta + t - T_0)}\right), & t > T_0. \end{cases}$$

Suppose Q_t follows the updating rule (32) *and* (33)*, then for any $T > T_0$, Q_T satisfies*

$$\mathrm{dist}(Q_T, V) \le C' \frac{M}{\rho_k} \sqrt{\frac{\log(Mk/\rho_k\delta)}{T - T_0}}$$

with probability at least $1 - \delta$, where C' is an absolute positive constant.

Proof. Using the same notation of Huang et al. (2021), we write $V \in \mathbb{R}^{d\times k}$ for the orthogonal matrix whose columns are the k leading eigenvectors of Σ, and $U \in \mathbb{R}^{d\times(d-k)}$ for the orthogonal matrix whose columns are the remaining eigenvectors. Let $Z_t := Y_t Y_{t-1} \cdots Y_1 Z_0$, where $Y_i = I_k + \eta_i A_i$, it suffices to prove

$$\mathrm{dist}(Q_t, V) \le \|W_t\|_{\mathrm{op}}, \quad W_t := U^\top Z_t (V^\top Z_t)^{-1}$$

The remaining proof are identical to Huang et al. (2021). Define $B_t := (R_t')^{-1} O_t$, we first claim that

$$Q_t = Y_t Y_{t-1} \cdots Y_1 Q_0 B_1 \cdots B_{t-1} B_t, t \ge 0, \tag{34}$$

Equation (34) holds for $t = 0$. We prove the claim by mathematical induction on t. Assume that (34) holds at time $t-1$, i.e., it holds for Q_{t-1}. We now show that (34) also holds for Q_t.

Since $Q_t' = QR[Y_t Q_{t-1}]$ has full column rank, there exists $R_t' \in \mathbb{R}^{k\times k}$ upper triangular and non-singular, such that $Q_t' R_t' = Y_t Q_{t-1}$, then

$$Q_t' = Y_t Q_{t-1} (R_t')^{-1} \tag{35}$$

Followed by (33), it follows that

$$Q_t = Q_t' O_t = Y_t Q_{t-1} (R_t')^{-1} O_t = Y_t Q_{t-1} B_t = Y_t Y_{t-1} \cdots Y_1 Q_0 B_1 \cdots B_{t-1} B_t, \tag{36}$$

Therefore, the claim holds for all Q_t. Recall that $Q_0 = QR[Z_0]$, there exist R_0 such that $Q_0 R_0 = Z_0$, we can write (34) as

$$Q_t = Z_t P_t, \quad P_t := R_0^{-1} B_1 \cdots B_t \in \mathbb{R}^{k \times k} \text{ non-singular} \tag{37}$$

We also note that $\|V^\top Q_t\|_{\mathrm{op}} \le 1$, then we complete the proof by showing that

$$\begin{aligned} \mathrm{dist}(Q_t, V) &= \|U^\top Q_t\|_{\mathrm{op}} \le \|U^\top Q_t (V^\top Q_t)^{-1}\|_{\mathrm{op}} \cdot \|V^\top Q_t\|_{\mathrm{op}} \le \|U^\top Q_t (V^\top Q_t)^{-1}\|_{\mathrm{op}} \\ &= \|U^\top Z_t P_t (V^\top Z_t P_t)^{-1}\|_{\mathrm{op}} \\ &= \|U^\top Z_t (V^\top Z_t)^{-1}\|_{\mathrm{op}} \end{aligned} \tag{38}$$

□

D.4 Proof for Lemma 2.4 in Section 2.3

By adapting the proof of Huang et al. (2021), we first establish an L_q-moment bound for the Oja update. Let

$$Y_i := I + \eta_i^{\mathrm{oja}} A_i, \qquad Z_t := Y_t Y_{t-1} \cdots Y_1 Q_0.$$

We write $V \in \mathbb{R}^{d \times k}$ and $U \in \mathbb{R}^{d \times (d-k)}$ be the orthogonal matrices whose columns are respectively the top k and remaining $d-k$ eigenvectors of Σ. Following Huang et al. (2021), we define

$$W_t := U^\top Z_t (V^\top Z_t)^{-1}$$

We then define the good events as

$$G_0 := \{\|W_0\|_{\mathrm{op}} \le 1\}, \qquad G_t := G_{t-1} \cap \{\|W_t\|_{\mathrm{op}} \le \sqrt{2e}\}, \qquad t \ge 1. \tag{39}$$

The following lemma gives the complete statement of Lemma 2.4.

Lemma D.11 (L^q-Moment Bound for Oja Update)**.** *Suppose Assumption 2.3 holds. Set* $\eta_t^{\mathrm{oja}} = \frac{\alpha}{(\beta+t)\rho_k}$ *and define the normalized gap*

$$\bar{\rho}_k := \min\left\{\frac{\rho_k}{M}, \frac{\rho_k}{\|\Sigma\|_{\mathrm{op}}}, 1\right\}.$$

If

$$\alpha \ge 8, \qquad \beta \ge \max\left\{\frac{4(1+\sqrt{2e})\alpha}{\bar{\rho}_k},\, C \cdot \frac{\alpha^2}{\bar{\rho}_k^2}\right\},$$

where C is an absolute constant. Then for every $q \ge 2$ and $t \ge 1$, it holds that

$$\|\mathrm{dist}(Q_t, V)\|_q \le k^{1/q} \left[\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} + C \cdot \frac{\alpha q^{1/2}}{\bar{\rho}_k} \cdot \frac{t^{1/2}}{\beta+t} + C^{2/q} \cdot \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{-\frac{\beta \bar{\rho}_k^2}{\alpha^2 C^2 q}\right\}\right].$$

Additionally, for any $t \ge 1$,we have

$$\mathbb{P}(G_t^c) \le k C^2 \alpha^2 \bar{\rho}_k^{-2} \exp\left\{-\frac{\beta \bar{\rho}_k^2}{C^2 \alpha^2}\right\}. \tag{40}$$

Proof. By Lemma 6 of Huang et al. (2021), it holds that

$$\mathrm{dist}(Q_t, V) = \|U^\top Q_t\|_{\mathrm{op}} \le \|U^\top Q_t (V^\top Q_t)^{-1}\|_{\mathrm{op}} = \|U^\top Z_t (V^\top Z_t)^{-1}\|_{\mathrm{op}} = \|W_t\|_{\mathrm{op}}. \tag{41}$$

We now verify that $\|W_0\|_{\mathrm{op}} < 1$ holds almost surely. Since

$$\mathrm{dist}(Q_0, V) = \|U^\top Q_0\|_{\mathrm{op}} \le \frac{1}{2},$$

we have

$$(V^\top Q_0)^\top (V^\top Q_0) = Q_0^\top V V^\top Q_0 = I_k - Q_0^\top U U^\top Q_0.$$

Therefore

$$\sigma_{\min}(V^\top Q_0)^2 = 1 - \|U^\top Q_0\|_{\mathrm{op}}^2 \geq 1 - \frac{1}{4} = \frac{3}{4},$$

and hence

$$\sigma_{\min}(V^\top Q_0) \geq \frac{\sqrt{3}}{2}.$$

It follows that

$$\|W_0\|_{\mathrm{op}} = \|U^\top Q_0 (V^\top Q_0)^{-1}\|_{\mathrm{op}} \leq \|U^\top Q_0\|_{\mathrm{op}} \cdot \|(V^\top Q_0)^{-1}\|_{\mathrm{op}} \leq \frac{1}{2} \cdot \frac{2}{\sqrt{3}} = \frac{1}{\sqrt{3}} < 1.$$

By Proposition 11 of Huang et al. (2021), for every $p \geq 2$, we have

$$\|W_t \mathbf{1}_{G_t}\|_{p,p}^2 \leq k^{2/p} \left(\frac{\beta+1}{\beta+t}\right)^\alpha + p k^{2/p} \left(\frac{C\alpha}{\bar{\rho}_k}\right)^2 \frac{t}{(\beta+t)^2}, \tag{42}$$

where, for a random matrix X,

$$\|X\|_{p,p} := \left(\mathbb{E}\|X\|_{S_p}^p\right)^{1/p}, \qquad \|X\|_{S_p} := \left(\sum_{i \geq 1} \sigma_i(X)^p\right)^{1/p},$$

and $\{\sigma_i(X)\}_{i \geq 1}$ are the singular values of X. Here $\|\cdot\|_{S_p}$ denotes the Schatten-p norm, and $\|\cdot\|_{p,p}$ denotes the corresponding L_p-norm of this random matrix norm.

Now we decompose

$$\|\operatorname{dist}(Q_t, V)\|_q \leq \|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t}\|_q + \|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t^c}\|_q. \tag{43}$$

First consider the good-event term. By (41) and $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_{Sq}$,

$$\|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t}\|_q \leq \|W_t \mathbf{1}_{G_t}\|_{\mathrm{op}, L_q} \leq \|W_t \mathbf{1}_{G_t}\|_{S_q, L_q} = \|W_t \mathbf{1}_{G_t}\|_{q,q}. \tag{44}$$

Applying (42) with $p = q$, we get

$$\|W_t \mathbf{1}_{G_t}\|_{q,q}^2 \leq k^{2/q} \left(\frac{\beta+1}{\beta+t}\right)^\alpha + q k^{2/q} \left(\frac{C\alpha}{\bar{\rho}_k}\right)^2 \frac{t}{(\beta+t)^2}.$$

Taking square roots and using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$,

$$\|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t}\|_q \leq k^{1/q} \left[\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} + \frac{C\alpha\sqrt{q\,t}}{\bar{\rho}_k(\beta+t)}\right]. \tag{45}$$

Next consider the bad-event term. Since

$$0 \leq \operatorname{dist}(Q_t, V) \leq 1,$$

we have

$$\|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t^c}\|_q \leq \mathbb{P}(G_t^c)^{1/q}. \tag{46}$$

Following the proof of Theorem 12 of Huang et al. (2021), it holds that

$$\mathbb{P}(G_t^c) \leq k C^2 \alpha^2 \bar{\rho}_k^{-2} \exp\left\{-\frac{\beta \bar{\rho}_k^2}{C^2\alpha^2}\right\}. \tag{47}$$

By (46) and (47), we then have

$$\|\operatorname{dist}(Q_t, V)\mathbf{1}_{G_t^c}\|_q \leq C^{2/q} k^{1/q} \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{-\frac{\beta\bar{\rho}_k^2}{C^2\alpha^2 q}\right\} \tag{48}$$

Finally, combining (43), (45), and (48), we obtain

$$\|\operatorname{dist}(Q_t, V)\|_q \leq k^{1/q} \left[\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} + C \cdot \frac{\alpha q^{1/2}}{\bar{\rho}_k} \cdot \frac{t^{1/2}}{\beta+t} + C^{2/q} \cdot \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{-\frac{\beta\bar{\rho}_k^2}{\alpha^2 C^2 q}\right\}\right].$$

□

Remark D.12. By taking

$$\beta \geq \frac{\alpha^2 C^2 q}{\bar{\rho}_k^2}\left[\frac{1}{2}\log d + \frac{1}{q}\log\left(\frac{k C^2 \alpha^2}{\bar{\rho}_k^2}\right)\right] = \Theta\Big(\bar{\rho}_k^{-2}\big(q \log d + \log \frac{k}{\bar{\rho}_k^2}\big)\Big),$$

the last term has the order of $\mathcal{O}(d^{-1/2})$.

D.5 Proofs for Proposition 3.5 in Section 3

Lemma D.13 (Stability of θ^*)**.** *Let $R_1, R_2 \in \mathbb{R}^{k\times k}$ be orthogonal matrices (possibly random), and suppose (f, y) is independent of (R_1, R_2). For $i = 1, 2$, define θ_i^* as*

$$F_i(\theta) := \mathbb{E}\big[l(\theta; R_i f, y) \,\big|\, R_i\big], \qquad \theta_i^* := \arg\min_\theta F_i(\theta).$$

Under Assumption 3.1 and Assumption 3.3, we have

$$\Big\| \big|\theta_1^* - \theta_2^*\big|_s \Big\|_q \;\le\; \frac{L_{s,f}}{\mu} \Big\| \big|(R_1 - R_2)f\big|_s \Big\|_q.$$

Proof. Note that that F_1 is l^s-strong convex with parameter μ. i.e. for all θ_1, θ_2,

$$\big\langle (\theta_1 - \theta_2)^{\circ(s-1)},\, \nabla F_1(\theta_1) - \nabla F_1(\theta_2) \big\rangle \ge \mu\, |\theta_1 - \theta_2|_s^s. \tag{49}$$

Then, we apply (49) with $(\theta_1, \theta_2) = (\theta_1^*, \theta_2^*)$. Note that $\nabla F_1(\theta_1^*) = 0$, it holds that

$$\mu\, |\theta_1^* - \theta_2^*|_s^s \le \big\langle (\theta_1^* - \theta_2^*)^{\circ(s-1)},\, \nabla F_1(\theta_1^*) - \nabla F_1(\theta_2^*) \big\rangle = -\big\langle (\theta_1^* - \theta_2^*)^{\circ(s-1)},\, \nabla F_1(\theta_2^*) \big\rangle.$$

Applying Hölder's inequality with the parameters s and $s/(s-1)$ yields

$$\mu\, |\Delta|_s^s \le \big|\big\langle \Delta^{\circ(s-1)},\, \nabla F_1(\theta_2^*) \big\rangle\big| \le \big|\Delta^{\circ(s-1)}\big|_{s/(s-1)}\, \big|\nabla F_1(\theta_2^*)\big|_s = |\Delta|_s^{s-1}\, \big|\nabla F_1(\theta_2^*)\big|_s,$$

where $\Delta := \theta_1^* - \theta_2^*$ for notation brevity. We only consider $|\Delta|_s \neq 0$, otherwise it's trivial. Dividing both sides by $|\Delta|_s^{s-1}$ gives

$$|\theta_1^* - \theta_2^*|_s \le \frac{1}{\mu} \big|\nabla F_1(\theta_2^*)\big|_s = \frac{1}{\mu} \big|\nabla F_1(\theta_2^*) - \nabla F_2(\theta_2^*)\big|_s, \tag{50}$$

where the last equality holds by definition of θ_2^*.

