Stochastic Regularized Majorization-Minimization with weakly convex and multi-convex surrogates

Hanbaek Lyu[*]   HLYU@MATH.WISC.EDU
Department of Mathematics
University of Wisconsin - Madison, WI 53706, USA

Editor: Pradeep Ravikumar

Abstract

Stochastic majorization-minimization (SMM) is a class of stochastic optimization algorithms that proceed by sampling new data points and minimizing a recursive average of surrogate functions of an objective function. The surrogates are required to be strongly convex and the existing convergence rate analysis for the general non-convex setting was not available. In this paper, we propose an extension of SMM where surrogates are allowed to be only weakly convex or block multi-convex, and the averaged surrogates are approximately minimized with proximal regularization or block-minimized within diminishing radii, respectively. For the general nonconvex constrained setting with non-i.i.d. data samples, we show that the first-order optimality gap of the proposed algorithm decays at the rate $\widetilde{O}(n^{-1/4})$ for the empirical loss and $\widetilde{O}(n^{-1/8})$ for the expected loss, where n denotes the number of data samples processed. Under some additional assumption, the latter convergence rate can be improved to $\widetilde{O}(n^{-1/4})$. As a corollary, we obtain the first convergence rate bounds for various optimization methods under general nonconvex non-i.i.d. data setting: Double-averaging projected gradient descent and its generalizations, proximal point empirical risk minimization, and online matrix/tensor decomposition algorithms. We also provide experimental validation of our results.

Keywords: Nonconvex stochastic optimization, stochastic majorization-minimization, convergence rate, Markovian data, double-averaging, online matrix/tensor factorization, deep learning

Contents

[*]. All codes are available at https://github.com/HanbaekLyu/SRMM

1. Introduction

Empirical loss minimization is a classical problem setting regarding parameter estimation with a growing number of observations, where one seeks to minimize a recursively defined empirical loss function as new data arrives. Some of its well-known applications include maximum likelihood estimation, or more generally, M-estimation Geyer (1994); Geer et al. (2000); Stefanski and Boos (2002), as well as the online dictionary learning literature Mairal et al. (2010); Mairal (2013b); Mensch et al. (2017); Lyu et al. (2020). On the other hand, the *expected loss minimization* seeks to estimate a parameter by minimizing the loss function with respect to random data. It provides a general framework for stochastic optimization literature Schneider and Kirkpatrick (2007); Marti (2005); Bottou and Bousquet (2008); Nemirovski et al. (2009). Optimization algorithms for empirical or expected loss minimization are in nature 'online', meaning that sampling new data points and adjusting the current estimation occurs recursively. Such online algorithms have proven to be particularly efficient in large-scale problems in statistics, optimization, and machine learning Bottou (1998); Duchi and Singer (2009); Ghadimi and Lan (2013); Kingma and Ba (2014).

First-order methods for expected loss minimization include (projected or proximal) *Stochastic Gradient Descent* (SGD) have been extensively studied for many decades, which usually consist of stochastically estimating the full gradient of the expected loss function and performing a gradient descent step followed by a projection onto the parameter space. For a general constrained and nonconvex (weakly convex) setting, global convergence to stationary (first-order optimal) points of such methods was recently established in Davis and Drusvyatskiy (2019) with a rate of convergence $O(\log n/\sqrt{n})$, where n denotes the number of iterations (processed samples). This result assumes arbitrary initialization, although faster convergence with special initialization is known in a matrix factorization setting Wang et al. (2017). Throughout this paper, we are concerned with the convergence of online algorithms with arbitrary initialization, which is often referred to as a 'global convergence' in the literature.

On the other hand, *Stochastic Majorization-Minimization* (SMM) is one of the most popular approaches that directly solve empirical loss minimization Mairal (2013b) by sampling data points from a target data distribution and minimizing a recursively defined majorizing surrogate of the empirical loss function. This method is an online extension of the classical majorization-minimization Lange et al. (2000) principle and encompasses Expectation-Minimization in statistics Neal and Hinton (1998); Cappé and Moulines (2009), and also generalizes the celebrated online matrix factorization algorithm in Mairal et al. (2010). While it was observed empirically that SMM shows competitive performance while requiring less parameter tuning than SGD for online dictionary learning problems Mairal et al. (2010), the theoretical convergence guarantee of SMM in the literature only ensures asymptotic convergence to stationary points and lacks any convergence rate bounds Mairal et al. (2010); Mairal (2013b,a); Mensch et al. (2017).

Most theoretical analyses on online optimization algorithms assume the ability to sample i.i.d. data points from the target distribution. This is a classical and convenient assumption for analysis, but it is violated frequently in practice, especially when the samples are accessed through Markov chain-based methods. A common practice is to first sample a single Markov chain trajectory of dependent data samples (after some burn-in period), which is

then thinned (subsampled) to reduce the dependence. However, such practice only weakens the dependence between data samples and does not make it truly independent in general. One may re-initialize independent Markov chains for every sample to attain true independence, but this approach suffers from a huge computational burden. Optimization algorithms with Markovian data samples were studied in Johansson et al. (2007, 2010) in the context of distributed optimization in networks. More recently, it was shown in Sun et al. (2018) that arbitrarily initialized SGD almost surely converges to critical points of unconstrained nonconvex objectives at rate $O((\log n)^2/\sqrt{n})$, even when the data samples have a Markovian dependence. Block Coordinate Descent with Markovian coordinate selection was also studied recently in Sun et al. (2020).

On the other hand, in Lyu et al. (2020), the online matrix factorization algorithm in Mairal et al. (2010) based on SMM is extended and shown to converge to stationary points for constrained matrix factorization problems in the Markovian setting. Based on the result and combined with an MCMC network sampling algorithm in Lyu et al. (2023), an online dictionary learning algorithm for learning latent motifs in networks is proposed in Lyu et al. (2020). More recently, an online algorithm for nonnegative tensor dictionary learning utilizing CANDECOMP/PARAFAC (CP) decomposition is developed in Lyu et al. (2022), where convergence to stationary points under the Markovian setting is established. The algorithm is based on SMM but uses new components such as block coordinate descent with diminishing radius Lyu and Li (2023). Although the works Lyu et al. (2020, 2022) are the first to establish global convergence of SMM-type methods to stationary points in the Markovian setting, there has not been a convergence rate bounds, which is still the case in the i.i.d. setting. A rate of convergence result is also important in practice since it provides bounds on the number of iterations and data samples sufficient to guarantee obtaining an approximate solution up to the desired precision.

Our goal in this paper is twofold. First, we generalize the framework of SMM so that not only strongly convex surrogates can be used, but also the more general class of weakly convex or multi-convex surrogates can be used with suitable regularization. Second, we intend to provide the missing convergence rate analysis of SMM for general constrained nonconvex objectives, both in the context of empirical and expected loss minimization. We call our algorithm *stochastic regularized majorization-minimization* (SRMM), which generalizes the original SMM Mairal (2013a), online matrix factorization Mairal et al. (2010); Lyu et al. (2020), stochastic approximate majorization-minimization and subsampled online matrix factorization Mensch et al. (2017), and the online CP-dictionary learning algorithm Lyu et al. (2022).

1.1 Contribution

Our algorithm and analysis of SMM consider three cases: 1) Strongly convex surrogates without regularization; 2) Multi-convex surrogates with block coordinate descent with diminishing radius; 3) Strongly convex surrogates with proximal regularization. A concise summary of our results is given in Table 1 and in the following bullet points.

SMM with strongly convex surrogates

- Rate of convergence of classical SMM with strongly convex surrogates and i.i.d. data samples is established: $O((\log n)^{1+\varepsilon} n^{-1/2})$ for empirical loss and $O((\log n)^{1+\varepsilon} n^{-1/4})$ for the expected loss.

- The results also hold against inexact computation of surrogates, inexact surrogate minimization, and hidden Markovian dependence in data samples.

SMM with multi-convex surrogates

- SMM with block multi-convex surrogates is developed, where averaged surrogates are approximately block-minimized within a diminishing radius, either with cyclically or randomly chosen block coordinates.

- Coordinate descent for surrogate minimization reduces computational cost in high dimension.

- Same convergence rate results as in the previous case hold.

SMM with weakly convex surrogates

- SMM with weakly convex surrogates and proximal regularization is developed. In particular, only L-smooth surrogates can be used without verifying convexity.

- Proximal regularization improves the conditioning of surrogate minimization without changing the surrogates.

- Same convergence rate results as in the previous case hold.

Surrogates	Coordinate Decent	Regularization	Asymptotic Convergence	Iteration Complexity	Data Sampling
Strongly convex	Full	None	✓	✓	Markovian
Multi-convex	Block	Diminishing Radius	✓	✓	Markovian
Weakly convex	Full	Proximal	✗	✓	i.i.d.

Table 1: Overview of theoretical results depending on the types of surrogates, coordinate descent, regularization, and data sampling. For the weakly convex case, asymptotic convergence to stationary points is only subsequential. See Theorem 2.

In a unified manner, we provide an extensive convergence analysis on the proposed algorithm in all three cases mentioned above, which we derive under possibly dependent data streams, relaxing the standard i.i.d. assumption on data samples. We obtain global convergence to stationary points of rate $O((\log n)^{1+\varepsilon}/n^{1/2})$, matching the optimal convergence rates for SGD-based methods Sun et al. (2018); Davis and Drusvyatskiy (2019), where $\varepsilon > 0$ is an arbitrary constant. Interestingly, our analysis shows that SRMM (and hence SMM) is more adapted to solve empirical loss minimization than expected loss minimization, in the sense that the aforementioned rate of convergence holds for the empirical loss functions, but for the empirical loss function, a slower rate of $O((\log n)^{1+\varepsilon}/n^{1/4})$ is obtained. This is the opposite of SGD-based methods, which converges almost surely for the expected loss at rate $O((\log n)^2/n^{1/2})$ and rate $O((\log n)^{1+\varepsilon}/n^{1/4})$ for the empirical loss. However, we also show that the optimal almost sure convergence rate $O((\log n)^{1+\varepsilon}/n^{1/2})$ is obtained when

the stationary points of the surrogate functions are in the interior of the constraint set. See Theorems 2, 3, and 4 for the full statements.

Our general framework of SRMM can be specialized to various stochastic optimization algorithms. The examples include

- Double-averaging projected gradient descent Nesterov and Shikhman (2015) and its generalizations (Sec. 5.1.1)
- Proximal point empirial loss minimization Frostig et al. (2015); Lin et al. (2015) (Sec. 5.1.2)
- Online (Nonnegative) Matrix Factorization Mairal et al. (2010); Lyu et al. (2020) (Sec. 5.1.3)
- Subsampled Online (Nonnegative) Matrix Factorization Mensch et al. (2017) (Sec. 5.1.4)
- Online Nonnegative CP-dictionary learning Lyu et al. (2022) (Sec. 5.1.5)

As an immediate corollary, our general results yield the first convergence rate bounds for the above algorithms in the general setting with nonconvex objectives with constraints and possibly non-i.i.d. data samples. For all of the above algorithms, there have not been any convergence rate results for nonconvex objectives even in the special case of i.i.d. data samples. For the first two examples above, even asymptotic convergence to stationary points for nonconvex objectives was not known.

In Section 5.2, we also experimentally validate the efficacy of SRMM for two tasks: Network Dictionary Learning Lyu et al. (2024) and image classification with deep convolutional neural networks for the CIFAR-10 dataset Krizhevsky et al. (2009).

1.2 Related work

In Mairal (2013b), convergence analysis of SMM with i.i.d. data samples are given. There, it is shown that when the objective function f is convex, SMM converges to the global minimum with a rate of convergence in expectation of order $O(\log n/\sqrt{n})$ (see (Mairal, 2013b, Prop 3.1)) and of order $O(1/n)$ when f is strongly convex (see (Mairal, 2013b, Prop 3.2)). Also, when f is nonconvex, SMM is shown to converge almost surely to the set of stationary points of f over a convex constraint set (see (Mairal, 2013b, Prop 3.4)) but no result on the rate of convergence is given. The latter nonconvex result was later extended to SMM with approximate surrogate functions in Mensch et al. (2017), which was applied to developing subsampled online matrix factorization algorithms.

In Lyu et al. (2020), the classical online matrix factorization algorithm in Mairal et al. (2010) is shown to converge when the input data matrices are given by a function of some underlying Markov chain. A similar almost sure global convergence in the Markovian setting for online tensor factorization is obtained in Lyu et al. (2022). In both works, bounds on the rate of convergence are not given. In Lyu and Li (2023), block coordinate descent with diminishing radius is proposed for deterministic nonconvex problems and shown to converge to the stationary points of the objective function. Also, a rate of convergence of order $O(\log n/\sqrt{n})$ is obtained.

Stochastic Gradient Descent (SGD) is another popular method for various optimization problems. In Sun et al. (2018), a convergence of SGD under the Markovian data assumption is obtained. For the convex case, (Sun et al., 2018, Thm. 1) shows the convergence of SGD

to a global minimum with the rate of convergence of order $O((\log n)^{1+\varepsilon}/\sqrt{n})$ for each fixed $\varepsilon > 0$. Moreover, (Sun et al., 2018, Thm. 2) shows that SGD for unconstrained nonconvex optimization converges almost surely to the stationary points with rate $O((\log n)^{1+\varepsilon}/\sqrt{n})$ for each fixed $\varepsilon > 0$. A similar rate of convergence for nonconvex and constrained projected SGD is also known in Davis and Drusvyatskiy (2019). In Zhao et al. (2017), an online NMF algorithm based on projected SGD with general divergence in place of the squared ℓ_2-loss is proposed, and convergence to stationary points to the expected loss function for i.i.d. data samples is shown.

One of the key features of SMM-type algorithms is that the chosen majorizing surrogates in each iteration are recursively averaged. However, MM-type algorithms without such recursive averaging have also been investigated extensively. For instance, for convex objectives, an iteration complexity of $O(\varepsilon^{-2})$ is known for block successive upper-bound minimization (BSUM) algorithm Hong et al. (2017).

1.3 Organization

In Section 2, we state the problem settings of empirical and expected loss minimization and introduce background on majorization-minimization (MM) and stochastic MM (SMM). We also introduce the proposed method of stochastic regularized MM (SRMM) at a high level. In Section 3, we state our SRMM algorithm in Algorithm 1. Next in Section 4, we state our main results (Theorems 2, 3, and 4) and their corollaries (Corollaries 5 and 6). In Section 5, we discuss applications of our general framework on various special instances of SRMM – double averaged PSGD and its generalizations, (subsampled) online matrix factorization, and online CP-dictionary learning. In Section 5.2, we provide numerical experiments of SRMM on network dictionary learning and image classification using deep convolutional neural networks.

The rest of the sections are devoted to convergence analysis for SRMM. In Section 6 we establish some preliminary lemmas. Following in Section 7, we state five key lemmas (Lemmas 20, 21, 22, 23, and 24) and derive all main results stated in Section 4 from them. The first three key lemmas will be proved in Section 8, whereas the other two key lemmas will be established in Sections 9 and 10.

Lastly, we give some backgrounds on Markov chains and Markov chain Monte Carlo (MCMC) sampling in Appendix A, provide examples of various surrogate functions in Appendix B, and give some auxiliary lemmas in Appendix C.

1.4 Notation

In this paper, $\mathbb{R}^p$ denote the ambient space for the parameter space $\boldsymbol{\Theta}$ equipped with standard dot product $\langle \cdot, \cdot \rangle$ and the induced Euclidean norm $\|\cdot\|$. We denote $\mathbf{1}(A)$ the indicator function of event A, which takes value 1 on A and 0 on A^c. For each $J \subseteq \{1, \ldots, p\}$, we denote $\mathbb{R}^J$ by the $|J|$-dimensional subspace of $\mathbb{R}^p$ generated by the coordinates in J. We identify $\mathbb{R}^{\{1,\ldots,p\}}$ and $\mathbb{R}^p$. We call a subset $\mathbb{J} \subseteq 2^{\{1,\ldots,p\}} \setminus \{\emptyset\}$ a set of *coordinate blocks* if $\bigcup \mathbb{J} = \{1, \ldots, p\}$. In this case, each element $J \in \mathbb{J}$ is called a *coordinate block* (or *block coordinate*). Note that two distinct coordinate blocks do not need to be disjoint. For $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$, and a block coordinate $J \subseteq \{1, \ldots, p\}$, denote $\boldsymbol{\Theta}^J := \mathrm{Proj}_{\mathbb{R}^J}(\boldsymbol{\Theta})$ and $\boldsymbol{\theta}^J := \mathrm{Proj}_{\mathbb{R}^J}(\boldsymbol{\theta})$.

2. Problem statement

2.1 Empirical Loss Minimization

Suppose we have a loss function $\ell : \mathfrak{X} \times \boldsymbol{\Theta} \to [0, \infty)$ that measures the fitness of a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ with respect to an observed data $\mathbf{x} \in \mathfrak{X}$. Consider a sequence of newly observed data $(\mathbf{x}_n)_{n \geq 1}$ in $\mathfrak{X}$, where $\mathfrak{X}$ can be a general topological space. We would like to estimate a sequence of parameters $(\boldsymbol{\theta}_n)_{n \geq 1}$ such that $\boldsymbol{\theta}_n$ is in some sense the best fit to all data $\mathbf{x}_1, \cdots, \mathbf{x}_n$ up to time n. In the *empirical loss minimization* framework, one seeks to minimize a recursively defined empirical loss function as new data arrives:

$$\text{Upon arrival of } \mathbf{x}_n: \quad \boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{f}_n(\boldsymbol{\theta}) := (1 - w_n)\bar{f}_{n-1}(\boldsymbol{\theta}) + w_n \, \ell(\mathbf{x}_n, \boldsymbol{\theta}) \right), \tag{2.1}$$

where $(w_n)_{n \geq 1}$ is a sequence of *adaptivity weights* in $(0, 1]$ and the function $\bar{f}_n$ is the *empirical loss function* recursively defined by the weighted average as in (2.1) with $f_0 \equiv 0$. More explicitly, we can write

$$\bar{f}_n(\boldsymbol{\theta}) = \sum_{k=1}^n \ell(\mathbf{x}_k, \boldsymbol{\theta})w_k^n, \qquad w_k^n := w_k \prod_{i=k+1}^n (1 - w_i). \tag{2.2}$$

The adaptivity weight w_n in (2.1) controls how much we want our new estimate $\boldsymbol{\theta}_n$ deviate from minimizing the previous empirical loss $\bar{f}_{n-1}$ to adapting to the newly observed tensor data $\mathbf{x}_n$. In the extreme case of $w_n \equiv 1$, $\boldsymbol{\theta}_n$ is a minimizer of the time-n loss $\ell(\mathbf{x}_n, \cdot)$ and ignores the past $\bar{f}_{n-1}$. If $w_n \equiv \alpha \in (0, 1)$ then the history is forgotten exponentially fast, that is, $\bar{f}_n(\cdot) = \sum_{k=1}^n \alpha(1 - \alpha)^{n-k} \ell(\mathbf{x}_k, \cdot)$. On the other hand, the 'balanced weight' $w_n = 1/n$ makes the empirical loss be the arithmetic mean: $\bar{f}_n(\cdot) = \frac{1}{n} \sum_{k=1}^n \ell(\mathbf{x}_s, \cdot)$, which is the canonical choice in the literature including maximum likelihood estimation and online NMF problem in Mairal et al. (2010). Hence, one may choose the sequence of weights $(w_n)_{n \geq 1}$ in (2.1) to decay fast for learning average features and decay slow (or keep it constant) for learning trending features). See Figure 2 for illustration.

Slow adaptation   Fast adaptation
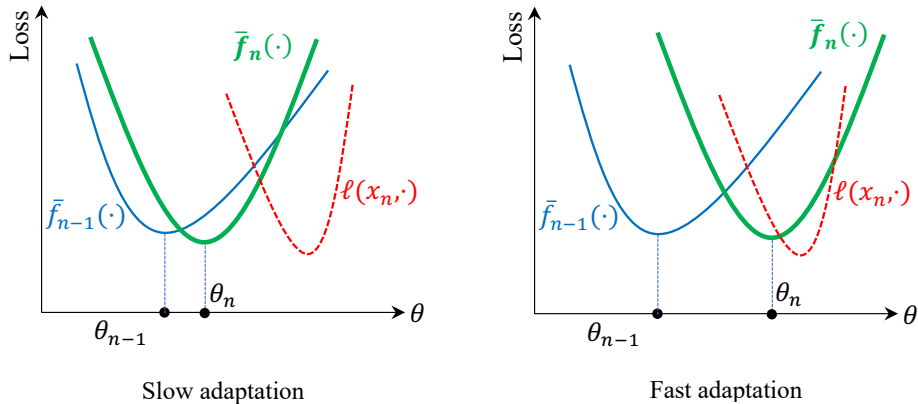
Figure 2: Schematic plot of empirical loss $\bar{f}_n(\cdot) = (1 - w_n)\bar{f}_{n-1}(\cdot) + w_n\ell(\mathbf{x}_n, \cdot)$ when the adaptivity weight w_n is small ('slow adaptation' regime) or large ('fast adaptation' regime).