Applying Jensen's inequality, it holds that

$$\big|\nabla F_1(\theta_2^*) - \nabla F_2(\theta_2^*)\big|_s \le \mathbb{E}\big[\big|\nabla_\theta l(\theta_2^*; R_1 f, y) - \nabla_\theta l(\theta_2^*; R_2 f, y)\big|_s \,\big|\, R_1, R_2\big],$$

where we can condition on (R_1, R_2), since F_1 depends only on R_1 not on R_2, and vice versa.

By applying condition Jensen inequality, $\nabla_\theta F_i(\theta_i^*) = 0$, and utilizing Assumption 3.3, we obtain

$$\begin{aligned}
\Big\| \big|\nabla F_1(\theta_2^*) - \nabla F_1(\theta_1^*)\big|_s \Big\|_q &\le \mathbb{E}^{\frac{1}{q}}\Big[\mathbb{E}^q\big(\big|\nabla_\theta l(\theta_2^*; R_1 f, y) - \nabla_\theta l(\theta_2^*; R_2 f, y)\big|_s \,\big|\, R_1, R_2\big)\Big] \\
&\le \mathbb{E}^{\frac{1}{q}}\Big[\mathbb{E}\big(\big|\nabla_\theta l(\theta_2^*; R_1 f, y) - \nabla_\theta l(\theta_2^*; R_2 f, y)\big|_s^q \,\big|\, R_1, R_2\big)\Big] \\
&= \mathbb{E}^{\frac{1}{q}}\Big[\big|\nabla_\theta l(\theta_2^*; R_1 f, y) - \nabla_\theta l(\theta_2^*; R_2 f, y)\big|_s^q\Big] \\
&= \Big\| \big|\nabla_\theta l(\theta_2^*; R_1 f, y) - \nabla_\theta l(\theta_2^*; R_2 f, y)\big|_s \Big\|_q \\
&\le L_{s,f} \Big\| |(R_1 - R_2)f|_s \Big\|_q
\end{aligned} \tag{51}$$

Finally, we complete the proof by combining (50) and (51):

$$\Big\| \big|\theta_1^* - \theta_2^*\big|_s \Big\|_q \le \frac{1}{\mu} \Big\| \big|\nabla F_1(\theta_2^*) - \nabla F_1(\theta_1^*)\big|_s \Big\|_q \;\le\; \frac{L_{s,f}}{\mu} \Big\| |(R_1 - R_2)f|_s \Big\|_q.$$

□

Lemma D.14 (Procrustes Lipschitz Bound)**.** *Let V, U be the matrices of the leading and remaining eigenvectors of Σ. For $Q \in \mathcal{O}_{d,k}$, define*

$$W(Q) := U^\top Q (V^\top Q)^{-1}$$

whenever $V^\top Q$ is invertible. For $i = 1, 2$, let

$$R_i := \arg\min_{R \in \mathbb{O}_k} \|Q_i - VR\|_F.$$

If

$$\|W(Q_i)\|_{\mathrm{op}} \le \kappa, \qquad i = 1, 2,$$

then

$$\|R_1 - R_2\|_{\mathrm{op}} \le C_\kappa \|V^\top (Q_1 - Q_2)\|_{\mathrm{op}} \le C_\kappa \|Q_1 - Q_2\|_{\mathrm{op}},$$

where C_κ depends only on κ. In particular, for $\kappa = \sqrt{2e}$, C_κ is an absolute constant.

Proof. Let

$$A_i := V^\top Q_i, \qquad B_i := U^\top Q_i.$$

Since Q_i has orthonormal columns, i.e. $Q_i^\top Q_i = I_k$, it holds that

$$A_i^\top A_i + B_i^\top B_i = I_k.$$

Moreover, $B_i = W(Q_i) A_i$ by definition of $W(Q_i)$. Hence

$$A_i^\top (I_k + W(Q_i)^\top W(Q_i)) A_i = I_k.$$

For any $x \in \mathbb{R}^k$, we note that

$$\|x\|_2^2 = \|(I_k + W(Q_i)^\top W(Q_i))^{1/2} A_i x\|_2^2 \le (1 + \kappa^2) \|A_i x\|_2^2.$$

Therefore

$$\sigma_{\min}(A_i) \ge (1 + \kappa^2)^{-1/2}, \qquad i = 1, 2.$$

The Procrustes solution is the polar factor of A_i:

$$R_i = A_i (A_i^\top A_i)^{-1/2}.$$

By the polar-factor perturbation bound (Lemma D.7),

$$\|R_1 - R_2\|_{\mathrm{op}} \le \frac{2}{\sigma_{\min}(A_1) + \sigma_{\min}(A_2)} \|A_1 - A_2\|_{\mathrm{op}}.$$

Using the lower bound on $\sigma_{\min}(A_i)$,

$$\|R_1 - R_2\|_{\mathrm{op}} \le \sqrt{1 + \kappa^2} \|A_1 - A_2\|_{\mathrm{op}} = \sqrt{1 + \kappa^2} \|V^\top (Q_1 - Q_2)\|_{\mathrm{op}}.$$

This proves the lemma with $C_\kappa = \sqrt{1 + \kappa^2}$. □

Lemma D.15 (Theorem 3.1 of Stewart (1977))**.** *Let $A = QR$, where $A \in \mathbb{R}^{d \times k}$ has rank k and $Q \in \mathbb{R}^{d \times k}$ with $Q^\top Q = I_k$. Let $E \in \mathbb{R}^{d \times k}$ satisfies $\|A^\dagger\|_F \|E\|_F < \frac{1}{2}$. Let the linear operator $T : \mathbb{R}^{k \times k} \to \mathbb{R}^{k \times k}$ that maps the space of upper triangular matrices into the space of symmetric matrices be defined by*

$$T(F) := RF + F^\top R,$$

and define

$$F_1 := T^{-1}\left[R^\top (Q^\top E) + (E^\top Q) R\right].$$

Under the condition

$$\tau \|F_{1,t}\|_F < \frac{1}{4}, \qquad \tau = k \|R^{-1}\|_F \Big(1 + \frac{1}{\sqrt{2}} \kappa(R)\Big)$$

Then there exist matrices $W \in \mathbb{R}^{d \times k}, F \in \mathbb{R}^{k \times k}$, such that $A + E = (Q + W)(R + F)$, $(Q + W)^\top (Q + W) = I_k$ is the QR decomposition of A, with

$$\|W\|_F \le \frac{3\kappa(A) \frac{\|E\|_F}{\|A\|_{\mathrm{op}}}}{1 - 2\kappa(A) \frac{\|E\|_F}{\|A\|_{\mathrm{op}}}},$$

where $\kappa(A) := \|A\|_{\mathrm{op}} \|A^\dagger\|_{\mathrm{op}}$.

The next Lemma controls the one-step difference for Oja's iteration. Note that, we adopt the same notation as introduced above Lemma D.10.

Lemma D.16 (One-step Difference Bound for Online PCA iteration)**.** *Consider the update* (32) *and* (33)*. Assume that A_t is symmetric with $\|A_t\|_{\mathrm{op}} \le M_A$ a.s. , and $\eta_t^{\mathrm{oja}} < \frac{1}{4(\sqrt{2}+1)M_A}\, k^{-2}$. Then for all $t \ge 1$,*

$$\|Q_t - Q_{t-1}\|_F \le 6\eta_t^{\mathrm{oja}} k^{\frac{1}{2}} M_A, \quad \|Q_t' - Q_{t-1}\|_F \le 6\eta_t^{\mathrm{oja}} k^{\frac{1}{2}} M_A \tag{52}$$

The same upper bounds also hold for $\|\cdot\|_{\mathrm{op}}$-norm since $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$. Moreover, the same argument applies to the standard Oja iteration (4)*.*

Proof. It suffices to consider the update (32) and (33) and prove the perturbation bound for $Q_t' - Q_{t-1}$, since $\|Q_t - Q_{t-1}\|_F \le \|Q_t' - Q_{t-1}\|_F$. Note that $Q_{t-1} = Q_{t-1}I_k$, is the Q-factor of QR decomposition itself. We first verify the condition of Lemma D.15 with $A = Q_{t-1}, Q = Q_{t-1}, R = I_k, E = \eta_t^{\mathrm{oja}} A_t Q_{t-1}$ and $W = Q_t' - Q_{t-1}$. Now the linear operator T can be written as

$$T(F) = R^\top F + F^\top R = F + F^\top,$$

Then, by definition,

$$\begin{aligned} F_1 &= T^{-1}\left[Q_{t-1}^\top E + E^\top Q_{t-1}\right] \\ &= \eta_t^{\mathrm{oja}}\, T^{-1}\left[Q_{t-1}^\top A_t Q_{t-1} + \left(Q_{t-1}^\top A_t Q_{t-1}\right)^\top\right] \end{aligned} \tag{53}$$

Let

$$S := Q_{t-1}^\top A_t Q_{t-1}, \qquad B := S + S^\top.$$

Restricting F to be upper triangular, T admits the closed-form inverse $T^{-1}(B) = \mathrm{triu}(B) - \frac{1}{2}\,\mathrm{diag}(B)$ for any symmetric B, i.e.

$$(T^{-1}B)_{ij} = \begin{cases} B_{ij}, & i < j, \\ \dfrac{1}{2}B_{ii}, & i = j, \\ 0, & i > j. \end{cases}$$

Then we have

$$\|T^{-1}B\|_F^2 = \sum_{i<j} B_{ij}^2 + \frac{1}{4}\sum_i B_{ii}^2 \le \frac{1}{2}\|B\|_F^2 \le \|S\|_F^2 + \|S^\top\|_F^2 \le 2k\,\|A_t\|_{\mathrm{op}}^2,$$

Consequently,

$$\|F_1\|_F = \eta_t^{\mathrm{oja}}\|T^{-1}B\|_F \le \sqrt{2}\eta_t^{\mathrm{oja}} k^{\frac{1}{2}}\|A_t\|_{\mathrm{op}} \le \sqrt{2}M_A k^{\frac{1}{2}}\eta_t^{\mathrm{oja}} \tag{54}$$

By combining (53), (54) and under the assumption $\eta_t^{\mathrm{oja}} < \frac{1}{4(\sqrt{2}+1)M_A}\, k^{-2}$, we obtain that

$$\tau\|F_1\|_F = k^{\frac{3}{2}}\left(1 + \frac{1}{\sqrt{2}}\right)\|F_1\|_F < \frac{1}{4},$$

where τ is defined in Lemma D.15. Additionally, it follows that

$$\|A^\dagger\|_F\|E\|_F \le \eta_t^{\mathrm{oja}} k M_A < \frac{1}{4(\sqrt{2}+1)}k^{-1} < \frac{1}{2}$$

Now we apply Lemma D.15 to get

$$\|Q_t' - Q_{t-1}\|_F = \|W\|_F \le \frac{3\kappa(A)\frac{\|E\|_F}{\|A\|_{\mathrm{op}}}}{1 - 2\kappa(A)\frac{\|E\|_F}{\|A\|_{\mathrm{op}}}} \le 6\eta_t^{\mathrm{oja}} k^{\frac{1}{2}} M_A,$$

where the last inequality holds after plugging in the definitions

$$\kappa(A) = 1, \quad \|A\|_{\mathrm{op}} := \|Q_{t-1}\|_{\mathrm{op}} = 1, \quad \|E\|_F \le \eta_t^{\mathrm{oja}} k^{\frac{1}{2}} M_A < \frac{1}{2}$$

Lastly, since Q_t satisfies (33), it immediately follows that

$$\|Q_t - Q_{t-1}\|_F \le \|Q_t' - Q_{t-1}\|_F \le 6\eta_t^{\mathrm{oja}} k^{\frac{1}{2}} M_A$$

□

By combining Lemma D.13, D.14, and Lemma D.16, we are now able to prove a complete statement of proposition 3.5.

Proposition D.17 (Stability of θ^* under Oja's update)**.** *Suppose Assumptions 2.3, 3.1, 3.3 and 3.4 hold. Fix a finite horizon T with $\log T = \mathcal{O}(\log d)$, and choose the Stage-II Oja parameter*

$$\beta \geq C\bar{\rho}_k^{-2}\left[q\log d + \log\left(\frac{k}{\bar{\rho}_k^2}\right)\right], \qquad \alpha \geq 8 \tag{55}$$

Then, Lemma D.11 gives that

$$\mathbb{P}(G_T^c) \leq d^{-q/2}T^{-q},$$

where Oja good event G_T is defined in (39)*. For every $1 \leq t \leq T$, it holds that*

$$\left\| \left|\theta_t^* - \theta_{t-1}^*\right|_s \right\|_q \leq C\frac{L_{s,f}}{\mu}M_f\left(M_A k^{1/2}\eta_t^{\mathrm{oja}} + d^{-1/2}T^{-1}\right),$$

where $M_f := \||f|_2\|_q$. In particular, if $\eta_t^{\mathrm{oja}} \leq c_k/t$ for some c_k independent of t, then

$$\left\| \left|\theta_t^* - \theta_{t-1}^*\right|_s \right\|_q \leq C\frac{L_{s,f}}{\mu}M_f(M_A k^{1/2}c_k + d^{-1/2})t^{-1}.$$

Proof. Let $\mathbb{E}_z[\cdot]$ denote the expectation taken only w.r.t the random vectors (f, y). Recall that θ_t^* is the minimizer of $\mathbb{E}_z\Big[l(\theta; R_t f, y)\Big|R_t\Big]$. Lemma D.13 provides the following inequality

$$\left\| \left|\theta_t^* - \theta_{t-1}^*\right|_s \right\|_q \;\leq\; \frac{L_{s,f}}{\mu}\left\| \left|(R_t - R_{t-1})f\right|_s \right\|_q \leq \frac{L_{s,f}}{\mu}\Big\|\|R_t - R_{t-1}\|_{\mathrm{op}}\Big\|_q\Big\||f|_2\Big\|_q \tag{56}$$

On the good event G_T, where G_T defined in (39), we have

$$\|W_t\|_{\mathrm{op}} \leq \kappa, \qquad \|W_{t-1}\|_{\mathrm{op}} \leq \kappa, \qquad \kappa := \sqrt{2e}.$$

Therefore, by Lemma D.14,

$$\|R_t - R_{t-1}\|_{\mathrm{op}} \leq C_\kappa\|Q_t - Q_{t-1}\|_{\mathrm{op}}.$$

Followed by Lemma D.16, it holds on event G_T that

$$\|R_t - R_{t-1}\|_{\mathrm{op}} \leq CM_A k^{\frac{1}{2}}\eta_t^{oja},$$

where C is an absolute constant.

On G_T^c, since $R_t, R_{t-1} \in \mathbb{O}_k$,

$$\|R_t - R_{t-1}\|_{\mathrm{op}} \leq 2.$$

Therefore,

$$\|R_t - R_{t-1}\|_{\mathrm{op}} \leq CM_A k^{1/2}\eta_t^{\mathrm{oja}} + 2\cdot\mathbf{1}_{G_T^c}.$$

Taking L_q-norm gives

$$\Big\|\|R_t - R_{t-1}\|_{\mathrm{op}}\Big\|_q \leq CM_A k^{1/2}\eta_t^{\mathrm{oja}} + 2\mathbb{P}(G_T^c)^{1/q}.$$

By Lemma D.11, we have

$$\mathbb{P}(G_t^c) \leq kC^2\alpha^2\bar{\rho}_k^{-2}\exp\left\{-\frac{\beta\bar{\rho}_k^2}{C^2\alpha^2}\right\}. \tag{57}$$

By taking

$$\beta \geq C\bar{\rho}_k^{-2}\left[2q\log d + \log\left(\frac{k}{\bar{\rho}_k^2}\right)\right] \geq C\bar{\rho}_k^{-2}\left[q\log d + q\log T + \log\left(\frac{k}{\bar{\rho}_k^2}\right)\right] \tag{58}$$

We have

$$\mathbb{P}(G_T^c)^{1/q} \leq d^{-1/2}T^{-1}.$$

Hence

$$\Big\|\|R_t - R_{t-1}\|_{\mathrm{op}}\Big\|_q \leq CM_A k^{1/2}\eta_t^{\mathrm{oja}} + Cd^{-1/2}T^{-1}. \tag{59}$$

Combine with (59) and (56), we obtain

$$\left\| \left| \theta_t^* - \theta_{t-1}^* \right|_s \right\|_q \;\le\; C \frac{L_{s,f}}{\mu} M_f \left(M_A k^{1/2} \eta_t^{\mathrm{oja}} + d^{-1/2} T^{-1} \right),$$

If $\eta_t^{\mathrm{oja}} \le c_k / t$, then for $1 \le t \le T$,

$$d^{-1/2} T^{-1} \le d^{-1/2} t^{-1}.$$

Therefore

$$M_A k^{1/2} \eta_t^{\mathrm{oja}} + d^{-1/2} T^{-1} \le (M_A k^{1/2} c_k + d^{-1/2}) t^{-1}.$$

which proves that second claim. By absorbing the absolute constant in (58), we obtain the same statement as (55) □

Remark D.18. When we set $\eta_t^{oja} = \frac{\alpha}{\rho_k(\beta + t)}$, we can choose $c_k = \frac{\alpha}{\rho_k}$.

D.6 Proofs for Theorem 3.6 in Section 3

Lemma D.19 (Factor Estimation Error in L_q-norm)**.** *Suppose Assumption 2.1 and Lemma D.11 holds. Set $\eta_t^{\mathrm{oja}} = \frac{\alpha}{(\beta+t)\rho_k}$. Define the factor estimator for a new observation x by $\hat{f} := d^{-1/2} Q_t^\top x$, where Q_t denotes the output of the t-th iteration of Oja's algorithm. Let $R \in \mathbb{O}_k$ be an orthogonal matrix satisfying the conclusion of Proposition 2.2. Then for $q \ge 2, s \ge 2$, we have*

$$\begin{aligned} \left\| \left| \hat{f} - Rf \right|_s \right\|_q \le \sqrt{2}\, k^{1/2+1/q} \left\| |f|_2 \right\|_q \Bigg[& \left(\frac{\beta+1}{\beta+t} \right)^{\alpha/2} + C \frac{\alpha q^{1/2}}{\bar{\rho}_k} \frac{t^{1/2}}{\beta+t} \\ & + C^{2/q} \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{ -\frac{\beta \bar{\rho}_k^2}{\alpha^2 C^2 q} \right\} \Bigg] \\ & + \left(C \frac{K_u^2}{\lambda_k(\Sigma_f)} \left\| |f|_2 \right\|_q + K_u \big(c k^{1/2} + C q^{1/2} \big) \right) d^{-1/2}, \end{aligned} \tag{60}$$

where $C > 0$ is absolute constant. Furthermore, by taking $\beta \ge \Theta\Big(\bar{\rho}_k^{-2} \big(q \log d + \log \frac{k}{\bar{\rho}_k^2} \big) \Big)$, and absorbing the absolute constant, we have

$$\begin{aligned} \left\| \left| \hat{f} - Rf \right|_s \right\|_q \le C k^{1/2+1/q} \left\| |f|_2 \right\|_q & \left[\left(\frac{\beta+1}{\beta+t} \right)^{\alpha/2} + \frac{q^{1/2}}{\bar{\rho}_k} \frac{t^{1/2}}{\beta+t} \right] \\ & + C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} \left\| |f|_2 \right\|_q + K_u (k^{1/2} + q^{1/2}) + k^{1/2} \left\| |f|_2 \right\|_q \right) d^{-1/2}. \end{aligned} \tag{61}$$

Proof. By Lemma D.11, we obtain

$$\left\| \mathrm{dist}(Q_t, V) \right\|_q \le k^{1/q} \left[\left(\frac{\beta+1}{\beta+t} \right)^{\alpha/2} + C \cdot \frac{\alpha q^{1/2}}{\bar{\rho}_k} \cdot \frac{t^{1/2}}{\beta+t} + C^{2/q} \cdot \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{ -\frac{\beta \bar{\rho}_k^2}{\alpha^2 C^2 q} \right\} \right]. \tag{62}$$

Plug (62) into Proposition 2.2 and absorb absolute constant, we obtain

$$\begin{aligned} \left\| \left| \hat{f} - Rf \right|_s \right\|_q \le \sqrt{2}\, k^{1/2+1/q} \left\| |f|_2 \right\|_q \Bigg[& \left(\frac{\beta+1}{\beta+t} \right)^{\alpha/2} + C \frac{\alpha q^{1/2}}{\bar{\rho}_k} \frac{t^{1/2}}{\beta+t} \\ & + C^{2/q} \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{ -\frac{\beta \bar{\rho}_k^2}{\alpha^2 C^2 q} \right\} \Bigg] \\ & + \left(C \frac{K_u^2}{\lambda_k(\Sigma_f)} \left\| |f|_2 \right\|_q + K_u \big(c k^{1/2} + C q^{1/2} \big) \right) d^{-1/2}, \end{aligned} \tag{63}$$

By taking

$$\beta \ge \frac{\alpha^2 C^2 q}{\bar{\rho}_k^2} \left[\frac{1}{2} \log d + \frac{1}{q} \log \left(\frac{k C^2 \alpha^2}{\bar{\rho}_k^2} \right) \right] = \Theta \left(\bar{\rho}_k^{-2} \left(q \log d + \log \frac{k}{\bar{\rho}_k^2} \right) \right),$$

where α and C are absolute constants, we have

$$k^{1/q} C^{2/q} \alpha^{2/q} \bar{\rho}_k^{-2/q} \exp\left\{-\frac{\beta \bar{\rho}_k^2}{\alpha^2 C^2 q}\right\} \le d^{-1/2}.$$

Therefore, the exponential residual term in Lemma D.11 can be absorbed into the $d^{-1/2}$ factor-estimation error term

$$\begin{aligned}\left\| |\hat{f} - Rf|_s \right\|_q \le \sqrt{2}\, k^{1/2+1/q} \left\| |f|_2 \right\|_q \left[\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} + C \frac{\alpha q^{1/2}}{\bar{\rho}_k} \frac{t^{1/2}}{\beta+t} \right] \\ + C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} \left\| |f|_2 \right\|_q + K_u (k^{1/2} + q^{1/2}) + k^{1/2} \left\| |f|_2 \right\|_q \right) d^{-1/2}.\end{aligned}$$

□

To simplify the term $(\frac{\beta+1}{\beta+t})^{\alpha/2}$, we provide the following lemma.

Lemma D.20. *Let $\beta \ge 0$ and $\alpha \ge 2$. Then, for all $t \ge 1$,*

$$\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} \le C_{\alpha,\beta} t^{-1/2},$$

where

$$C_{\alpha,\beta} = \begin{cases} 1, & 0 \le \beta \le \alpha - 1, \\ \dfrac{(\beta+1)^{\alpha/2} \beta^{(1-\alpha)/2} (\alpha-1)^{(\alpha-1)/2}}{\alpha^{\alpha/2}}, & \beta > \alpha - 1. \end{cases} \tag{64}$$

Moreover, this is the smallest possible constant for the inequality to hold uniformly over $t \ge 1$. In particular, when $\alpha = 8$,

$$C_{8,\beta} = \begin{cases} 1, & 0 \le \beta \le 7, \\ \dfrac{7^{7/2}}{8^4} \dfrac{(\beta+1)^4}{\beta^{7/2}}, & \beta > 7. \end{cases}$$

Proof. It suffices to find the smallest constant $C_{\alpha,\beta}$ such that

$$\left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2} \le C_{\alpha,\beta} t^{-1/2}$$

for all $t \ge 1$. Equivalently,

$$C_{\alpha,\beta} = \sup_{t \ge 1} t^{1/2} \left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2}.$$

Define

$$f(t) = t^{1/2} \left(\frac{\beta+1}{\beta+t}\right)^{\alpha/2}, \qquad t \ge 1.$$

Then

$$\log f(t) = \frac{1}{2} \log t + \frac{\alpha}{2} \log(\beta+1) - \frac{\alpha}{2} \log(\beta+t).$$

Therefore,

$$\frac{d}{dt} \log f(t) = \frac{1}{2t} - \frac{\alpha}{2(\beta+t)} = \frac{\beta - (\alpha-1)t}{2t(\beta+t)}.$$

If $0 \le \beta \le \alpha - 1$, then for all $t \ge 1$,

$$\beta - (\alpha-1)t \le \beta - (\alpha-1) \le 0.$$

Hence $f(t)$ is non-increasing on $[1, \infty)$. Thus

$$\sup_{t \ge 1} f(t) = f(1) = 1.$$

Therefore,

$$C_{\alpha,\beta} = 1.$$

If $\beta > \alpha - 1$, then the derivative changes sign at

$$t_* = \frac{\beta}{\alpha - 1} > 1.$$

Thus $f(t)$ is increasing on $[1, t_*]$ and decreasing on $[t_*, \infty)$. Hence

$$\sup_{t \geq 1} f(t) = f(t_*).$$

A direct calculation gives

$$f(t_*) = \left(\frac{\beta}{\alpha - 1}\right)^{1/2} \left(\frac{\beta + 1}{\beta + \frac{\beta}{\alpha - 1}}\right)^{\alpha/2} = \frac{(\beta+1)^{\alpha/2} \beta^{(1-\alpha)/2} (\alpha - 1)^{(\alpha-1)/2}}{\alpha^{\alpha/2}}.$$

Therefore,

$$C_{\alpha,\beta} = \frac{(\beta+1)^{\alpha/2} \beta^{(1-\alpha)/2} (\alpha - 1)^{(\alpha-1)/2}}{\alpha^{\alpha/2}}.$$