For instance, consider the linear regression problem where $\ell(\mathbf{x}, \boldsymbol{\theta}) = \|\mathbf{x} - \mathbf{D}\boldsymbol{\theta}\|^2$, where $\mathbf{x} \in \mathbb{R}^q$ and $\mathbf{D} \in \mathbb{R}^{q \times p}$ is a fixed matrix of basis features. Then by writing $\ell(\mathbf{x}, \boldsymbol{\theta}) = $

$\boldsymbol{\theta}^T \mathbf{D}^T \mathbf{D} \boldsymbol{\theta} - 2 \boldsymbol{\theta}^T \mathbf{D}^T \mathbf{x} + \mathbf{x}^T \mathbf{x}$, it is easy to see that

$$\bar{f}_n(\boldsymbol{\theta}) = \|\bar{\mathbf{x}}_n - \mathbf{D} \boldsymbol{\theta}\|^2 + C_n, \tag{2.3}$$

where $\bar{\mathbf{x}}_n := \sum_{k=1}^{n} \mathbf{x}_k w_k^n$ and C_n depends only on $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Hence $\boldsymbol{\theta}_n$ in (2.1) in this case is the vector of best linear regression coefficients that fits the basis $\mathbf{D}$ to the averaged data point $\bar{\mathbf{x}}_n$.

2.2 Expected Loss Minimization

Instead of fitting the model underlying the loss function ℓ to a sequence of data points $(\mathbf{x}_n)_{n \geq 1}$, consider fitting it to a single but random data point $\mathbf{x}$ following some probability distribution π. Then we would be seeking a single estimate $\boldsymbol{\theta}^*$ that minimizes the expected loss function f as below:

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg \min} \left(f(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x} \sim \pi} \left[\ell(\mathbf{x}, \boldsymbol{\theta}) \right] \right). \tag{2.4}$$

We call the above problem setting *expected loss minimization*. This is a popular setting in the optimization literature for stochastic programs. A popular optimization algorithm for solving (2.4) is *projected stochastic gradient descent*, which proceeds by first drawing a sample $\mathbf{x}_n \sim \pi$, estimating the full gradient $\nabla f(\boldsymbol{\theta}_n)$ by the stochastic gradient $\nabla \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1})$, and then updating $\boldsymbol{\theta}_n \leftarrow \text{Proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_{n-1} - \alpha_n \nabla \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}))$, where α_n is a chosen stepsize and $\text{Proj}_{\boldsymbol{\Theta}}$ is the projection operator onto the parameter space $\boldsymbol{\Theta}$.

In the context of linear regression we discussed before, we have

$$f(\boldsymbol{\theta}) = \|\mathbb{E}_{\mathbf{x} \sim \pi}[\mathbf{x}] - \mathbf{D} \boldsymbol{\theta}\|^2 + C_n', \tag{2.5}$$

where C_n' does not depend on $\boldsymbol{\theta}$ and $\mathbf{D}$. By comparing the empirical (2.3) and the expected (2.5) loss functions for linear regression, one can see that the two problem settings are asymptotically equivalent when $\bar{\mathbf{x}}_n \to \mathbb{E}_{\mathbf{x} \sim \pi}[\mathbf{x}]$ as the sample size n for the empirical loss minimization tends to infinity. This is certainly true when $\mathbf{x}_k$'s are i.i.d. from π and $w_k = 1/k$ by the strong law of large numbers. This holds true in a more general setting where $\mathbf{x}_n$ form a Markov chain with stationary distribution π and the weights w_n are non-increasing and decay sufficiently faster than $1/\sqrt{n}$ (see Lemma 42).

2.3 Stochastic Regularized Majorization-Minimization

Majorization-minimization (MM) is a large class of classical approaches for solving nonconvex optimization problems (see Lange et al. (2000) for a recent review) that includes classical gradient descent algorithms as well as expectation-maximization (EM) for solving maximum likelihood estimate (MLE) problems in statistics Cappé and Moulines (2009); Neal and Hinton (1998). A key idea is local convex relaxation. For an illustration, suppose we would like to minimize a differentiable function f with L-Lipschitz gradient (i.e., L-smooth) and the current estimate is $\boldsymbol{\theta}_{n-1}$. In order to compute the new estimate $\boldsymbol{\theta}_n$, instead of directly minimizing f near $\boldsymbol{\theta}_{n-1}$, we may minimize its quadratic expansion

$$g_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}_{n-1}) + \langle \nabla f(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{1}{2\eta} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2,$$

where $\eta > 0$ is a parameter. It is easy to see that g_n is minimized at $\boldsymbol{\theta}_n := \boldsymbol{\theta}_{n-1} - \eta \nabla f(\boldsymbol{\theta}_{n-1})$, yielding the classical gradient descent update with stepsize η. In fact, if one chooses $\eta \leq L$, then g_n majorizes f (i.e., $g_n \geq f$, see Lemma 34) and the resulting gradient descent algorithm converges to the global minimum at rate $O(1/k)$ when f is convex (see, e.g., Bottou (2010)). In general, MM algorithms proceed by iteratively minimizing a majorizing surrogate (such as g_n above) of the objective function f.

Next, consider using a similar idea for solving the ELM problem (2.1). There, instead of a single objective function f, we have a sequence of empirical loss functions $\bar{f}_n$ to be minimized online. It would be natural to first find a majorizing surrogate g_n of the one-point loss function $f_n(\cdot) := \ell(\mathbf{x}_n, \cdot)$ at the current estimate $\boldsymbol{\theta}_n$ and then undergo the same recursive averaging step as in (2.1) to compute an averaged surrogate $\bar{g}_n$ so that $\bar{g}_n$ is a majorizing surrogate of the current objective $\bar{f}_n$. We can then minimize $\bar{g}_n$ to compute the next estimate $\boldsymbol{\theta}_n$. This method is called the *stochastic marjorization-minimization* (SMM) Mairal (2013b), which we state concisely as follow:

(**SMM**)   Upon arrival of $\mathbf{x}_n$: $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (2.6)

$$\begin{cases} g_n \leftarrow \text{Strongly convex majorizing surrogate of } f_n(\cdot) = \ell(\mathbf{x}_n, \cdot) \\ \boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - w_n)\bar{g}_{n-1}(\boldsymbol{\theta}) + w_n\, g_n(\boldsymbol{\theta}) \right). \end{cases}$$

SMM has been successful in a variety of settings including online versions of dictionary learning, matrix factorization, proximal gradient, and DC programming (see Mairal et al. (2010); Mairal (2013a); Mensch et al. (2017)).

A premise of SMM is that $\bar{g}_t$ is strongly convex so that it is easy to find a unique minimizer, which is the case for online matrix factorization problems Mairal et al. (2010); Mairal (2013a). When the surrogate $\bar{g}_n$ is only block multi-convex, that is, $\bar{g}_n : \boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \to [0, \infty)$ and it is convex in each of the m blocks, then finding a global minimum of $\bar{g}_n$ is not easy, if not impossible. Moreover, even when $\bar{g}_n$ is convex, one can still exploit its block multi-convex structure and use a more efficient coordinate descent method to minimize $\bar{g}_n$. After all, it may be enough to minimize $\bar{g}_n$ only approximately at each n using a cheap coordinate descent method since we are interested in solving an online problem.

To this end, we propose the following generalization of SMM, where the surrogate g_n functions can only be block-convex (not necessarily jointly convex) and the averaged surrogate $\bar{g}_n$ can only be approximately minimized (e.g., by using a single round of block coordinate descent) within a trust region. We call our algorithm *Stochastic Block Marjorization-Minimizaiton* (SRMM):

(**SRMM**)   Upon arrival of $\mathbf{x}_n$: $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (2.7)

$$\begin{cases} g_n & \leftarrow \text{(Weakly convex or multi-convex) Majorizing surrogate of } f_n(\cdot) = \ell(\mathbf{x}_n, \cdot) \\ \bar{g}_n(\boldsymbol{\theta}) & \leftarrow (1 - w_n)\bar{g}_{n-1}(\boldsymbol{\theta}) + w_n\, g_n(\boldsymbol{\theta}) \\ \boldsymbol{\theta}_n & \approx \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|) \right) \end{cases}$$

where $\Psi_n(\cdot)$ is a regularizer that penalizes having a large value $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|$ of parameter change.

A motivating example for SRMM is the recent work on online nonnegative CP[1]-dictionary learning in Lyu et al. (2022), which includes online nonnegative tensor CP-decomposition as a special case. There, the variational surrogate is only convex in each of the n loading matrices. The method developed in Lyu et al. (2022) in order to handle a similar issue for a tensor factorization setting is that, at each round, we only approximately minimize the surrogate $\bar{g}_n$ by a single round of cyclic block coordinate descent (BCD) in the n loading matrices. This additional layer of relaxation causes a number of technical difficulties in convergence analysis. One of the key innovations used to ensure convergence of the algorithm is the use of "search radius restriction" developed in Lyu and Li (2023), which can be regarded as a trust region method Yuan (2015); Conn et al. (2000) with diminishing radius. In this paper, we generalize and significantly improve this approach and analysis in Lyu et al. (2022). Most importantly, we will establish a rate of convergence of SRMM in the general nonconvex, constrained, and Markovian data setting. As a corollary, we obtain a rate of convergence results for the classical SMM in the Markovian data case, which has not been available even in the i.i.d. data case.

3. Statement of the algorithm

To state the main algorithm, we first define the class of surrogate functions we use in this paper. Recall the notations in Subsection 1.4.

Definition 1 (First-order ρ-multi-convex Surrogates) Fix a set $\mathbb{J}$ of coordinate blocks and parameters $L > 0$, $\rho \in \mathbb{R}$, and $\varepsilon \geq 0$. A function $g : \mathbb{R}^p \to \mathbb{R}$ is a (first-order) ρ-multi-convex ε-approximate surrogate of $f : \mathbb{R}^p \to \mathbb{R}$ at $\boldsymbol{\theta}^* \in \mathbb{R}^p$ on block coordinates in $\mathbb{J}$ if the following hold:

(i) (ε-majorization) $g(\boldsymbol{\theta}) + \varepsilon \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$;

(ii) (Smoothness of error) The approximation error $h := g - f$ is differentiable and ∇h is L-Lipschitz continuous. Furthermore, $h(\boldsymbol{\theta}^*) \leq \varepsilon$ and $\|\nabla h^*(\boldsymbol{\theta})\| \leq \varepsilon$.