Combining the two cases proves the desired inequality. □

Lemma D.21 (Moment decay of $\nabla_\theta l$)**.** *Let $q \geq 2$ and $s \geq 2$. For i.i.d. samples (f_i, y_i), $i = 1, \ldots, m$, suppose*

$$\left\| \, |\nabla_\theta l(\theta; f_i, y_i)|_q \, \right\|_q < \infty.$$

Then there exists a constant $C_{s,q} > 0$, independent of m, such that

$$\left\| \left| \frac{1}{m} \sum_{i=1}^{m} \Big(\nabla_\theta l(\theta; f_i, y_i) - \mathbb{E}_z[\nabla_\theta l(\theta; f_i, y_i)] \Big) \right|_s \right\|_q \leq C_{s,q} m^{-1/2}.$$

Moreover, one may take

$$C_{s,q} = C_q \left(\sum_{j=1}^{p} \Big[(\mathbb{E} X_{1,j}^2)^{1/2} + (\mathbb{E}|X_{1,j}|^q)^{1/q} \Big]^r \right)^{1/r}, \qquad r := q \wedge s,$$

where

$$X_1 := \nabla_\theta l(\theta; f_1, y_1) - \mathbb{E}_z[\nabla_\theta l(\theta; f_1, y_1)].$$

Proof. For simplicity, define the $\mathbb{R}^p$-valued random vectors

$$X_i := \nabla_\theta l(\theta; f_i, y_i) - \mathbb{E}_z[\nabla_\theta l(\theta; f_i, y_i)], \qquad S_m := \frac{1}{m} \sum_{i=1}^{m} X_i.$$

Let $r := q \wedge s$. Since $r \leq s$, we have $|v|_s \leq |v|_r$ for all $v \in \mathbb{R}^p$. Therefore,

$$\big\| |S_m|_s \big\|_q \leq \big\| |S_m|_r \big\|_q.$$

Moreover, since $r \leq q$, we have $q/r \geq 1$. Hence, by Minkowski's inequality applied with exponent q/r,

$$\begin{aligned} \big\| |S_m|_r \big\|_q &= \left\| \left(\sum_{j=1}^{p} |S_{m,j}|^r \right)^{1/r} \right\|_q = \left\| \sum_{j=1}^{p} |S_{m,j}|^r \right\|_{q/r}^{1/r} \\ &\leq \left(\sum_{j=1}^{p} \big\| |S_{m,j}|^r \big\|_{q/r} \right)^{1/r} = \left(\sum_{j=1}^{p} \|S_{m,j}\|_q^r \right)^{1/r}. \end{aligned}$$

For each $j \in \{1, \dots, p\}$, the random variables $\{X_{i,j}\}_{i=1}^m$ are i.i.d. and mean-zero. By Rosenthal's inequality (Lemma D.3), there exists a constant $C_q > 0$, depending only on q, such that

$$\left\| \sum_{i=1}^m X_{i,j} \right\|_q^q \le C_q \left[\sum_{i=1}^m \mathbb{E}|X_{i,j}|^q + \left(\sum_{i=1}^m \mathbb{E}X_{i,j}^2 \right)^{q/2} \right] = C_q \left[m\mathbb{E}|X_{1,j}|^q + \left(m\mathbb{E}X_{1,j}^2 \right)^{q/2} \right].$$

Taking the q-th root and dividing by m, and absorbing constants into C_q, we obtain

$$\|S_{m,j}\|_q \le C_q \left[m^{-1+1/q} (\mathbb{E}|X_{1,j}|^q)^{1/q} + m^{-1/2} (\mathbb{E}X_{1,j}^2)^{1/2} \right].$$

Since $q \ge 2$, $m^{-1+1/q} \le m^{-1/2}$. Hence

$$\|S_{m,j}\|_q \le C_q m^{-1/2} \left[(\mathbb{E}X_{1,j}^2)^{1/2} + (\mathbb{E}|X_{1,j}|^q)^{1/q} \right].$$

Substituting this coordinate-wise bound into the previous display gives

$$\big\| |S_m|_s \big\|_q \le \left(\sum_{j=1}^p \|S_{m,j}\|_q^r \right)^{1/r} \le C_q m^{-1/2} \left(\sum_{j=1}^p \left[(\mathbb{E}X_{1,j}^2)^{1/2} + (\mathbb{E}|X_{1,j}|^q)^{1/q} \right]^r \right)^{1/r}.$$

Define

$$C_{s,q} := C_q \left(\sum_{j=1}^p \left[(\mathbb{E}X_{1,j}^2)^{1/2} + (\mathbb{E}|X_{1,j}|^q)^{1/q} \right]^r \right)^{1/r}, \qquad r = q \wedge s.$$

The assumed q-th moment condition implies $C_{s,q} < \infty$. Therefore,

$$\big\| |S_m|_s \big\|_q \le C_{s,q} m^{-1/2},$$

which completes the proof. □

We are now in position to prove Theorem 3.6.

Proof of Theorem 3.6. Let $\mathcal{F}_{t-1}$ denote the σ-field generated by all random variables revealed up to iterations $0, 1, 2, \dots, t-1$. We denote by $\mathbb{E}_z[\cdot]$ the expectation taken only with respect to the random vectors $(f_{t,i}, y_{t,i})$, which are independent of $\mathcal{F}_{t-1}$. Note that in the proof we introduce constants with slightly different notation from that used in the statement of Theorem 3.6, for clarity.

Step 1. Recall

$$\theta_t = \theta_{t-1} - \eta_{t-1} \nabla_\theta \frac{1}{m} \sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}),$$

where $\hat{f}_{t,i} = d^{-\frac{1}{2}} \hat{Q}_{t-1}^\top x_{t,i}$. It holds that

$$\begin{aligned}
\left\| |\theta_t - \theta_{t-1}^*|_s \right\|_q^2 &= \left\| |\theta_{t-1} - \eta_{t-1} \nabla_\theta \frac{1}{m} \sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) - \theta_{t-1}^*|_s \right\|_q^2 \\
&= \Big\| |\theta_{t-1} - \theta_{t-1}^* - \eta_{t-1} \nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] \\
&\qquad + \eta_{t-1} \big(\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m} \sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) \big) |_s \Big\|_q^2
\end{aligned}$$

Applying Lemma D.1, it follows that

$$\left\| |\theta_t - \theta_{t-1}^*|_s \right\|_q^2 \le \left\| |\theta_{t-1} - \theta_{t-1}^* - \eta_{t-1} \nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})]|_s \right\|_q^2 \tag{65}$$

$$+ \eta_{t-1}^2 \cdot (\max\{q, s\} - 1) \cdot \left\| |\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m} \sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s \right\|_q^2 \tag{66}$$

$$+ 2 \left\| |w|_s \right\|_q^{2-q} \mathbb{E}\big(|w|_s^{q-s} \langle w^{\odot(s-1)}, v \rangle \big), \tag{67}$$

where $w := \theta_{t-1} - \theta^*_{t-1} - \eta_{t-1}\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]$, and $v := \eta_{t-1}\big(\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})\big)$.

Step 2. Now we control the first term on the RHS of Equation (65). Note that $\nabla_\theta \mathbb{E}_z[l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] = 0$, by applying Lemma D.1 again, we have

$$
\begin{aligned}
&\Big\||\theta_{t-1} - \theta^*_{t-1} - \eta_{t-1}\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]|_s\Big\|_q^2 \\
&= \Big\||\theta_{t-1} - \theta^*_{t-1} - \eta_{t-1}\Big(\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta \mathbb{E}_z[l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]\Big)|_s\Big\|_q^2 \\
&\le \Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2 + \eta_{t-1}^2(\max\{q,s\}-1)\Big\||\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta \mathbb{E}_z[l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]|_s\Big\|_q^2 \\
&\quad - 2\eta_{t-1}\Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^{2-q}\mathbb{E}\Big[|\theta_{t-1} - \theta^*_{t-1}|_s^{q-s}\cdot \\
&\qquad \langle(\theta_{t-1} - \theta^*_{t-1})^{\odot(s-1)}, \mathbb{E}_z[\nabla_\theta l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i}) - \nabla_\theta l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]\rangle\Big]
\end{aligned}
\tag{68}
$$

Due to Assumption 3.2, it yields

$$
\Big\||\mathbb{E}_z\big[\nabla_\theta l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})\big]|_s\Big\|_q^2 \le \Big\|L_{s,\theta}|\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2 \tag{69}
$$

Along with Assumption 3.1, we can bound

$$
\langle(\theta_{t-1} - \theta^*_{t-1})^{\odot(s-1)}, \mathbb{E}_z[\nabla_\theta l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i}) - \nabla_\theta l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]\rangle \ge \mu|\theta_{t-1} - \theta^*_{t-1}|_s^s \tag{70}
$$

Inserting inequalities (69), (70) into Equation (68), we obtain

$$
\begin{aligned}
&\Big\||\theta_{t-1} - \theta^*_{t-1} - \eta_{t-1}\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]|_s\Big\|_q^2 \\
&\le \Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2 + \eta_{t-1}^2(\max\{q,s\}-1)L_{s,\theta}^2\Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2 - 2\eta_{t-1}\mu\Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2 \\
&= \big(1 - 2\mu\eta_{t-1} + \eta_{t-1}^2(\max\{q,s\}-1)L_{s,\theta}^2\big)\Big\||\theta_{t-1} - \theta^*_{t-1}|_s\Big\|_q^2
\end{aligned}
\tag{71}
$$

Step 3. Now we control the term on the RHS of Equation (66).

$$
\begin{aligned}
&\Big\||\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s\Big\|_q \\
&\le \underbrace{\Big\||\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1}f_{t,i}, y_{t,i})] - \nabla_\theta \mathbb{E}_z[l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]|_s\Big\|_q}_{(i)} \\
&\quad + \underbrace{\Big\||\nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i}) - \nabla_\theta \mathbb{E}_z[l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})]|_s\Big\|_q}_{(ii)} \\
&\quad + \underbrace{\Big\||\nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta^*_{t-1}; \hat{f}_{t,i}, y_{t,i}) - \nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta^*_{t-1}; R_{t-1}f_{t,i}, y_{t,i})|_s\Big\|_q}_{(iii)} \\
&\quad + \underbrace{\Big\||\nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta^*_{t-1}; \hat{f}_{t,i}, y_{t,i}) - \nabla_\theta \frac{1}{m}\sum_{i=1}^m l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s\Big\|_q}_{(iv)}.
\end{aligned}
\tag{72}
$$

Rely on Assumption 3.2, we control $(i), (iv)$ as

$$(i) + (iv) \leq 2L_{s,\theta} \Big\| |\theta_{t-1} - \theta^*_{t-1}|_s \Big\|_q \tag{73}$$

The upper bound of (ii) can be derived using Lemma D.21:

$$(ii) = \Big\| |\frac{1}{m} \sum_{i=1}^{m} \big(\nabla_\theta l(\theta^*_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) - \mathbb{E}_z[\nabla_\theta l(\theta^*_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] \big) |_s \Big\|_q \leq C_{s,q} m^{-\frac{1}{2}} \tag{74}$$

(iii) depends on the estimation of $\hat{f}_i^t$. Along with Assumption 3.3, we obtain

$$(iii) = \Big\| \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} l(\theta^*_{t-1}; \hat{f}_{t,i}, y_{t,i}) - \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} l(\theta^*_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) |_s \Big\|_q \leq L_{s,f} \Big\| |\hat{f}_{t,i} - R_{t-1} f_{t,i}|_s \Big\|_q \tag{75}$$

Combining Equations (73), (74), (75) and (72), it holds that

$$\begin{aligned} &\Big\| |\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) |_s \Big\|_q^2 \\ &\leq 4\big(2L_{s,\theta}^2 \Big\| |\theta_{t-1} - \theta^*_{t-1}|_s \Big\|_q^2 + C_{s,q}^2 m^{-1} + L_{s,f}^2 \Big\| |\hat{f}_{t,i} - R_{t-1} f_{t,i}|_s \Big\|_q^2 \big) \end{aligned}$$

An upper bound of (66) follows immediately,

$$\begin{aligned} &\eta_{t-1}^2 \cdot (\max\{q, s\} - 1) \cdot \Big\| |\nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] - \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) |_s \Big\|_q^2 \\ &\leq 4\eta_{t-1}^2 \cdot (\max\{q, s\} - 1) \cdot \big(2L_{s,\theta}^2 \Big\| |\theta_{t-1} - \theta^*_{t-1}|_s \Big\|_q^2 + C_{s,q}^2 m^{-1} + L_{s,f}^2 \Big\| |\hat{f}_{t,i} - R_{t-1} f_{t,i}|_s \Big\|_q^2 \big) \end{aligned}$$

Step 4: Now we control the term on the RHS of (67). Note that $w = \theta_{t-1} - \theta^*_{t-1} - \eta_{t-1} \nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})] \in \mathcal{F}_{t-1}$, we can condition on $\mathcal{F}_{t-1}$ and rewrite the equation as

$$\Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\big[|w|_s^{q-s} \langle w^{\odot(s-1)}, v \rangle \big] = \Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\Big[|w|_s^{q-s} \langle w^{\odot(s-1)}, \mathbb{E}(v|\mathcal{F}_{t-1}) \rangle \Big] \tag{76}$$

By Hölder inequality and Jensen Inequality,

$$|\langle w^{\odot(s-1)}, \mathbb{E}(v|\mathcal{F}_{t-1}) \rangle| \leq |w^{\odot(s-1)}|_{s^*} \cdot |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \leq |w|_s^{s-1} \cdot |\mathbb{E}(v|\mathcal{F}_{t-1})|_s,$$

where $s^* = \frac{s}{s-1}$ satisfying $\frac{1}{s^*} + \frac{1}{s} = 1$. Consequently,

$$\Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\big[|w|_s^{q-s} \langle w^{\odot(s-1)}, v \rangle \big] \leq \Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\Big[|w|_s^{q-1} \cdot |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \Big] \tag{77}$$

Again, we apply Hölder inequality, $\mathbb{E}|ab| \leq \mathbb{E}^{1/q^*} |a|^{q^*} \cdot \mathbb{E}^{1/q} |b|^q$, with $q^* = \frac{q}{q-1}$ to get

$$\mathbb{E}\Big[|w|_s^{q-1} \cdot |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \Big] \leq \mathbb{E}^{\frac{q-1}{q}} \Big[|w|_s^q \Big] \cdot \mathbb{E}^{\frac{1}{q}} \Big[|\mathbb{E}(v|\mathcal{F}_{t-1})|_s^q \Big] = \Big\| |w|_s \Big\|_q^{q-1} \Big\| |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \Big\|_q \tag{78}$$

Combining (77) and (78), we actually show that if $w \in \mathcal{F}_{t-1}$, then

$$\Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\big[|w|_s^{q-s} \langle w^{\odot(s-1)}, v \rangle \big] \leq \Big\| |w|_s \Big\|_q \cdot \Big\| |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \Big\|_q \tag{79}$$

We focus on $\mathbb{E}[v \,|\, \mathcal{F}_{t-1}]$ first. Substitute the definition of v, we have

$$\begin{aligned} \mathbb{E}[v \,|\, \mathcal{F}_{t-1}] &= \eta_{t-1} \cdot \mathbb{E}\Big[\mathbb{E}_z\big[\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) \big] \Big| \mathcal{F}_{t-1} \Big] - \eta_{t-1} \cdot \mathbb{E}\Big[\nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) | \mathcal{F}_{t-1} \Big] \\ &= \eta_{t-1} \cdot \Big(\mathbb{E}_z\big[\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) \big] - \mathbb{E}_z[\nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})] \Big) \\ &= \eta_{t-1} \cdot \mathbb{E}_z\Big[\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) - \nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i}) \Big], \end{aligned} \tag{80}$$

where the second equality is due to $\theta_{t-1}, R_{t-1} \in \mathcal{F}_{t-1}$. Taking l_s-norm and L_q-norm of random variable on both side of (80), we have

$$\begin{aligned}
\Big\| |\mathbb{E}(v|\mathcal{F}_{t-1})|_s \Big\|_q &\le \eta_{t-1} \cdot \mathbb{E}^{\frac{1}{q}} \Big[\mathbb{E}_z^q \big(|\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) - \nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s \big) \Big] \\
&\le \eta_{t-1} \cdot \mathbb{E}^{\frac{1}{q}} \Big[\mathbb{E}_z \big(|\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) - \nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s^q \big) \Big] \\
&= \eta_{t-1} \cdot \Big\| |\nabla l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i}) - \nabla l(\theta_{t-1}; \hat{f}_{t,i}, y_{t,i})|_s \Big\|_q \\
&\le \eta_{t-1} L_{s,f} \Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q,
\end{aligned}$$

where the second inequality holds by applying Jensen's inequality since $q \ge 2$.

Replacing this upper bound into inequality (78), we have

$$\Big\| |w|_s \Big\|_q^{2-q} \mathbb{E}\big[|w|_s^{q-s} \langle w^{\odot(s-1)}, v \rangle \big] \le \eta_{t-1} \Big\| |w|_s \Big\|_q L_{s,f} \Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q$$

Note that $w = \theta_{t-1} - \theta_{t-1}^* - \eta_{t-1} \nabla_\theta \mathbb{E}_z[l(\theta_{t-1}; R_{t-1} f_{t,i}, y_{t,i})]$, Equation (71) already provides a bound for $\Big\| |w|_s \Big\|_q$, i.e.

$$\Big\| |w|_s \Big\|_q \le \big(1 - 2\mu\eta_{t-1} + \eta_{t-1}^2 (\max\{q, s\} - 1) L_{s,\theta}^2 \big)^{\frac{1}{2}} \Big\| |\theta_{t-1} - \theta_{t-1}^*|_s \Big\|_q$$

The term on the RHS of (67) can then be controlled by

$$2\big(1 - 2\mu\eta_{t-1} + \eta_{t-1}^2 (\max\{q, s\} - 1) L_{s,\theta}^2 \big)^{\frac{1}{2}} \eta_{t-1} L_{s,f} \Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q \Big\| |\theta_{t-1} - \theta_{t-1}^*|_s \Big\|_q \quad (81)$$

We denote $\alpha_{t-1} := \big(1 - 2\mu\eta_{t-1} + \eta_{t-1}^2 (\max\{q, s\} - 1) L_{s,\theta}^2 \big)^{\frac{1}{2}}$ and $C_1^2 := \max\{q, s\} - 1$ also for simplicity of notation. Lemma D.19 provides the factor estimation error for the output of the t-th iteration of Oja's algorithm, i.e.,

$$\begin{aligned}
\Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q \le\ & C k^{1/2+1/q} \Big\| \, |f|_2 \, \Big\|_q \left[\left(\frac{\beta + 1}{\beta + t} \right)^{\alpha/2} + \frac{q^{1/2}}{\bar{\rho}_k} \frac{t^{1/2}}{\beta + t} \right] \\
& + C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} \Big\| \, |f|_2 \, \Big\|_q + K_u (k^{1/2} + q^{1/2}) + k^{1/2} \Big\| \, |f|_2 \, \Big\|_q \right) d^{-1/2}.
\end{aligned} \quad (82)$$

Note that $\frac{t^{1/2}}{\beta + t} \le t^{-1/2}$ and applying Lemma D.20, it simplifies to

$$\begin{aligned}
\Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q \le\ & C k^{1/2} \Big\| \, |f|_2 \, \Big\|_q (C_{\alpha,\beta} + \frac{q^{1/2}}{\bar{\rho}_k}) t^{-\frac{1}{2}} \\
& + C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} \Big\| \, |f|_2 \, \Big\|_q + K_u (k^{1/2} + q^{1/2}) + k^{1/2} \Big\| \, |f|_2 \, \Big\|_q \right) d^{-1/2}.
\end{aligned} \quad (83)$$

Equivalently,

$$\Big\| |R_{t-1} f_{t,i} - \hat{f}_{t,i}|_s \Big\|_q \le C_2 M_f \, t^{-\frac{1}{2}} + C_d, \quad (84)$$

where

$$\begin{aligned}
C_2 &:= C k^{1/2} (C_{\alpha,\beta} + \frac{q^{1/2}}{\bar{\rho}_k}), \quad M_f := \Big\| \, |f|_2 \, \Big\|_q, \\
C_d &:= C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} M_f + K_u (k^{1/2} + q^{1/2}) + k^{1/2} M_f \right) d^{-1/2}.
\end{aligned}$$

and C is absolute constants.

To conclude the results in Steps 2,3,4, we obtain that $\left\| |\theta_t - \theta^*_{t-1}|_s \right\|_q^2$ admits the following bound:

$$\begin{aligned}\left\| |\theta_t - \theta^*_{t-1}|_s \right\|_q^2 \le \alpha_{t-1}^2 \left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q^2 + 4\eta_{t-1}^2 C_1^2 \cdot \Big(2L_{s,\theta}^2 \left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q^2 \\ + C_{s,q}^2 m^{-1} + L_{s,f}^2 (C_2 M_f\, t^{-\frac{1}{2}} + C_d)^2 \Big) \\ + 2\alpha_{t-1}\eta_{t-1} L_{s,f}\, (C_2 M_f\, t^{-\frac{1}{2}} + C_d) \left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q,\end{aligned} \tag{85}$$

Step 5: Now we have derived that $\left\| |\theta_t - \theta^*_{t-1}|_s \right\|_q$ can be bounded by a function of $\left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q$. In this step, we consider the minimizer shift $\left\| |\theta^*_{t-1} - \theta^*_{t-2}|_s \right\|_q$.

Apply Proposition D.17, we obtain that

$$\left\| \, |\theta^*_{t-1} - \theta^*_{t-2}|_s \, \right\|_q \;\le\; C \frac{L_{s,f}}{\mu} M_f (M_A k^{1/2} c_k + d^{-1/2}) t^{-1}, \tag{86}$$

where C is an absolute constant, $M_f := \left\| |f|_2 \right\|_q$, and $c_k = \frac{\alpha}{\rho_k}$ since $\eta_t^{oja} = \frac{\alpha}{\rho_k(\beta + t)}$.

Apply Lemma D.1 to $\left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q = \left\| |(\theta_{t-1} - \theta^*_{t-2}) + (\theta^*_{t-2} - \theta^*_{t-1})|_s \right\|_q$ with $w = \theta_{t-1} - \theta^*_{t-2}$ and $v = \theta^*_{t-2} - \theta^*_{t-1}$, it holds that

$$\begin{aligned}\left\| |\theta_{t-1} - \theta^*_{t-1}|_s \right\|_q^2 &\le \left\| |w|_s \right\|_q^2 + 2 \left\| |w|_s \right\|_q \left\| |v|_s \right\|_q + (\max\{q, s\} - 1) \cdot \left\| |v|_s \right\|_q^2 \\ &\le \left\| |w|_s \right\|_q^2 + 2\, C_4 \left\| |w|_s \right\|_q t^{-1} + C_1^2 C_4^2 \cdot t^{-2},\end{aligned} \tag{87}$$

where $C_4 := C \frac{L_{s,f}}{\mu} M_f (M_A k^{1/2} \rho_k^{-1} + d^{-1/2})$.

Step 5.1: Derive the Recursion

To abbreviate our notation, we denote

$$F_{t-1} := C_2 M_f t^{-\frac{1}{2}} + C_3 d^{-1/2}, \quad \Xi_{t-1}^2 := C_{s,q}^2 m^{-1} + L_{s,f}^2 (C_2 M_f\, t^{-\frac{1}{2}} + C_d)^2,$$

where $C_3 := C \left(\frac{K_u^2}{\lambda_k(\Sigma_f)} M_f + K_u (k^{1/2} + q^{1/2}) + k^{1/2} M_f \right)$. Rewrite shift of minimizer satisfies (as in (86)) as

$$\left\| \, |\theta^*_{t-2} - \theta^*_{t-1}|_s \, \right\|_q \;\le\; \kappa_{t-1}, \qquad \kappa_{t-1} := C_4 t^{-1}, \tag{88}$$

and that Lemma D.1 implies the inequality (87)

$$U_{t-1}^2 := \left\| \, |\theta_{t-1} - \theta^*_{t-1}|_s \, \right\|_q^2 \;\le\; X_{t-1}^2 + 2\kappa_{t-1} X_{t-1} + C_1^2 \kappa_{t-1}^2, \qquad t \ge 2, \tag{89}$$

where $X_{t-1} := \| \, |\theta_{t-1} - \theta^*_{t-2}|_s \|_{L_q}^2$. Note that $C_1 \ge 1$ and

$$X_{t-1}^2 + 2\kappa_{t-1} X_{t-1} + C_1^2 \kappa_{t-1}^2 \le (X_{t-1} + C_1 \kappa_{t-1})^2,$$

we have

$$U_{t-1} \le X_{t-1} + C_1 \kappa_{t-1}, \tag{90}$$

Then the bound (85) implies, for all $t \ge 2$,

$$X_t^2 \le \alpha_{t-1}^2\, U_{t-1}^2 + 8\eta_{t-1}^2 C_1^2 L_{s,\theta}^2\, U_{t-1}^2 + 4\eta_{t-1}^2 C_1^2\, \Xi_{t-1}^2 + 2\alpha_{t-1}\eta_{t-1} L_{s,f} F_{t-1}\, U_{t-1}, \tag{91}$$

Substituting (89) and (90) into (91) yields

$$\begin{aligned}X_t^2 \le \underbrace{\left(\alpha_{t-1}^2 + 8\eta_{t-1}^2 C_1^2 L_{s,\theta}^2\right)}_{=:\beta_{t-1}^2} \left(X_{t-1}^2 + 2\kappa_{t-1} X_{t-1} + C_1^2 \kappa_{t-1}^2 \right) + 4\eta_{t-1}^2 C_1^2 \Xi_{t-1}^2 \\ + 2\alpha_{t-1}\eta_{t-1} L_{s,f} F_{t-1}\, (X_{t-1} + C_1 \kappa_{t-1}),\end{aligned} \tag{92}$$

where $\beta_{t-1}^2 := 1 - 2\mu\eta_{t-1} + 9\eta_{t-1}^2(\max\{q,s\}-1)L_{s,\theta}^2 \le 1$, by assumption.

Now, we control the cross terms.