(iii) (Block ρ-multi-convexity) For each $J \in \mathbb{J}$, $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}) - \frac{\rho}{2}\|\boldsymbol{\theta}\|^2$ is convex on $\mathbb{R}^J$. (We allow $\rho < 0$.)

Note that the parameter ρ is not necessarily positive and we allow it to be any value in $\mathbb{R}$. Then **(iii)** states that in each block J, the function g is ρ-*strongly convex* If $\rho > 0$; convex if $\rho = 0$; $(-\rho)$-*weakly convex* if $\rho < 0$. To conveniently refer to all these cases, we call a function g ρ-*convex* for each $\rho \in \mathbb{R}$ if the function $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}) - \frac{\rho}{2}\|\boldsymbol{\theta}\|^2$ is convex. If this holds on $\mathbb{R}^J$ for each coordinate block $J \in \mathbb{J}$, we call g ρ-*multi-convex*. Denote $\mathtt{Srg}_{L,\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^*, \varepsilon)$ for the set of all ρ-multi-convex ε-approximate surrogates of f at $\boldsymbol{\theta}^*$ with parameters L and ρ and coordinate blocks $\mathbb{J}$. If $\mathbb{J}$ consists of the single coordinate block $\{1, \ldots, p\}$, then we write $\mathtt{Srg}_{L,\rho}(f, \boldsymbol{\theta}^*, \varepsilon) := \mathtt{Srg}_{L,\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^*, \varepsilon)$, which consists of ρ-convex surrogates g of f at $\boldsymbol{\theta}^*$. Lastly, we denote $\mathtt{Srg}_{L,\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^*) := \mathtt{Srg}_{L,\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^*, 0)$.

---

1. CANDECOMP/PARAFAC for tensor decomposition

Algorithm 1 Stochastic Regularized Majorization-Minimization (SRMM)

1: **Input:** $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ (initial estimate); N (number of iterations); $(w_n)_{n\geq 1}$, (non-increasing weights in $(0,1]$); $(\varepsilon_n)_{n\geq 1}$, (non-increasing surrogate tolerance in $[0,1]$); $\bar{\mathbb{J}} \in 2^{\{1,\dots,p\}}$ (set of coordinate blocks); $L > 0$ (surrogate smoothness parameter); $(\rho_n)_{n\geq 1}$ (surrogate convexity parameters in $[0,\infty)$);

2:    Initialize the approximate surrogate $\bar{g}_0 : \boldsymbol{\theta} \mapsto \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$; $\bar{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$; $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$;

3:    **for** $n = 1, \cdots, N$ **do**

4:       sample a training point $\mathbf{x}_n \in \mathfrak{X}$; define $f_n : \boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$;

5:       choose a surrogate function $g_n \in \mathtt{Srg}_{L,\rho_n}^{\mathbb{J}}(f_n, \boldsymbol{\theta}_{n-1}, \varepsilon_n)$;

6:       update the average surrogate: $\bar{g}_n = (1 - w_n)\bar{g}_{n-1} + w_n g_n$;

7:       use Algorithm 2 or 3 to compute updated estimate $\boldsymbol{\theta}_n$ near $\boldsymbol{\theta}_{n-1}$:

$$\boldsymbol{\theta}_n \approx \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)\right); \quad \left(\triangleright \begin{array}{c}\text{Inexact surrogate minimization} \\ \text{with regularization}\end{array}\right) \quad (3.1)$$

8:    **end for**

9: **output:** $\boldsymbol{\theta}_N$

Now we state our main algorithm, Algorithm 1, which iteratively executes the following: 1) Sample a new data point $\mathbf{x}_n \in \mathfrak{X}$ using an a priori sampling algorithm (e.g., MCMC); 2) Choose a new surrogate g_n of the loss function $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$; 3) Update and aggregate surrogate $\bar{g}_n$ by taking a weighted average of $\bar{g}_{n-1}$ and g_n; 4) Find an approximate minimizer of $\bar{g}_n$ plus a regularizer $\Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$. In the simplest case when g_n's are strongly convex, we directly minimize the strongly convex function $\bar{g}_n$ over $\boldsymbol{\Theta}$; When g_n's are ρ-weakly convex, then we minimize the strongly convex function $\bar{g}_n(\boldsymbol{\theta}) + \frac{\lambda_n}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2$ over $\boldsymbol{\Theta}$, where $\lambda_n > \rho$. These two cases are covered by Algorithm 2.

Algorithm 2 Surrogate Minimization with Proximal Regularization

---

1: **Input:** all input of Algorithm 1; $\boldsymbol{\theta}_{n-1} \in \boldsymbol{\Theta}$ (current estimate); $m \in \mathbb{N}$ (number of sub-iterations); $\hat{\rho} > -\rho$ (proximal regularization parameter);

2: **Require:** $\rho \leq 0$ (i.e., $\boldsymbol{\theta} \mapsto g_n(\boldsymbol{\theta}) + \frac{|\rho|}{2}\|\boldsymbol{\theta}\|^2$ is convex for each $n \geq 1$)

3: **Do:**

$$\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\bar{g}_n(\boldsymbol{\theta}) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2\right]; \quad (3.2)$$

4: **output:** $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$

Algorithm 3 Block Surrogate Minimization with Diminishing Radius

---

1: **Input:** all input of Algorithm 1; $\boldsymbol{\theta}_{n-1} \in \boldsymbol{\Theta}$ (current estimate); $m \in \mathbb{N}$ (number of sub-iterations);

2: **Require:** g_n is block multi-convex with block structure $\mathbb{J}$ for each $n \geq 1$;

3: $\boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1}$;

4: **for** $i = 1, \cdots, m$ **do**:

5: choose a block coordinate $J = J_i(n) \in \mathbb{J}$ independently of everything else;

6: freeze coordinates in J^c:

$$\boldsymbol{\Theta}_n^J \leftarrow \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \,\middle|\, \boldsymbol{\theta}^{J^c} = (\boldsymbol{\theta}_n)^{J^c} \right\} \tag{3.3}$$

7: update $\boldsymbol{\theta}_n$ on coordinates in J while holding the rest of coordinates in J^c:

$$\boldsymbol{\theta}_n^{(i)} \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_n^J} \left[\bar{g}_n(\boldsymbol{\theta}) + (\infty \cdot \mathbf{1}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| > w_n/m)) \right]; \quad (\triangleright \operatorname{Set} \infty \cdot 0 = 0) \tag{3.4}$$

8: **end for**

9: **output**: $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n^{(m)} \in \boldsymbol{\Theta}$

On the other hand, when the surrogates g_n are block multi-convex with respect to the coordinate blocks in $\mathbb{J}$, then so is their weight average $\bar{g}_n$. Unless $\bar{g}_n$ itself is convex, finding an exact minimizer of $\bar{g}_n$ over $\boldsymbol{\Theta}$ in step (3.1) is infeasible. In fact, computing an exact minimizer of $\bar{g}_n$ in every step may not be necessary for the convergence of the algorithm, as it was observed for the problem of online CP-dictionary learning Lyu et al. (2022). Instead, by exploiting the block multi-convex structure of $\bar{g}_n$, we may solve a fixed number of convex sub-problems over deterministically or randomly chosen blocks in $\mathbb{J}$.

In each step of solving (3.1), it is crucial to ensure that the new estimate $\boldsymbol{\theta}_n$ obtained by approximately minimizing $\bar{g}_n$ is not too far from the previous estimate $\boldsymbol{\theta}_{n-1}$. When $\bar{g}_n$ is strongly convex on full coordinates and if $\boldsymbol{\theta}_n$ is an exact minimizer of $\bar{g}_n$, a simple argument shows that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ (see (Mairal, 2013b, Lem. B.8)). In this case, we may directly find an exact minimizer $\boldsymbol{\theta}_n$ of $\bar{g}_n$ over $\boldsymbol{\Theta}$ as in Algorithm 2 with $\hat{\rho} = 0$. This specialization corresponds to the original SMM algorithm in Mairal (2013b). However, such property is not a priori satisfied in the general case when $\bar{g}_n$ is nonconvex or $\boldsymbol{\theta}_n$ is an inexact minimizer of $\bar{g}_n$ over $\boldsymbol{\Theta}$.

A key idea behind Algorithms 2 and 3 for averaged surrogate minimization is to use an additional regularization that penalizes large values of $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|$ in (3.1). We consider two such regularization schemes stated in Algorithm 3: A 'hard' regularization of *Diminishing Radius* (DR) and a 'soft' regularization of *Proximal Regularization* (PR). DR can also be viewed as a trust region method, where our trust-region takes the form of the Euclidean ball of diminishing radius $r_n = O(w_n)$ in the order of adaptivity weights w_n used for iterated averaging of the objective functions. PR is a standard regularization scheme that quadratically penalizes the distance from the old estimate Parikh and Boyd (2014). The first-order optimality conditions for these two regularization schemes are equivalent assuming the solution of the latter lies in the interior of the trust region, but not necessarily in general. Another difference between the two regularization methods is that DR does not change the objective gradient but PR does. Nonetheless, in the context of block coordinate descent (BCD), both

regularization methods guarantee convergence to stationary points Grippo and Sciandrone (2000); Xu and Yin (2013); Lyu and Li (2023).

Note that each of the block minimization problems in (3.2) and (3.4) is a constrained convex optimization problem so it can be easily solved by a number of known algorithms (e.g., projected gradient descent Beck (2017), LARS Efron et al. (2004), LASSO Tibshirani (1996), and feature-sign search Lee et al. (2007)). Indeed, for (3.2), the averaged surrogate $\bar{g}_n$ is $-\rho$-weakly convex (recall $\rho < 0$) and so its proximal point modification with $\hat{\rho} > -\rho$ is $(\hat{\rho} + \rho)$-strongly convex; Also, (3.4) is equivalent to minimizing $\bar{g}_n$ over the convex set $\Theta_n \cap \{\boldsymbol{\theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq w_n/m\}$ restricted on $\mathbb{R}^J$, where the retriction of $\bar{g}_n$ on R^J is convex. In particular, using standard projected gradient descent algorithms, one can decrease the optimality gap sub-linearly for convex sub-problems and linearly if the restricted objectives are strongly convex (see, e.g., (Beck, 2017, Thm. 10.29)).