(i) The term $2 \cdot \kappa_{t-1}\beta_{t-1}^2 \cdot X_{t-1}$. For any $\varepsilon > 0$, Young's inequality gives $2\kappa B \le \varepsilon B^2 + \varepsilon^{-1}\kappa^2$. Hence

$$2 \cdot \kappa_{t-1}\beta_{t-1}^2 \cdot X_{t-1} \le \varepsilon_{t-1}X_{t-1}^2 + \varepsilon_{t-1}^{-1}\kappa_{t-1}^2. \tag{93}$$

(ii) The term $2 \cdot \alpha_{t-1}L_{s,f}F_{t-1}\eta_{t-1} \cdot X_{t-1}$. Again, for any $\varepsilon > 0$,

$$2 \cdot \alpha_{t-1}L_{s,f}F_{t-1}\eta_{t-1} \cdot X_{t-1} \le \varepsilon_{t-1}X_{t-1}^2 + \varepsilon_{t-1}^{-1}L_{s,f}^2F_{t-1}^2\eta_{t-1}^2 \tag{94}$$

(iii) The term $2 \cdot \alpha_{t-1}C_1L_{s,f}\kappa_{t-1}\eta_{t-1} \cdot F_{t-1}$. Again, for any $\varepsilon > 0$,

$$2 \cdot \alpha_{t-1}C_1L_{s,f}\kappa_{t-1}\eta_{t-1} \cdot F_{t-1} \le \varepsilon_{t-1}F_{t-1}^2 + \varepsilon_{t-1}^{-1}C_1^2L_{s,f}^2\kappa_{t-1}^2\eta_{t-1}^2 \tag{95}$$

We set $\varepsilon_{t-1} = \frac{\mu}{8}\eta_{t-1}$ in (93), (94). With this choice, the coefficient of X_t^2 satisfies

$$\beta_{t-1}^2 + 2\varepsilon_{t-1} = 1 - 2\mu\eta_{t-1} + 9\eta_{t-1}^2(\max\{q,s\}-1)L_{s,\theta}^2 + \frac{\mu}{4}\eta_{t-1} \le 1 - \frac{\mu}{2}\eta_{t-1} \tag{96}$$

Also, we set $\varepsilon_{t-1} = \eta_{t-1}$ in (95), so that

$$2 \cdot \alpha_{t-1}C_1L_{s,f}\kappa_{t-1}\eta_{t-1} \cdot F_{t-1} \le \eta_{t-1}F_{t-1}^2 + C_1^2L_{s,f}^2\kappa_{t-1}^2\eta_{t-1}, \tag{97}$$

which the factor error F_{t-1} can be absorbed by previous terms.

Now combine (92) with (93), (94), and using $\alpha_{t-1} \le \beta_{t-1} \le 1$, we obtain

$$\begin{aligned} X_t^2 \le & \Big(1 - \frac{\mu}{2}\eta_{t-1}\Big)X_{t-1}^2 + \frac{8}{\mu}\eta_{t-1}^{-1}\kappa_{t-1}^2 + C_1^2\kappa_{t-1}^2 \\ & + 4C_1^2\eta_{t-1}^2\Xi_{t-1}^2 + (\frac{8}{\mu}L_{s,f}^2 + 1)\eta_{t-1}F_{t-1}^2 + C_1^2L_{s,f}^2\kappa_{t-1}^2\eta_{t-1}. \end{aligned} \tag{98}$$

Step 5.2: Solving the recursion. We consider the recursion of the form

$$Z_t \le \Big(1 - \frac{\mu}{2}\eta_{t-1}\Big)Z_{t-1} + r_{t-1}, \tag{99}$$

where $Z_t = X_t^2$ and r_{t-1} is the corresponding remainder term. Let $S_t := \sum_{j=0}^{t-1}\eta_j$. Multiplying (99) by $\exp\big(\frac{\mu}{2}S_t\big)$ and using that $\exp(x)\big(1-x\big) \le 1$ for $x = \frac{\mu}{2}\eta_{t-1} > 0$, we obtain

$$\exp\Big(\frac{\mu}{2}S_t\Big)Z_t \le \exp\Big(\frac{\mu}{2}S_{t-1}\Big)Z_{t-1} + \exp\Big(\frac{\mu}{2}S_t\Big)r_{t-1}.$$

Iterating from 0 to t yields

$$\exp\Big(\frac{\mu}{2}S_t\Big)Z_t \le Z_0 + \sum_{k=0}^{t-1}\exp\Big(\frac{\mu}{2}S_{k+1}\Big)\,r_k$$

Reorganizing terms yields

$$Z_t \le \exp\Big(-\frac{\mu}{2}S_t\Big)Z_0 + \sum_{k=0}^{t-1}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big)\,r_k. \tag{100}$$

Step 5.3: Bounding the remainder.

For (98), we apply (100) with $r_{t-1} = \frac{8}{\mu}\eta_{t-1}^{-1}\kappa_{t-1}^2 + C_1^2\kappa_{t-1}^2 + 4C_1^2\eta_{t-1}^2\Xi_{t-1}^2 + (\frac{8}{\mu}L_{s,f}^2+1)\eta_{t-1}F_{t-1}^2 + C_1^2L_{s,f}^2\kappa_{t-1}^2\eta_{t-1}$. Then, it holds that

$$\begin{aligned}
X_t^2 \le{}& \exp\Big(-\frac{\mu}{2}S_t\Big)X_0^2 + \frac{8}{\mu}C_4^2\sum_{k=0}^{t-1}\eta_k^{-1}\frac{1}{(k+1)^2}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ C_1^2C_4^2\sum_{k=0}^{t-1}\frac{1}{(k+1)^2}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ 4C_1^2(C_{s,q}^2m^{-1} + 2L_{s,f}^2C_d^2)\sum_{k=0}^{t-1}\eta_k^2\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ 8C_1^2L_{s,f}^2C_2^2M_f^2\sum_{k=0}^{t-1}\eta_k^2\frac{1}{k+1}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ \big(2(\frac{8}{\mu}L_{s,f}^2+1)C_2^2M_f^2\big)\sum_{k=0}^{t-1}\eta_k\frac{1}{k+1}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ 2(\frac{8}{\mu}L_{s,f}^2+1)C_d^2\sum_{k=0}^{t-1}\eta_k\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big) \\
&+ L_{s,f}^2C_1^2C_4^2\sum_{k=0}^{t-1}\eta_k\frac{1}{(k+1)^2}\exp\Big(-\frac{\mu}{2}(S_t - S_{k+1})\Big)
\end{aligned} \tag{101}$$

Combine with Lemma D.23 and take $\lambda = \frac{\mu}{2}$, we have

$$\Big\||\theta_t - \theta_{t-1}^*|_s\Big\|_q^2 \le \exp\Big(-\frac{\mu}{2}S_t\Big)\Big\||\theta_0 - \theta^*|_s\Big\|_q^2 + \frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)C_d^2 + o(1) \tag{102}$$

Furthermore, we recall the assumption $\eta_t = \mathcal{O}((t+1)^{-\gamma})$ for some $\gamma \in (0,1)$. Then $S_t = \Theta(t^{1-\gamma})$ and Lemma D.24 gives

$$\begin{aligned}
X_t^2 \le{}& \exp\Big(-\frac{\mu}{2}S_t\Big)X_0^2 + \frac{2}{\mu}\cdot\Big(\frac{8}{\mu}C_4^2\mathcal{O}(t^{-(2-2\gamma)}) + C_1^2C_4^2\mathcal{O}(t^{-(2-\gamma)}) \\
&+ 4C_1^2(C_{s,q}^2m^{-1} + 2L_{s,f}^2C_d^2)\mathcal{O}(t^{-\gamma}) + 8C_1^2L_{s,f}^2C_2^2M_f^2\mathcal{O}(t^{-(1+\gamma)}) \\
&+ \big(2(\frac{8}{\mu}L_{s,f}^2+1)C_2^2M_f^2\big)\mathcal{O}(t^{-1}) + 2(\frac{8}{\mu}L_{s,f}^2+1)C_d^2 \\
&+ L_{s,f}^2C_1^2C_4^2\mathcal{O}(t^{-2})\Big)
\end{aligned} \tag{103}$$

Thus,

$$\begin{aligned}
\Big\||\theta_t - \theta_{t-1}^*|_s\Big\|_q^2 \le{}& \frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)C_d^2 + \frac{8}{\mu}C_1^2(C_{s,q}^2m^{-1} + 2L_{s,f}^2C_d^2)\cdot O\big(t^{-\gamma}\big) \\
&+ \frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)C_2^2M_f^2\cdot\mathcal{O}(t^{-1}) + \frac{16}{\mu^2}C_4^2\cdot\mathcal{O}(t^{-(2-2\gamma)}) + \tilde{\mathcal{R}}_t^2,
\end{aligned} \tag{104}$$

where the remainder term $\tilde{\mathcal{R}}_t^2$ collects all error terms that decay faster in t or grow slower in d than the rates displayed above, i.e.

$$\tilde{\mathcal{R}}_t^2 := \exp\Big(-\frac{\mu}{2}S_t\Big)X_0^2 + \frac{2}{\mu}\Big(C_1^2C_4^2\mathcal{O}(t^{-(2-\gamma)}) + 8C_1^2L_{s,f}^2C_2^2M_f^2\mathcal{O}(t^{-(1+\gamma)}) + L_{s,f}^2C_1^2C_4^2\mathcal{O}(t^{-2})\Big). \tag{105}$$

We keep terms related to C_2, C_4, C_d are they are related to dimension d. Taking square roots on both side, and applying inequality $(\sum a_i)^{\frac{1}{2}} \le \sum a_i^{\frac{1}{2}}$, we obtain

$$\begin{aligned}
\Big\||\theta_t - \theta_{t-1}^*|_s\Big\|_q \le{}& \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)}\,C_d + \sqrt{\frac{8}{\mu}C_1^2(C_{s,q}^2m^{-1} + 2L_{s,f}^2C_d^2)}\cdot O\big(t^{-\frac{\gamma}{2}}\big) \\
&+ \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)}C_2M_f\cdot\mathcal{O}(t^{-\frac{1}{2}}) + \frac{4}{\mu}C_4\cdot\mathcal{O}(t^{-(1-\gamma)}) + \tilde{\mathcal{R}}_t,
\end{aligned} \tag{106}$$

Substituting the definitions of all constants C_1, C_2, C_4 and C_d, we obtain

$$\begin{aligned}\Big\||\theta_t - \theta^*_{t-1}|_s\Big\|_q &\le \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)}\, C\left(\frac{K_u^2}{\lambda_k(\Sigma_f)}M_f + K_u(k^{1/2}+q^{1/2}) + k^{1/2}M_f\right)\cdot d^{-\frac{1}{2}} \\ &+ \sqrt{\frac{8}{\mu}(\max\{q,s\}-1)\Big(C_{s,q}^2 m^{-1} + 2L_{s,f}^2 C^2\big[\frac{K_u^2}{\lambda_k(\Sigma_f)}M_f + K_u(k^{1/2}+q^{1/2}) + k^{1/2}M_f\big]^2\cdot d^{-1}\Big)}\cdot O\big(t^{-\frac{\gamma}{2}}\big) \\ &+ \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2+1)}\cdot Ck^{1/2}(C_{\alpha,\beta}+\frac{q^{1/2}}{\bar\rho_k})\cdot M_f\cdot\mathcal{O}(t^{-\frac{1}{2}}) \\ &+ \frac{4}{\mu}\cdot C\frac{L_{s,f}}{\mu}M_f(M_A k^{1/2}\rho_k^{-1} + d^{-1/2})\cdot\mathcal{O}(t^{-(1-\gamma)}) + \tilde{\mathcal{R}}_t,\end{aligned} \tag{107}$$

To make the dependence on d, m, and t explicit, we denote

$$\begin{aligned}A_1 &:= C\sqrt{\frac{1}{\mu}\left(\frac{1}{\mu}L_{s,f}^2+1\right)}\left(\frac{K_u^2}{\lambda_k(\Sigma_f)}M_f + K_u(k^{1/2}+q^{1/2}) + k^{1/2}M_f\right), \\ A_2 &:= C\sqrt{\frac{\max\{q,s\}-1}{\mu}}\left(C_{s,q} + L_{s,f}\left[\frac{K_u^2}{\lambda_k(\Sigma_f)}M_f + K_u(k^{1/2}+q^{1/2}) + k^{1/2}M_f\right]\right), \\ A_3 &:= C\sqrt{\frac{1}{\mu}\left(\frac{1}{\mu}L_{s,f}^2+1\right)}\cdot M_f\, k^{1/2}, \\ A_4 &:= C\frac{L_{s,f}}{\mu^2}M_f,\end{aligned} \tag{108}$$

where C is absolute constant.

Combining the above notations, we obtain the stated bound in Theorem 3.6, i.e.

$$\begin{aligned}\Big\||\theta_t - \theta^*_{t-1}|_s\Big\|_q &\le A_1 d^{-1/2} + A_2(m^{-1/2}+d^{-1/2})t^{-\gamma/2} \\ &\quad + A_3(C_{\alpha,\beta}+\frac{q^{1/2}}{\bar\rho_k})\,t^{-1/2} + A_4(M_A k^{1/2}\rho_k^{-1}+d^{-1/2})t^{-(1-\gamma)} + \widetilde{\mathcal{R}}_t,\end{aligned}$$

□

Remark D.22. The terms involving C_2 reflect the improvement of subspace distance estimates along the trajectory of the online PCA updates, whereas the terms involving the growing factor C_4 arise from the minimizer shift induced by online PCA updates.

Lemma D.23. *Let $\{\eta_k\}_{k\ge 0}$ be a non-increasing sequence of nonnegative numbers and define $S_t := \sum_{j=0}^{t-1}\eta_j$, $t \ge 1$. Fix $\lambda > 0$, then for all $t \ge 1$,*

$$\sum_{k=0}^{t-1}\eta_k \exp\Big(-\lambda(S_t - S_{k+1})\Big) \le \frac{1}{\lambda}. \tag{109}$$

Suppose $\sum_{k=0}^{\infty}\eta_k = \infty$ and $\eta_k \to 0$, we then have

$$\sum_{k=0}^{t-1}\eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) \to 0\,, \text{ as } t\to\infty. \tag{110}$$

$$\sum_{k=0}^{t-1}\eta_k \frac{1}{k+1}\exp\Big(-\lambda(S_t - S_{k+1})\Big) \to 0\,, \text{ as } t\to\infty. \tag{111}$$

Additionally assuming $(k+1)\eta_k \to \infty$, it then holds that

$$\sum_{k=0}^{t-1}\eta_k^{-1}\frac{1}{(k+1)^2}\exp\Big(-\lambda(S_t - S_{k+1})\Big) \to 0\,, \text{ as } t\to\infty. \tag{112}$$

Proof. For any $u \in [S_k, S_{k+1})$, we have $u \le S_{k+1}$, hence

$$S_t - u \;\ge\; S_t - S_{k+1}, \text{ then } \quad e^{-\lambda(S_t - S_{k+1})} \;\le\; e^{-\lambda(S_t-u)}.$$

Consequently

$$\eta_k\, e^{-\lambda(S_t-S_{k+1})} = \Big(S_{k+1}-S_k\Big)\, e^{-\lambda(S_t-S_{k+1})} = \int_{S_k}^{S_{k+1}} e^{-\lambda(S_t-S_{k+1})}\, du \le \int_{S_k}^{S_{k+1}} e^{-\lambda(S_t-u)}\, du.$$

Summing over $k = 0, \dots, t-1$, it holds that

$$\sum_{k=0}^{t-1} \eta_k\, e^{-\lambda(S_t-S_{k+1})} \le \int_0^{S_t} e^{-\lambda(S_t-u)}\, du = \int_0^{S_t} e^{-\lambda v}\, dv = \frac{1-e^{-\lambda S_t}}{\lambda} \le \frac{1}{\lambda},$$

which proves (109).

To show (110), we first fix $K > 0$, then

$$\sum_{k=0}^{t-1} \eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) = \sum_{k=0}^{K-1} \eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) + \sum_{k=K}^{t-1} \eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) \tag{113}$$

Since $S_t \to \infty$, the first term on RHS of (113) goes to 0 as $t \to \infty$. Note that $\{\eta_k\}$ is non-increasing, for all $k \ge K$ we have $\eta_k^2 \le \eta_k\, \eta_K$. Thus followed by (113), the second term on the RHS can be controlled as

$$\sum_{k=K}^{t-1} \eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) \le \eta_K \sum_{k=K}^{t-1} \eta_k \exp\Big(-\lambda(S_t - S_{k+1})\Big) \le \frac{\eta_K}{\lambda}$$

We complete the proof of (110) holds by taking $K \to \infty$. With the same argument, (111) holds since $\frac{1}{K+1} \to 0$.

For (112), similarly, for fixed $K > 0$,

$$\begin{aligned}
&\sum_{k=0}^{t-1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\lambda(S_t - S_{k+1})\Big) \\
&\quad = \sum_{k=0}^{K-1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\lambda(S_t - S_{k+1})\Big) + \sum_{k=K}^{t-1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\lambda(S_t - S_{k+1})\Big)
\end{aligned} \tag{114}$$

The first term goes to 0 as $t \to \infty$, and the second term

$$\sum_{k=K}^{t-1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\lambda(S_t - S_{k+1})\Big) = \sum_{k=K}^{t-1} \frac{1}{\eta_k^2 (k+1)^2} \eta_k \exp\Big(-\lambda(S_t - S_{k+1})\Big) \tag{115}$$

$$\le \frac{1}{\eta_K^2 (K+1)^2} \cdot \frac{1}{\lambda} \tag{116}$$

The last inequality holds since $\eta_k^2 (k+1)^2$ is monotone increasing to infinity. We complete the proof of (112) holds by taking $K \to \infty$.

□

Lemma D.24 (Polynomial step size and exponentially weighted sums)**.** *Suppose $\eta_k = \Theta((k+1)^{-\gamma})$ for some $0 < \gamma < 1$, i.e. there exist absolute constants $0 < a \le b < \infty$ such that for all $k \ge 0$,*

$$a(k+1)^{-\gamma} \le \eta_k \le b(k+1)^{-\gamma}. \tag{117}$$

Then, $S_t = \Theta(t^{1-\gamma})$, and the three weighted sums admit the following rates as $t \to \infty$:

$$\sum_{k=0}^{t-1} \eta_k^2 \exp\Big(-\lambda(S_t - S_{k+1})\Big) = \frac{1}{\lambda} \cdot O\big(t^{-\gamma}\big), \tag{118}$$

$$\sum_{k=0}^{t-1} \eta_k \frac{1}{k+1} \exp\Big(-\lambda(S_t - S_{k+1})\Big) = \frac{1}{\lambda} \cdot O\big(t^{-1}\big), \tag{119}$$

$$\sum_{k=0}^{t-1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\lambda(S_t - S_{k+1})\Big) = \frac{1}{\lambda} \cdot O\big(t^{-(2-2\gamma)}\big). \tag{120}$$

In particular, all three are $o(1)$ since $\gamma \in (0,1)$.

Proof. Recall $S_t - S_{k+1} = \sum_{j=k+1}^{t-1} \eta_j$. We first derive an uniform lower bound on the tail sum $S_t - S_{k+1}$, by splitting indices into "early" $k \le \lfloor t/2 \rfloor$ and "late" $k > \lfloor t/2 \rfloor$.

(i) Early indices $0 \le k \le \lfloor t/2 \rfloor$. Then $k+1 \le t/2+1$ and

$$S_t - S_{k+1} \;\ge\; \sum_{j=\lfloor t/2 \rfloor}^{t-1} \eta_j \;\ge\; a \sum_{j=\lfloor t/2 \rfloor}^{t-1} (j+1)^{-\gamma} \;\ge\; a \cdot \frac{t}{2} \cdot t^{-\gamma} \;=\; \frac{a}{2}\, t^{1-\gamma}.$$

Hence, for all such k,

$$\exp\big(-\lambda(S_t - S_{k+1})\big) \;\le\; \exp\Big(-\frac{\lambda a}{2}\, t^{1-\gamma}\Big). \tag{121}$$

(ii) Late indices $\lfloor t/2 \rfloor < k \le t-1$. For $j \in \{k+1, \dots, t-1\}$ we have $j+1 \le t$, hence $(j+1)^{-\gamma} \ge t^{-\gamma}$, and thus

$$S_t - S_{k+1} \;\ge\; a \sum_{j=k+1}^{t-1} (j+1)^{-\gamma} \;\ge\; a\,(t-k-1)\, t^{-\gamma}.$$

Consequently, for all such k,

$$\exp\big(-\lambda(S_t - S_{k+1})\big) \;\le\; \exp\big(-\lambda a (t-k-1) t^{-\gamma}\big). \tag{122}$$

We will repeatedly use the elementary bound

$$\sum_{u=0}^{\infty} e^{-c\,u\,t^{-\gamma}} \;=\; \frac{1}{1-e^{-ct^{-\gamma}}} \;\le\; \frac{2}{c}\, t^{\gamma} \qquad \text{for all sufficiently large } t, \tag{123}$$

which follows from $1 - e^{-x} \ge x/2$ for small $x > 0$. For notation simplicity, we denote C as absolute constants which may differ from line to line.

Now we prove (118). Let

$$A_t := \sum_{k=0}^{t-1} \eta_k^2\, e^{-\lambda(S_t - S_{k+1})} = \sum_{k=0}^{\lfloor t/2 \rfloor} \eta_k^2\, e^{-\lambda(S_t - S_{k+1})} + \sum_{k=\lfloor t/2 \rfloor + 1}^{t-1} \eta_k^2\, e^{-\lambda(S_t - S_{k+1})} =: A_t^{\mathrm{early}} + A_t^{\mathrm{late}},$$

where the split is at $\lfloor t/2 \rfloor$.

Early part. By (121) and $\eta_k^2 \le b^2 (k+1)^{-2\gamma} \le b^2$,

$$A_t^{\mathrm{early}} \le \exp\Big(-\frac{\lambda a}{2}\, t^{1-\gamma}\Big) \sum_{k=0}^{\lfloor t/2 \rfloor} \eta_k^2 \le b^2 t \exp\Big(-\frac{\lambda a}{2}\, t^{1-\gamma}\Big) = o(t^{-m}) \quad \forall m > 0.$$

Late part. For $\lfloor t/2 \rfloor < k \le t-1$, we have $(k+1) \asymp t$, hence $\eta_k^2 \asymp b^2 (k+1)^{-2\gamma} \le b^2 t^{-2\gamma} = C t^{-2\gamma}$. Using (122) and the change of variables $u := t - k - 1 \in \{0, 1, \dots\}$,

$$A_t^{\mathrm{late}} \le C t^{-2\gamma} \sum_{k=\lfloor t/2 \rfloor + 1}^{t-1} e^{-\lambda a (t-k-1) t^{-\gamma}} = C t^{-2\gamma} \sum_{u=0}^{t - \lfloor t/2 \rfloor - 2} e^{-\lambda a u t^{-\gamma}} \le C t^{-2\gamma} \sum_{u=0}^{\infty} e^{-\lambda a u t^{-\gamma}}.$$

Applying (123) with $c = \lambda a$ gives

$$A_t^{\mathrm{late}} \le C t^{-2\gamma} \cdot \frac{2}{c} t^{\gamma} = \frac{2b^2}{\lambda a} t^{-\gamma}.$$

Combining early and late parts yields $A_t = \frac{1}{\lambda} \cdot O(t^{-\gamma})$, proving (118).

For (119), let

$$B_t := \sum_{k=0}^{t-1} \eta_k \frac{1}{k+1} e^{-\lambda(S_t - S_{k+1})} = B_t^{\text{early}} + B_t^{\text{late}},$$

where the split is at $\lfloor t/2 \rfloor$.

Early part. Using (121) and $\eta_k/(k+1) \le b(k+1)^{-(1+\gamma)} \le b$,

$$B_t^{\text{early}} \le \exp\Big(-\frac{\lambda a}{2} t^{1-\gamma}\Big) \sum_{k=0}^{\lfloor t/2 \rfloor} \eta_k \frac{1}{k+1} \le Ct \exp\Big(-\frac{\lambda a}{2} t^{1-\gamma}\Big) = o(t^{-m}) \quad \forall m > 0.$$

Late part. For $k > \lfloor t/2 \rfloor$ we have $(k+1) \asymp t$, hence

$$\eta_k \frac{1}{k+1} \le b(k+1)^{-(1+\gamma)} \le Ct^{-(1+\gamma)}.$$

Therefore, by (122) and (123),

$$B_t^{\text{late}} \le Ct^{-(1+\gamma)} \sum_{k=\lfloor t/2 \rfloor + 1}^{t-1} e^{-\lambda a (t-k-1) t^{-\gamma}} \le Ct^{-(1+\gamma)} \sum_{u=0}^{\infty} e^{-\lambda a u t^{-\gamma}} \le \frac{C}{\lambda} t^{-(1+\gamma)} \cdot t^{\gamma} = \frac{C}{\lambda} t^{-1}.$$

Now we obtain (119).

For (120), let

$$C_t := \sum_{k=0}^{t-1} \eta_k^{-1} \frac{1}{(k+1)^2} e^{-\lambda(S_t - S_{k+1})} = C_t^{\text{early}} + C_t^{\text{late}},$$

where the split is at $\lfloor t/2 \rfloor$.

Early part. Using (121) and $\eta_k^{-1}/(k+1)^2 \le a^{-1}(k+1)^{\gamma-2} \le a^{-1}$,

$$C_t^{\text{early}} \le \exp\Big(-\frac{\lambda a}{2} t^{1-\gamma}\Big) \sum_{k=0}^{\lfloor t/2 \rfloor} \eta_k^{-1} \frac{1}{(k+1)^2} \le Ct \exp\Big(-\frac{\lambda a}{2} t^{1-\gamma}\Big) = o(t^{-m}) \quad \forall m > 0.$$

Late part. For $k > \lfloor t/2 \rfloor$ we have $(k+1) \asymp t$, hence

$$\eta_k^{-1} \frac{1}{(k+1)^2} \le a^{-1}(k+1)^{\gamma-2} \le Ct^{\gamma-2} = Ct^{-(2-\gamma)}.$$

Therefore, by (122) and (123),

$$C_t^{\text{late}} \le Ct^{-(2-\gamma)} \sum_{k=\lfloor t/2 \rfloor + 1}^{t-1} e^{-\lambda a (t-k-1) t^{-\gamma}} \le Ct^{-(2-\gamma)} \sum_{u=0}^{\infty} e^{-\lambda a u t^{-\gamma}} \le \frac{C}{\lambda} t^{-(2-\gamma)} \cdot t^{\gamma} = \frac{C}{\lambda} t^{-(2-2\gamma)}.$$

This proves (120). □

D.7 Proofs for Corollary 3.9 in Section 3

An immediate corollary of Theorem 3.6 is that an improved convergence rate can be established when the output of Oja's iteration is frozen after iteration $\lfloor t/2 \rfloor$. Note that the same rate of convergence can also be achieved when the factor subspace is estimated using offline PCA.