4. Main results

4.1 Optimality conditions and convergence measures

In this subsection, we introduce some notions on optimality conditions and related quantities. Here we denote f to be a general objective function $\Theta \to \mathbb{R}$, but elsewhere f will denote the expected loss function in (2.4) unless otherwise mentioned.

Recall that we say $\boldsymbol{\theta}^* \in \Theta$ is a *stationary point* of f over Θ if

$$\sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0, \tag{4.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot project on $\mathbb{R}^p \supseteq \Theta$. (The restriction $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1$ in (4.1) is vacuous by rescaling $\boldsymbol{\theta} - \boldsymbol{\theta}^*$.) Equivalently, let $\mathcal{N}_\Theta(\boldsymbol{\theta}) := \{\eta \,:\, \langle \eta, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle \leq 0 \;\forall \boldsymbol{\theta}' \in \Theta\}$ denote the normal cone of Θ at $\boldsymbol{\theta}$. Then $\boldsymbol{\theta}^*$ is a stationary point of f over Θ if and only if $-\nabla f(\boldsymbol{\theta}^*) \in \mathcal{N}_\Theta(\boldsymbol{\theta})$. If $\boldsymbol{\theta}^*$ is in the interior of Θ, then it implies $\|\nabla f(\boldsymbol{\theta}^*)\| = 0$. For iterative algorithms, such a first-order optimality condition may hardly be satisfied exactly in a finite number of iterations, so it is important to know how the worst-case number of iterations required to achieve an ε-approximate solution scales with the desired precision ε. More precisely, we say $\boldsymbol{\theta}^* \in \Theta$ is an *ε-approximate stationary point* of f over Θ if

$$\sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq \varepsilon. \tag{4.2}$$

This notion of ε-approximate solution is consistent with the corresponding notion for unconstrained problems. Indeed, if $\boldsymbol{\theta}^*$ lies sufficiently deep the interior of Θ, then

$$\sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle = \|f(\boldsymbol{\theta}^*)\|. \tag{4.3}$$

There are alternative ways to define ε-stationary points for constrained problems. We have the equivalence between the following two stationary measures (see (Alacaoglu and Lyu, 2023, Prop. B.1) and (Rockafellar and Wets, 2009, Prop. 8.32)):

$$\sup_{\boldsymbol{\theta} \in \Theta \setminus \{\boldsymbol{\theta}^*\}} \left\langle -\nabla f(\boldsymbol{\theta}^*), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle = \text{dist}(\mathbf{0}, \nabla f(\boldsymbol{\theta}) + \mathcal{N}_\Theta(\boldsymbol{\theta})), \tag{4.4}$$

where the right-hand side is a standard measure of stationarity (e.g., see Nesterov (2013)). The stationary measures in (4.4) do not behave continuously as one approaches stationary points at the boundary of the constraint set from the interior. Indeed, when $\boldsymbol{\theta}^*$ is in the interior of $\boldsymbol{\Theta}$, then the quantities in (4.4) always equal to $\|\nabla f(\boldsymbol{\theta}^*)\|$, whereas the left-hand side of (4.3) can be much smaller than $\|\nabla f(\boldsymbol{\theta}^*)\|$ when $\boldsymbol{\theta}^*$ is close to the boundary of $\boldsymbol{\Theta}$. This was partly the motivation for Davis and Drusvyatskiy to use a 'near-stationary measure' incorporating Moreu envelop in their seminal work for constrained stochastic optimization Davis and Drusvyatskiy (2019). Namely, if the gradient norm of the Moreau envelope is small at a parameter $\boldsymbol{\theta}$, then the proximal point $\hat{\boldsymbol{\theta}}$ is close to being stationary and is near $\boldsymbol{\theta}$. In comparison, the stationarity measure we introduced in (4.2) is a 'direct' measure of approximate stationarity (since it does not involve an additional proximal operator) and it behaves continuously near the boundary.

Next, for each $\varepsilon > 0$ we define the *iteration complexity* N_ε of an algorithm for minimizing f over $\boldsymbol{\Theta}$ with initialization $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ as

$$N_\varepsilon = N_\varepsilon(f, \boldsymbol{\theta}_0) := \inf \{n \geq 1 \,|\, \boldsymbol{\theta}_n \text{ is an } \varepsilon\text{-approximate stationary point of } f \text{ over } \boldsymbol{\Theta}\},$$

where $(\boldsymbol{\theta}_n)_{n \geq 0}$ is a sequence of estimates produced by the algorithm with an initial estimate $\boldsymbol{\theta}_0$. An upper bound on N_ε can be regarded as the *worst-case* bound on the number of iterations for an algorithm to achieve an ε-approximate solution.

4.2 Assumptions

In this subsection, we state all assumptions we use for establishing the main results. Throughout this paper, we denote by $\mathcal{F}_n$ the σ-algebra generated by the data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as well as the choice of coordinate blocks $J_1(k), \ldots, J_m(k)$ for $1 \leq k \leq m$. Clearly $(\mathcal{F}_n)_{n \geq 1}$ defines a filtration, that is, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$.

(A1) *(Per-sample loss function) There exists constants $R, L > 0$ such that for each data point $\mathbf{x} \in \mathfrak{X}$, the function $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is R-Lipscthiz continuous and its gradient is L-Lipschitz continuous.*

(A2) *(Data sampling) The observed data points $\mathbf{x}_n \in \mathfrak{X}$ are given by $\mathbf{x}_n = \varphi(Y_n)$, where Y_n is Markov chain on state space Ω and $\varphi : \Omega \to \mathbb{R}^p$ is a function with compact image. Furthermore, Y_n has a unique stationary distribution π and satisfies exponential mixing with some rate $\lambda \in [0, 1)$:*

$$\sup_{\mathbf{y} \in \Omega} \|P^m(\mathbf{y}, \cdot) - \pi\|_{TV} \leq \lambda^m. \tag{4.5}$$

(A3) *(Constraint sets for parameters) The constraint set $\boldsymbol{\Theta}$ is a compact and convex subset of $\mathbb{R}^p$. Let $\mathbb{J} \subseteq 2^{\{1,\ldots,p\}} \setminus \{\emptyset\}$ denote the set of coordinate blocks used in Algorithm 1. Then for each block coodinate $J \in \mathbb{J}$, $\boldsymbol{\Theta}^J$ contains an open ball in $\mathbb{R}^J$.*

(A4) *(Adaptivity weight decay) The weights $w_n \in (0, 1]$ are non-increasing and satisfy $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all sufficiently large $n \geq 1$, $\sum_{n=1}^\infty w_n = \infty$, and $w_n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$*

for some $\varepsilon > 0$. Also, there exists a constant $C > 0$ such that for $a_n = \lfloor C \log n \rfloor$, the following hold:

$$\sum_{n=1}^{\infty} w_n^2 \sum_{k=1}^{n+1} \mathbb{E}[\varepsilon_k] < \infty, \qquad \sum_{n=1}^{\infty} a_n w_n w_{n-a_n} < \infty, \qquad \sum_{n=1}^{\infty} w_n \lambda^{a_n} < \infty,$$

$$\left(\text{optionally,} \ \sum_{n=1}^{\infty} w_n w_{n-a_n} \sqrt{n} < \infty \right),$$

where $(\varepsilon_n)_{n \geq 1}$ is given in Algorithm 1 and $\lambda \in (0, 1]$ is given in (A2).

(A5) *(Surrogate minimization) If Algorithm 2 is used for (3.1), define*

$$\Delta_n = \Delta_n^{(1)} := \bar{G}_n(\boldsymbol{\theta}_n^{(i)}) - \bar{G}_n(\boldsymbol{\theta}_n^{(i\star)}), \quad \boldsymbol{\theta}_n^{(i\star)} := \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \bar{G}_n(\boldsymbol{\theta}),$$

$$\bar{G}_n(\boldsymbol{\theta}) := \bar{g}_n(\boldsymbol{\theta}) + \frac{\hat{\rho}}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2.$$

If Algorithm 3 is used for (3.1), define

$$\Delta_n := \sum_{i=1}^{m} \Delta_n^{(i)}, \qquad \Delta_n^{(i)} := \bar{g}_n(\boldsymbol{\theta}_n^{(i)}) - \bar{g}_n(\boldsymbol{\theta}_n^{(i\star)}), \qquad \boldsymbol{\theta}_n^{(i\star)} := \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_n^{J_i}} \bar{g}_n(\boldsymbol{\theta}),$$

where $\boldsymbol{\Theta}_n^{J_i}$ is defined in (3.3). Then in both cases, $\mathbb{E}\left[\sum_{n=1}^{\infty} \Delta_n\right] < \infty$.

(A6) *(Block coordinate sampling) If Algorithm 3 is used for (3.1), then the joint distribution of coordinate blocks $(J_1(n), \ldots, J_m(n))$ chosen in Algorithm 3 at iteration n of Algorithm 1 does not depend on n. Furthermore, the coordinate blocks are disjoint, and the expected number of each coordinate in $\{1, \ldots, p\}$ appearing in all coordinate blocks $J_1(n), \ldots, J_m(n)$ is constant.*

(A7) *(Parameterized surrogates) The averaged surrogates $\bar{g}_n$ are parameterized by some variable κ_n in some compact set $\mathcal{K}$. That is, there exists a function $\bar{g} : \mathcal{K} \times \boldsymbol{\Theta} \to [0, \infty)$ such that $\bar{g}_n(\boldsymbol{\theta}) = \bar{g}(\kappa_n, \boldsymbol{\theta}, \varepsilon_n)$ for some $\kappa_n \in \mathcal{K}$. Furthermore, $\bar{g}$ is Lipschitz in the first coordinate.*

Assumptions (A1) and (A3) are standard in the literature of constrained stochastic nonconvex optimization and online dictionary learning Mairal et al. (2010); Mairal (2013b); Mensch et al. (2017); Lyu et al. (2020, 2022).

(A2) states a relaxation of the standard i.i.d. data sampling assumption in Mairal et al. (2010); Mairal (2013b); Mensch et al. (2017), where we assume the data samples are obtained by some underlying Markov chain with exponential mixing to the stationary distribution π (geometric ergodicity). It states that the m-step empirical distribution converges to the stationary distribution π at a geometric rate λ, which holds trivially (with $\lambda = 0$) when the underlying Markov chain Y_n is in fact drawn i.i.d. from π. When the Markov chain is irreducible and aperiodic on a finite state space, then (4.5) holds with some constant $\lambda > 0$ Levin and Peres (2017). In the case when the underlying Markov chain Y_n has countably

infinite or uncountable state space, there are a number of sufficient conditions for geometric ergodicity (see, e.g., (Meyn and Tweedie, 2012, Sec. 15)). See Lyu et al. (2020) and Sun et al. (2018) for concrete applications and sampling methods that satisfy this assumption. This assumption is common in the literature Bhandari et al. (2018); Sun et al. (2018); **?**); Lyu et al. (2020, 2022); Nagaraj et al. (2020); Alacaoglu and Lyu (2023) and i.i.d. sampling is another special case.

Assumption (A4) states that the sequence of weights $w_n \in (0,1]$ we use to recursively define the empirical loss (2.1) and surrogate loss (2.7) does not decay too fast so that $\sum_{n=1}^{\infty} w_n = \infty$ but decay fast enough so that $\sum_{n=1}^{\infty} w_n^2 < \infty$. This is analogous to requirements for stepsizes in stochastic gradient descent algorithms, where the stepsizes are usually required to be non-summable but square-summable (see, e.g., Sun et al. (2018)). Note that our general results do not require the stronger assumption $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$, which is standard in the literature Mairal et al. (2010); Mairal (2013b); Mensch et al. (2017); Lyu et al. (2020, 2022). Also, the condition $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all suficiently large n is equivalent for the recursively defined weights w_k^n in (2.2) being non-decreasing in k for all sufficiently large k, which is required to use Lemma 42. We also remark that (A4) is implied by the following simpler condition:

(A4') *The sequence of non-increasing weights $w_n \in (0,1]$ in Algorithm 1 satisfy either $w_n = n^{-1}$ for $n \geq 1$ or $w_n = \Theta(n^{-\beta}(\log n)^{-\delta})$ for some $\delta > 1$ and $\beta \in [1/2, 1)$ (Optionally, $\beta \in [3/4, 1)$). Also, the sequence $(\varepsilon_n)_{n \geq 1}$ in Algorithm 1 satisfies $\sum_{n=1}^{\infty} w_{n+1}^2 \sum_{k=1}^{n+1} \mathbb{E}[\varepsilon_{k+1}] < \infty$.*

If we consider the weight $w_n = n^{-\beta}/(\log n)^\delta$ for some $\beta \in [1/2, 1)$ and $\delta > 1$ as in (A4'), we have

$$\left(\sum_{k=1}^{n} w_n \right)^{-1} = O(n^{\beta-1}(\log n)^\delta). \tag{4.6}$$

Hence the above bound is optimized when $\beta = 1/2$ for each fixed $\delta > 1$.

Next, we give some remarks on (A5). When the averaged surrogate $\bar{g}_n$ is strongly convex on each coordinate block $J_i(n)$, then the expected optimality gap in (A5) decays exponentially fast in the number of iterations of standard constrained convex optimization algorithms such as projected gradient descent (see, e.g., (Beck, 2017, Thm. 10.29)). See also stochastic gradient descent or random coordinate descent (see, e.g., (Bottou et al., 2018, Thm 4.6) and (Wright, 2015, Thm. 1)) for unconstrained cases. Hence, for instance, one can ensure $\mathbb{E}[\Delta_n] = O(n^{-2})$ in $O(m \log n)$ sub-iterations for computing $\boldsymbol{\theta}_n^{(1)}, \ldots, \boldsymbol{\theta}_n^{(m)}$ in Algorithm 3. Consequently, the total computational cost of Algorithm 1 would be the iteration complexity times a log factor, which is negligible.

On the other hand, when Algorithm 2 is used for (3.1) in Theorem 2, we will always be minimizing a strongly convex function over $\boldsymbol{\Theta}$ to find $\boldsymbol{\theta}_n$, so the same remark applies. In a special case when the surrogates g_n are strongly convex and Algorithm 2 is used with $\hat{\rho} = 0$, then one can ensure $\mathbb{E}[\Delta_n] = O(w_n^2)$ in $O(1)$ sub-iterations, instead of $O(\log n)$ (see Lemma 39). More precisely, (A5) is implied by the following simple condition, which was also used in (Mensch et al., 2017, Assumption (**I**)) to analyze SMM with inexact surrogate minimization:

(A5') *(Surrogate optimality gap decay) Suppose the surrogates g_n are ρ-strongly convex for some $\rho > 0$ Algorithm 2 is used with $\hat{\rho} = 0$. Then there exists a constant $\mu \in (0, 1]$ such that for all $n \geq 1$,*

$$\mathbb{E}\left[\bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n^\star) \,\middle|\, \mathcal{F}_{n-1}\right] \leq (1 - \mu)\left[\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n^\star)\right],$$

where $\boldsymbol{\theta}_n^{(\star)}$ denotes the exact minimizer of $\bar{g}_n$ over $\boldsymbol{\Theta}$.

(A6) asserts some properties of a random sampling of block coordinates $J_1(n), \ldots, J_m(n)$ in Algorithm 3, which are crucially used in the proof of Lemma 28 that is pivotal to establishing a rate of convergence results in Theorems 3 and 4. We note that (A6) is trivially satisfied if the deterministic cyclic block coordinate descent is used. Namely, if each surrogate g_n is block multi-convex with block structure $\mathbb{J} = \{J_1, \ldots, J_m\}$ where $J_1 \cup \cdots \cup J_m$ gives a partition of full coordinates $\{1, \ldots, p\}$, then we can deterministically cycle through the m coordinate blocks by setting $J_i(n) = J_i$ for $1 \leq i \leq m$. Then (A6) is satisfied. Such cyclic block coordinate descent was used in the online CP-dictionary learning Lyu et al. (2022). Hence the present work generalizes such a deterministic block coordinate schedule to a possibly randomized schedule.

Lastly, (A7) asserts that the averaged surrogates $\bar{g}_n$ can be parameterized by a compact index set $\mathcal{K}$, which is satisfied by most practical use cases of SMM-type algorithms Mairal (2013b) including online matrix factorization (see Subsection 5.1.3) and online CP-dictionary learning (see Subsection 5.1.5 as well as (Lyu et al., 2022, Alg. 2)). While (A7) is crucially used in deriving some of the main results (Theorem 2), it is also of practical importance since it allows one to store averaged surrogates $\bar{g}_n$ only by storing some sufficient statistics living in a compact set $\mathcal{K}$, without needing to store all past data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This is the case for the double-averaging PSGD (Sec. 5.1.1) and online matrix/tensor factorization (Sec. 5.1.3 and 5.1.5).

4.3 Statement of main results

Throughout this section, let $(\boldsymbol{\theta}_n)_{n\geq 1}$ denote the output of Algorithm 1 with arbitrary initialization $\boldsymbol{\theta}_0$. We will consider one of the following instances:

C1 *(Strongly convex surrogates without regularization) $g_n \in \mathtt{Srg}_{L,\rho}(f_n, \boldsymbol{\theta}_{n-1}, \varepsilon_n)$ for $n \geq 1$ for some $L, \rho > 0$. Use Algorithm 2 with $\hat{\rho} = 0$ for solving (3.1). In this case, assume (A5').*

C2 *(Multi-convex surrogates with radius restriction) $g_n \in \mathtt{Srg}^{\mathbb{J}}_{L,0}(f_n, \boldsymbol{\theta}_{n-1}, \varepsilon_n)$ for $n \geq 1$ for some set $\mathbb{J}$ of coordinate blocks and for some $L > 0$. Use Algorithm 3 for solving (3.1). Assume (A5)-(A6) and $r_{n+1}/r_n = O(1)$.*

C3 *(Weakly convex surrogates with proximal regularization) $g_n \in \mathtt{Srg}_{L,\rho}(f_n, \boldsymbol{\theta}_{n-1}, \varepsilon_n)$ for $n \geq 1$ for some $L > 0$ and $\rho \leq 0$. Use Algorithm 2 with $\hat{\rho} > -\rho$ for solving (3.1). Assume (A5) and that the data sequence $(\mathbf{x}_n)_{n\geq 1}$ is i.i.d. from the stationary distribution π.*

Now we state our first main result, Theorem 2, which states that Algorithm 1 converges globally (w.r.t. initialization) to the set of stationary points of both the empirical loss $\bar{f}_n$ and

the expected loss f. Moreover, it also states that the averaged surrogate $\bar{g}_n$ is asymptotically an accurate approximation of the empirical and the expected loss functions at $\boldsymbol{\theta}_n$ both in the function values and gradients.

Theorem 2 (Global Convergence) *Assume (A1)-(A4) and (A7) hold. Then for cases C1-C2, the following hold:*

(i) *(Empirical Loss Minimization) Both $|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)|$ and $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|$ converge to zero almost surely. Furthermore, $\boldsymbol{\theta}_n$ is asymptotically a stationary point of $\bar{f}_n$ over $\boldsymbol{\Theta}$ almost surely.*

(ii) *(Expected Loss Minimization) $|\bar{g}_n(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_n)|$, $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|$ and $\|\mathbb{E}[\nabla \bar{g}_n(\boldsymbol{\theta}_n)] - \nabla f(\boldsymbol{\theta}_n)\|$ converge to zero alsmot surely. Furthermore, $\boldsymbol{\theta}_n$ converges to the set of a stationary points of f over $\boldsymbol{\Theta}$ almost surely.*

For the case C3, the following subsequential versions of **(i)**-**(ii)** *hold:*

(iii) *All five quantities in* **(i)**-**(ii)** *converge to zero almost surely on some subsequence of $(\boldsymbol{\theta}_n)_{n\geq 1}$. Furthermore, $\boldsymbol{\theta}_n$ is asymptotically a stationary point of $\bar{g}_n$ over $\boldsymbol{\Theta}$ almost surely.*

We note that standard convergence results in the literature of SMM Mairal et al. (2010); Mairal (2013a); Lyu et al. (2020) asserts that the SMM algorithm converges globally to the set of the stationary points of the expected loss function f almost surely, which is recovered by Theorem 2 **(ii)** for the first case of strongly convex surrogates without radius restriction. Note that we establish Theorem 2 without the optional condition in (A4), which essentially states that $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$. Such condition was standard in the literature (see, e.g., (Mairal, 2013b, Assumption (E))). The same statement for the second case of block multi-convex surrogates with radius restriction was recently obtained in Lyu et al. (2022) for the context of online CP-dictionary learning.