Proof for Corollary 3.9. It suffices to consider the case $t_1 = \lfloor t/2 \rfloor$, as the case $t_1 = \lfloor t/c_0 \rfloor$ follows from identical arguments.

We keep using the proof constants C_1, C_2, C_3, C_4 in the proof of Theorem 3.6. By following the same proof strategy as Theorem 3.6, we obtain the recursive inequality holds for all $1 \le i \le t$,

$$\begin{aligned} X_i^2 \le & \Big(1 - \frac{\mu}{2}\eta_{i-1}\Big) X_{i-1}^2 + \frac{8}{\mu}\eta_{i-1}^{-1}\kappa_{i-1}^2 + C_1^2 \kappa_{i-1}^2 \\ & + 4C_1^2 \eta_{i-1}^2 \Xi_{i-1}^2 + (\frac{8}{\mu} L_{s,f}^2 + 1)\eta_{i-1} F_{i-1}^2 + C_1^2 L_{s,f}^2 \kappa_{i-1}^2 \eta_{i-1}. \end{aligned} \tag{124}$$

For iterations $i \geq \lfloor \frac{t}{2} \rfloor$, the minimizer shift no longer occurs, so that $\theta_i^* = \theta_{\lfloor \frac{t}{2} \rfloor - 1}^*$. Under this regime, we can rewrite κ, F, Ξ as: For minimizer shift $\kappa_i (\geq \left\| |\theta_{i-2}^* - \theta_{i-1}^*|_s \right\|_q)$,

$$\kappa_{i-1} = \begin{cases} C_4 i^{-1}, & i \leq \lfloor \frac{t}{2} \rfloor, \\ 0, & i > \lfloor \frac{t}{2} \rfloor. \end{cases}$$

For factor estimation error

$$F_{i-1} = \begin{cases} C_2 M_f i^{-1/2} + C_3 d^{-1/2}, & i \leq \lfloor \frac{t}{2} \rfloor, \\ \sqrt{2} C_2 M_f t^{-1/2} + C_3 d^{-1/2}, & i > \lfloor \frac{t}{2} \rfloor. \end{cases}$$

And for Ξ which is related to mini-batch concentration

$$\Xi_{i-1}^2 = C_{s,q}^2 m^{-1} + L_{s,f}^2 F_{i-1}^2$$

Now we plug-in the definition of κ, F, Ξ and apply (100) to solve the recursive bound from $k = 0$.

Note that for the summation up to index $t_a := \lfloor ta \rfloor$, where $\frac{1}{2} < a \leq 1$, we can split the sum at $\lfloor t/2 \rfloor$. The early part has the same structure as in the proof of Lemma D.24, while the late part involves only the terms F_k and Ξ_k since $\kappa_k = 0$. Therefore, we obtain

$$\begin{aligned}
X_{t_a}^2 \leq & \exp\Big(-\frac{\mu}{2} S_{t_a} \Big) X_0^2 + \frac{8}{\mu} C_4^2 \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k^{-1} \frac{1}{(k+1)^2} \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + C_1^2 C_4^2 \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \frac{1}{(k+1)^2} \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + 4 C_1^2 (C_{s,q}^2 m^{-1} + 2 L_{s,f}^2 C_d^2) \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k^2 \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + 8 C_1^2 L_{s,f}^2 C_2^2 M_f^2 \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k^2 \frac{1}{k+1} \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + \big(2 (\frac{8}{\mu} L_{s,f}^2 + 1) C_2^2 M_f^2 \big) \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k \frac{1}{k+1} \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + 2 (\frac{8}{\mu} L_{s,f}^2 + 1) C_d^2 \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + L_{s,f}^2 C_1^2 C_4^2 \sum_{k=0}^{\lfloor \frac{t}{2} \rfloor - 1} \eta_k \frac{1}{(k+1)^2} \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + 4 C_1^2 \sum_{k=\lfloor \frac{t}{2} \rfloor}^{t_a - 1} \Xi_k^2 \eta_k^2 \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big) \\
& + (\frac{8}{\mu} L_{s,f}^2 + 1) \sum_{k=\lfloor \frac{t}{2} \rfloor}^{t_a - 1} F_k^2 \eta_k \exp\Big(-\frac{\mu}{2} (S_{t_a} - S_{k+1}) \Big)
\end{aligned} \tag{125}$$

According to the proof of Lemma D.24, all early-stage terms are of order $\Theta\Big(t_a \exp\{-\frac{\mu}{4}{t_a}^{1-\gamma}\}\Big)$, so that decay faster than any polynomial rate. Thus, we could rewrite (125) as

$$\begin{aligned} X_{t_a}^2 \leq & \exp\Big(-\frac{\mu}{2}S_{t_a}\Big)X_0^2 + \Theta\Big(t_a \exp\{-\tfrac{\mu}{4}{t_a}^{1-\gamma}\}\Big) \\ & + 4C_1^2 \sum_{k=\lfloor \frac{t}{2} \rfloor}^{t_a-1} \Xi_k^2 \eta_k^2 \exp\Big(-\frac{\mu}{2}(S_{t_a} - S_{k+1})\Big) \\ & + (\frac{8}{\mu}L_{s,f}^2 + 1) \sum_{k=\lfloor \frac{t}{2} \rfloor}^{t_a-1} F_k^2 \eta_k \exp\Big(-\frac{\mu}{2}(S_{t_a} - S_{k+1})\Big) \end{aligned} \tag{126}$$

When $k \geq \lfloor \frac{t}{2} \rfloor$, it holds that

$$\begin{aligned} F_k &= \sqrt{2}C_2 M_f t^{-1/2} + C_3 d^{-1/2} = F_{\lfloor \frac{t}{2} \rfloor} \\ \Xi_k^2 &= C_{s,q}^2 m^{-1} + L_{s,f}^2(\sqrt{2}C_2 M_f t^{-1/2} + C_3 d^{-1/2})^2 =: \Xi_{\lfloor \frac{t}{2} \rfloor}^2 \end{aligned}$$

which are only related to t.

Thus, it holds for (126) that

$$\begin{aligned} X_{t_a}^2 \leq & \exp\Big(-\frac{\mu}{2}S_{t_a}\Big)X_0^2 + 4C_1^2 \Xi_{\lfloor \frac{t}{2} \rfloor}^2 \sum_{k=0}^{t_a-1} \eta_k^2 \exp\Big(-\frac{\mu}{2}(S_{t_a} - S_{k+1})\Big) \\ & + (\frac{8}{\mu}L_{s,f}^2 + 1)F_{\lfloor \frac{t}{2} \rfloor}^2 \sum_{k=0}^{t_a-1} \eta_k \exp\Big(-\frac{\mu}{2}(S_{t_a} - S_{k+1})\Big) + \Theta\Big(t_a \exp\{-\tfrac{\mu}{4}{t_a}^{1-\gamma}\}\Big), \end{aligned} \tag{127}$$

where, after extracting the constants $F_{\lfloor \frac{t}{2} \rfloor}$ and $\Xi_{\lfloor \frac{t}{2} \rfloor}$ from the summation, we extend the lower summation index to 0.

Applying Lemma D.24 with $\lambda = \frac{\mu}{2}$, we have

$$\Big\| |\theta_{t_a} - \theta^*_{\lfloor \frac{t}{2} \rfloor - 1}|_s \Big\|_q^2 \leq \frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2 + 1)F_{\lfloor \frac{t}{2} \rfloor}^2 + 4C_1^2 \Xi_{\lfloor \frac{t}{2} \rfloor}^2 \cdot O(t_a^{-\gamma}) + \exp\Big(-\frac{\mu}{2}S_{t_a}\Big)X_0^2 + \Theta\Big(t_a \exp\{-\tfrac{\mu}{4}{t_a}^{1-\gamma}\}\Big)$$

We complete the proof by taking square root on both side, and applying inequality $\sqrt{\sum a_i} \leq \sum \sqrt{a_i}$.

$$\Big\| |\theta_{t_a} - \theta^*_{\lfloor \frac{t}{2} \rfloor - 1}|_s \Big\|_q \leq \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2 + 1)}F_{\lfloor \frac{t}{2} \rfloor} + 2C_1 \Xi_{\lfloor \frac{t}{2} \rfloor} \cdot O\big(t_a^{-\frac{\gamma}{2}}\big) + \mathcal{R}_t, \tag{128}$$

where $\mathcal{R}_t$ decays faster than any polynomial rate. Specifically, when $a = 1$, we obtain that

$$\Big\| |\theta_t - \theta^*_{\lfloor \frac{t}{2} \rfloor - 1}|_s \Big\|_q \leq \sqrt{\frac{4}{\mu}(\frac{8}{\mu}L_{s,f}^2 + 1)}F_{\lfloor \frac{t}{2} \rfloor} + 2C_1 \Xi_{\lfloor \frac{t}{2} \rfloor} \cdot O\big(t^{-\frac{\gamma}{2}}\big) + \mathcal{R}_t, \tag{129}$$

□

Remark D.25 (Proof constants). For clarity, we collect the constants appearing in the proof. Let

$$M_f := \Big\| \, |f|_2 \, \Big\|_q, \qquad C_1^2 := \max\{q, s\} - 1.$$

Furthermore, define

$$\begin{aligned} C_2 &:= Ck^{1/2}(C_{\alpha,\beta} + \frac{q^{1/2}}{\bar{\rho}_k}), \\ C_3 &:= C\left(\frac{K_u^2}{\lambda_k(\Sigma_f)}M_f + K_u(k^{1/2} + q^{1/2}) + k^{1/2}M_f\right)d^{-1/2}, \\ C_4 &:= C\frac{L_{s,f}}{\mu}M_f(M_A k^{1/2}\rho_k^{-1} + d^{-1/2}), \\ C_d &:= C_3 d^{-1/2}. \end{aligned}$$

We also note that $C_{s,q}$ is a constant of order $\mathcal{O}(p^{1/s})$ depending on s, q, and $\mathbb{E}[|\nabla l(\theta; f, y)|_q]$, introduced in Lemma D.21.

These constants are independent of the iteration index t and are used only to simplify the notation in the proof.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions of the paper: the proposed FSGD algorithm, streaming online factor learning, and the theoretical analysis of FSGD. These claims are supported by the formal algorithmic description in Section 2, the convergence results under the stated assumptions in Section 3, and the synthetic and real-data experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our work. For example, we clarify that our theory is developed for strongly convex population objectives in Section 1. Neural network simulations are provided in Appendix B.2 only as empirical evidence beyond the convex setting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results state or reference the required assumptions explicitly. Complete proofs of the main propositions and theorems, together with the auxiliary lemmas used in the analysis, are provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: For the synthetic experiments, the paper provides the data generating process, optimization details, hyperparameters, evaluation metrics, and number of repetitions. For the real-data experiment, the dataset, prediction task, train/test period, evaluation metric, and baseline configurations are also described. Full experimental details are provided in Appendices B and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation code is included in the supplementary materials. The synthetic data-generating processes are described in Appendix B. The real-data experiments use publicly available MNIST, CIFAR-10, and NCEP/NCAR Reanalysis 1 data, with experimental settings provided in Appendix C.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the main experimental settings, including data generation, model dimensions, optimization details, hyperparameters, baselines, evaluation metrics, and repetitions. Further details are provided in Appendix B and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Synthetic experiments report variability across repeated runs. For MNIST and CIFAR-10, we report means and standard deviations across 30 split means, each averaging five training seeds, together with paired confidence intervals and significance tests. The NCEP table reports a single run at four selected locations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computational environment, memory allocation, and approximate compute budget used for the experiments in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have all reviewed the NeurIPS Code of Ethics and striven to maintain and preserve anonymity.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is primarily a theoretical study of online latent factor learning and stochastic gradient descent. It does not introduce a deployed system, a dataset involving human subjects, or an application specific decision making pipeline. We therefore do not identify direct societal impacts beyond the general possibility that more scalable optimization methods may be used in a wide range of downstream applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: As a theoretical and algorithmic study of online latent factor learning and stochastic gradient descent, this paper does not release any high risk models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The experiments use the publicly available MNIST, CIFAR-10, and NCEP/NCAR Reanalysis 1 datasets. Their experimental use and preprocessing are described in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the full FSGD implementation as a new asset in the supplementary materials to support reproducibility of the experiments, with plans to publish it on GitHub upon acceptance.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. It uses only synthetic data and a publicly available climate reanalysis dataset, so participant instructions and compensation details are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper uses only synthetic data and a publicly available climate reanalysis dataset, so IRB approval, participant risk disclosure, and consent procedures are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The theoretical and methodology advancements in this work are built on well-established mathematical tools, and LLM is not utilized for developing them.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.