The convergence of gradient norms $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|$ and $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|$ are new and the asymptotic stationarity for the empirical loss functions in Theorem 2 **(i)** has not been elaborated very much in the aforementioned literature, since the main focus of using SMM was to solve the expected loss minimization. It is worth noting that the hypothesis for the expected loss minimization stated in Theorem 2 **(ii)** is a bit stronger than that for the empirical loss minimization stated in Theorem 2 **(i)**. This is an indication that SRMM (and hence SMM) is generically more suited to solve the empirical loss minimization than the expected loss minimization, which we will elaborate on in the forthcoming results.

In the following two results, Theorems 3 and 4, we establish bounds on the rate of convergence of surrogate gaps as well as approximate optimality. Below, we denote by $\mathbb{E}[\cdot]$ the coditional expectation $\mathbb{E}[\cdot \,|\, \mathcal{F}_0]$ with respect to the time-0 information that contains the initial estimate $\boldsymbol{\theta}_0$.

Theorem 3 (Rate of Convergence of Surrogate Gaps and Variation) *Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1. Make the same assumption as in Theorem 2. Then the following hold:*

(i) *(Empirical Loss Minimization) Asymptotically almost surely,*

$$\min_{1\leq k\leq n} \left(\left|\bar{g}_k(\boldsymbol{\theta}_k) - \bar{f}_k(\boldsymbol{\theta}_k)\right| + \|\nabla\bar{g}_k(\boldsymbol{\theta}_k) - \nabla\bar{f}_n(\boldsymbol{\theta}_k)\|^2 \right) = O\left(\left(\sum_{k=1}^n w_k\right)^{-1} \right). \quad (4.7)$$

Furthermore, if the optional condition in (A4) holds, then asymptotically almost surely,

$$\min_{1\leq k\leq n} \left(\left|\mathbb{E}[\bar{g}_k(\boldsymbol{\theta}_k)] - \bar{f}_k(\boldsymbol{\theta}_k)\right| + \|\mathbb{E}[\nabla\bar{g}_k(\boldsymbol{\theta}_k)] - \nabla\bar{f}_k(\boldsymbol{\theta}_k)\|^2 \right) = O\left(\left(\sum_{k=1}^n w_k\right)^{-1} \right). \quad (4.8)$$

(ii) *(Expected Loss Minimization) Asymptotically almost surely,*

$$\min_{1\leq k\leq n} \left(\left|\mathbb{E}[\bar{g}_k(\boldsymbol{\theta}_k)] - f(\boldsymbol{\theta}_k)\right| + \|\mathbb{E}[\nabla\bar{g}_k(\boldsymbol{\theta}_k)] - \nabla f(\boldsymbol{\theta}_k)\|^2 \right) = O\left(\left(\sum_{k=1}^n w_k\right)^{-1} \right). \quad (4.9)$$

Furthermore, if the optional condition in (A4) holds, then asymptotically almost surely,

$$\min_{1\leq k\leq n} \left(\left|\bar{g}_k(\boldsymbol{\theta}_k) - f(\boldsymbol{\theta}_k)\right| + \|\nabla\bar{g}_k(\boldsymbol{\theta}_k) - \nabla f(\boldsymbol{\theta}_k)\|^2 \right) = O\left(\left(\sum_{k=1}^n w_k\right)^{-1} \right). \quad (4.10)$$

Theorem 4 (Rate of Convergence to Stationarity) *Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1. Make the same assumption as in Theorem 2. Then the following hold:*

(i) *(Surrogate and Empirical Loss Stationarity) Asymptotically almost surely,*

$$\min_{1\leq k\leq n} \left[\sup_{\boldsymbol{\theta}\in\Theta, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla\bar{g}_k(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k\rangle \right] = O\left(\left(\sum_{k=1}^n w_k\right)^{-1} \right), \quad (4.11)$$

$$\min_{1\leq k\leq n} \left[\sup_{\boldsymbol{\theta}\in\Theta, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla\bar{f}_k(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k\rangle \right] = O\left(\left(\sum_{k=1}^n w_k\right)^{-1/2} \right). \quad (4.12)$$

(ii) *(Expected Loss Stationarity) It holds that*

$$\min_{1\leq k\leq n} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k\rangle \right] = O\left(\left(\sum_{k=1}^n w_k\right)^{-1/2} + w_n\sqrt{n} \right). \quad (4.13)$$

Further assume the optional condition in (A4) holds. Then asymptotically almost surely,

$$\min_{1\leq k\leq n} \left[\sup_{\boldsymbol{\theta}\in\Theta, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k\rangle \right] = O\left(\left(\sum_{k=1}^n w_k\right)^{-1/2} \right). \quad (4.14)$$

To our best knowledge, the rate of convergence results in Theorems 3 and 4 (for nonconvex objectives) are entirely new for SMM-type algorithms even under the classical setting with i.i.d. data and strongly convex surrogates. Only almost sure convergence to stationary points was known before Mairal et al. (2010); Mairal (2013b); Zhao et al. (2017); Lyu et al. (2020, 2022) in such cases. Moreover, (4.8) and (4.9) in Theorem 3 give bounds on the variation of the random objective values $\bar{f}_n(\boldsymbol{\theta}_n)$ and $f(\boldsymbol{\theta}_n)$ against the deterministic quantity $\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)]$ with respect to the randomness of data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and of Algorithm 1 (e.g., possibly randomized choice of blocks in Algorithm 3). We remark that (A7) in fact is not necessary for deriving Theorems 3 and 4.

Next, we state a corollary of Theorems 3 and 4, which specializes these results to a more familiar setting of unconstrained nonconvex optimization.

Corollary 5 *Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1. Make the same assumption as in Theorem 2. Further, assume that for some $\varepsilon > 0$, open ε-ball centered at $\boldsymbol{\theta}_n$ lies in the interior of Θ for all $n \geq 1$. Then asymptotically almost surely,*

$$\min_{1\leq k\leq n} \|\nabla \bar{g}_k(\boldsymbol{\theta}_k)\|^2 = O\left(\left(\sum_{k=1}^n w_k\right)^{-2}\right), \quad \min_{1\leq k\leq n} \|\nabla \bar{f}_k(\boldsymbol{\theta}_k)\|^2 = O\left(\left(\sum_{k=1}^n w_k\right)^{-1}\right), \quad (4.15)$$

$$\min_{1\leq k\leq n} \|\nabla f(\boldsymbol{\theta}_k)\|^2 = O\left(\left(\sum_{k=1}^n w_k\right)^{-1}\right).$$

Below we unpack the above results and give some remarks. For a more direct interpretation of our results, we will consider the weight $w_k = k^{-\beta}/(\log k)^{1+\varepsilon}$ for some $\beta \in [1/2, 1)$ and $\varepsilon > 0$ as in (A4'). Then the bound in (4.6) is optimized at $\beta = 1/2$ for each fixed $\varepsilon > 0$ to

$$\left(\sum_{k=1}^n w_k\right)^{-1} = O(n^{-1/2}(\log n)^{1+\varepsilon}). \quad (4.16)$$

Suppose all iterate $\boldsymbol{\theta}_n$ are in the interior of Θ. This would be a reasonable assumption when all stationary points of the averaged surrogates $\bar{g}_n$ are in the interior of Θ. Then by Corollary 5,

$$\min_{1\leq k\leq n} \|\nabla \bar{g}_k(\boldsymbol{\theta}_k)\| = O\left(\frac{(\log n)^{1+\varepsilon}}{n^{1/2}}\right), \quad \min_{1\leq k\leq n} \|\nabla \bar{f}_k(\boldsymbol{\theta}_k)\| = O\left(\frac{(\log n)^{(1+\varepsilon)/2}}{n^{1/4}}\right), \quad (4.17)$$

$$\min_{1\leq k\leq n} \|\nabla f(\boldsymbol{\theta}_k)\| = O\left(\frac{(\log n)^{(1+\varepsilon)/2}}{n^{1/4}}\right).$$

The last asymptotic expresses a well-known rate of convergence bound for nonconvex and unconstrained SGD Sun et al. (2018); Ward et al. (2019). A similar rate of convergence for nonconvex and constrained projected SGD is also known in Davis and Drusvyatskiy (2019), although a different measure (using Moreu envelope) of the rate of convergence for the constrained nonconvex problems was used. We also remark that, in the finite-horizon

setting where one knows the total number n of iterations ahead of the time, one can choose $w_k = \Theta(n^{-1})$ for $1 \le k \le n$ and eliminate the $(\log n)^{1+\varepsilon}$ factors in (4.16)-(4.17).

For the general case when $\bar{g}_n$ may have its stationary points at the boundary of $\boldsymbol{\Theta}$, we cannot measure first-order optimality simply by the gradient norms as above. In this case, for the empirical loss minimization, (4.12) in Theorem 4 states that

$$\min_{1 \le k \le n} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| \le 1} \left\langle -\nabla \bar{f}_k(\boldsymbol{\theta}_k), \, \boldsymbol{\theta} - \boldsymbol{\theta}_k \right\rangle \right] = O\left(\frac{(\log n)^{(1+\varepsilon)/2}}{n^{1/4}} \right),$$

which obtains the same asymptotic rate of convergence as in the interior stationary points case above. As for the expected loss minimization, we need to assume the strong er condition $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$ (i.e., the optional condition in (A4)). In this case, the optimal value of β is $3/4$, in which case we obtain a rate

$$\min_{1 \le k \le n} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| \le 1} \left\langle -\nabla f(\boldsymbol{\theta}_k), \, \boldsymbol{\theta} - \boldsymbol{\theta}_k \right\rangle \right] = O\left(\frac{(\log n)^{(1+\varepsilon)/2}}{n^{1/8}} \right)$$

which is slower than the bound in (4.15). While the $n^{-1/4}$ bound on the rate of convergence for the expected loss in the general case may not be sharp, this discussion gives a more quantitative indication that SRMM (and hence SMM) is generically more suited to solve the empirical loss minimization than the expected loss minimization for constrained nonconvex objectives.

Lastly, we state a corollary of the earlier results on the iteration complexity of Algorithm 1. Corollary 6 gives a "worst-case" rate of convergence of Algorithm 1 until reaching an ε-stationary points of the objective functions.

Corollary 6 (Iteration Complexity) *Let $(\boldsymbol{\theta}_n)_{n \ge 1}$ be an output of Algorithm 1. Make the same assumption as in Theorem 2. Then the following hold:*

(i) *(Empirical Loss Minimization) Suppose $w_n = n^{-1/2}(\log n)^{\delta}$ for some $\delta > 1$. Then we have the following worst-case iteration complexity for Algorithm 1 with the empirical loss objective:*

$$N_{\varepsilon}(\bar{f}_n) = O(\varepsilon^{-4}(\log \varepsilon^{-1})^{\delta}).$$

(ii) *(Expected Loss Minimization) Suppose $w_n = n^{-3/4}(\log n)^{\delta}$ for some $\delta > 1$. Then we have the following worst-case iteration complexity for Algorithm 1 with the expected loss objective:*

$$N_{\varepsilon}(f) = O(\varepsilon^{-8}(\log \varepsilon^{-1})^{\delta}).$$

Furthermore, suppose $\boldsymbol{\theta}_n$ is in the interior of $\boldsymbol{\Theta}$ for $n \ge 1$. Then we may choose $w_n = n^{-1/2}(\log n)^{\delta}$ for some $\delta > 1$ and obtain

$$N_{\varepsilon}(f) = O(\varepsilon^{-4}(\log \varepsilon^{-1})^{\delta}).$$

4.4 Remarks on main results

Here we give some remarks regarding the main results stated in the previous subsection.

Remark 7 (SRMM vs. PSGD on empirical and expected loss minimization)
Theorem 3 gives an upper bound on the rate of convergence of our proposed algorithm of SRMM (Algorithm 1) both for empirical and expected loss minimization. It is interesting to compare the rate of convergence bounds for SRMM and projected stochastic gradient descent (PSGD) algorithms with respect to both empirical and expected loss minimization. In this discussion, we omit log factors in all error bounds for simplicity and use $\widetilde{O}$ notation for big-O modulo ploy-log factors.
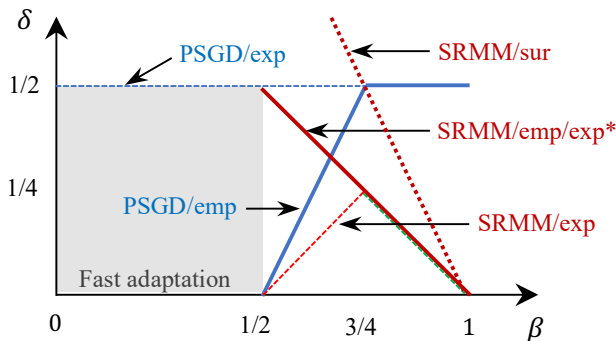

Figure 3: Diagram of upper bounds of the rate of convergence $\widetilde{O}(n^{-\delta})$ of algorithms SRMM (red) and PSGD (blue) for empirical loss minimization (denoted 'emp' in solid lines) with adaptivity weight $w_n = n^{-\beta}$, for expected loss minimization (denoted 'exp' in dashed lines), and for surrogate loss minimization (denoted 'sur' in dotted line). The red solid line SRMM/exp* is the case when the iterates are assumed to be in the interior of the constraint set. The red dotted line linearly extends to $\beta = 1/2$.

It is known that the optimal asymptotic rate of convergence of SGD for nonconvex problems measured in terms of gradient norm squared is $\widetilde{O}(1/\sqrt{n})$ (Bottou et al., 2018; Xu et al., 2019). Similarly, PSGD for constrained nonconvex problems also has the same optimal rate of convergence (Davis and Drusvyatskiy, 2019), although a direct comparison is not immediate in this case since a near-stationarity measure using gradient norm squared of Moreau envelope is used (see Sec. 4.1). It should be noted that all these convergence rate bounds are with respect to the expected loss functions and considering (P)SGD rate of convergence for empirical loss is less standard in the literature. Nonetheless, one can transfer the rate of convergence between these two objectives, at least in expectation, using the following inequality:

$$\mathbb{E}\left[\left|\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq1}\left\langle-\nabla\bar{f}_k(\boldsymbol{\theta}_k),\,\boldsymbol{\theta}-\boldsymbol{\theta}_k\right\rangle - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq1}\left\langle-\nabla f(\boldsymbol{\theta}_k),\,\boldsymbol{\theta}-\boldsymbol{\theta}_k\right\rangle\right|\right]$$

$$\leq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla\bar{f}(\boldsymbol{\theta})-\nabla f(\boldsymbol{\theta})\|\right] = O(w_n\sqrt{n}).$$

Indeed, the inequality follows from the Cauchy-Schwarz inequality and the equality is from Lemma 20. Assuming $w_n = n^{-\beta}$ for $\beta \in [0, 1]$, this implies

$$
\min_{1 \le k \le n} \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| \le 1} \left\langle -\nabla \bar{f}_k(\boldsymbol{\theta}_k),\, \boldsymbol{\theta} - \boldsymbol{\theta}_k \right\rangle \right]
$$

$$
\le \min_{1 \le k \le n} \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| \le 1} \left\langle -\nabla f(\boldsymbol{\theta}_k),\, \boldsymbol{\theta} - \boldsymbol{\theta}_k \right\rangle \right] + O(n^{-\beta + 1/2}).
$$

For PSGD (also for SGD), the first term on the right-hand side is of $\widetilde{O}(n^{-1/4})$, so this implies that the rate of convergence for PSGD with respect to the empirical loss function is bounded by $\tilde{O}(n^{-\delta})$ with $\delta = \min(1/2, 2\beta - 1)$. Note that δ stays constant $1/2$ for $\beta \in [3/4, 1]$ and decays linearly to 0 when β decreases from $3/4$ to $1/2$ (see the solid blue line in Figure 3). This means the upper bound on the rate of convergence of PSGD for empirical loss minimization becomes degenerate when β is closed to $1/2$.

On the contrary, according to Theorem 4, the upper bound on the rate of convergence of SRMM for the empirical loss function in fact improves linearly as β decays from 1 to $1/2$, with a better rate than PSGD over the interval $\beta \in [1/2, 2/3)$, achieving rate $\widetilde{O}(n^{-1/2})$ at $\beta = 1/2$. (See the solid red line in Figure 3). In our analysis, for SRMM, we obtain twice the rate of convergence $\widetilde{O}(n^{-2(1-\beta)})$ when the averaged surrogates are minimized approximately (see the red dotted line in Figure 3), and then we transferred it to a rate bound for the empirical loss and then to the expected loss. This comparison perhaps indicates that SRMM could be more adapted to and also more effective in minimizing empirical loss function than PSGD is, but PSGD may in general be more adapted to minimizing the expected loss than SRMM is.

Our analysis of SRMM critically relies on the fact that we can compare the averaged surrogate loss $\bar{g}_n$ with the empirical loss $\bar{f}_n$, as stated in Lemma 20. This result is only valid in what we call the 'slow adaptation regime' $\beta \in (1/2, 1]$. This is where we can use CLT-type arguments to connect the empirical and the expected loss minimization. Hence, it is an interesting open problem to analyze the rate of convergence of SRMM in the 'fast adaptation regime' of $\beta \in [0, 1/2]$, which is depicted as the grey region in Figure 3. Empirical loss in this regime will quickly adapt to newly observed data, so minimizing it will capture more fast-paced, short-time-scale features from streaming data. We speculate that SRMM is more effective than PSGD in such a regime.

Remark 8 (Iterate stability and regularization) We give some remarks on the use of proximal regularization and diminishing radius in Algorithms 3 and 2, and also why we need to assume i.i.d. data sampling for the case C3 of weakly convex surrogates with proximal regularization.

As we mentioned earlier in Section 3, a key property of Algorithm 1 for cases C1-C2 stated in Theorem 2 is the iterate stability $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ (see Lemma 17), which was critically used in the SMM literature Mairal et al. (2010); Mairal (2013b); Mensch et al. (2017); Lyu et al. (2020, 2022). It is crucial in deducing asymptotic convergence to stationary points stated in Theorem 2 by using Lemma 19. Since w_n's are square summable (see (A4)), it implies the 'weak iterate stability', $\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 < \infty$. Our analysis shows that we can still retain the rate of convergence results stated in Theorems 3 and 4 if we

only had this weak iterate stability. While in the classical block coordinate descent setting of non-stochastic optimization the weak iterate stability is sufficient to derive asymptotic stationarity of the iterates (see, e.g., Grippo and Sciandrone (2000), Xu and Yin (2013)), it does not seem to be immediate for the stochastic setting we consider in this work.

The technique of diminishing radius constraints we used in Algorithm 3 to handle multi-convex surrogates was first introduced in Lyu and Li (2023) to analyze convergence and complexity of cyclic BCD algorithms for minimizing block multi-convex functions on convex constraint sets. Enforcing additional trust region condition $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ with diminishing radii $O(w_n)$ bakes iterate stability directly into the algorithm, although that this auxiliary constraint is weak enough to retain asymptotic stability with respect to the original objective function needs to be argued (Lemma 23). In comparison, with proximal regularization in Algorithm 2, we were able to only show weak asymptotic stability of the form $\mathbb{E}[\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2] < \infty$ (see Lemma 21 **(vi)**). This was assuming i.i.d. data sampling assumption in comparison to the more general Markovian data assumption for cases C1-C2. In order to handle Markovian dependence (especially in proving Lemma 21), we use the technique of 'conditioning on distant past' first introduced in Lyu et al. (2020). It is to bound the error involving $\boldsymbol{\theta}_{n+1}$ by conditioning on the past σ-algebra $\mathcal{F}_{n-N}$ and using Markov chain mixing during the interval $[n-N, n]$. This requires to control $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-N}\|$, which is difficult without a priori iterate stability with proximal regularization. However, when the data samples are i.i.d., then we can condition directly on $\mathcal{F}_n$ so that we avoid using iterate stability.

Remark 9 (Asymptotic stationarity and inexact surrogate minimization) We discuss the effect of using inexact minimization of block multi-convex surrogates on convergence rates. First, consider the case C1 in Theorem 2 together with identically zero optimality gaps (see (A5)). In this case, each $\boldsymbol{\theta}_n$ is the minimizer of the strongly convex surrogates $\bar{g}_n$ over $\boldsymbol{\Theta}$ so that $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ is in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$. Hence Theorem 3 can be directly used to obtain (4.12) and (4.14). In particular, $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n)\|$ is identically zero when there are no boundary stationary points of $\bar{g}_n$.

In all other cases covered in C1-C3 in Theorem 2, each estimate $\boldsymbol{\theta}_n$ is only an approximate minimizer of the averaged surrogate $\bar{g}_n$ over $\boldsymbol{\Theta}$ due to the combination of the following factors: non-convexity of $\bar{g}_n$; the use of regularization in surrogate minimization in Algorithms 2 and 3; possible nonzero optimality gap (see (A5)) in solving convex subproblems in surrogate minimization. Hence, $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ is not necessarily in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$ in the general case. However, (4.11) shows that $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ should subsequentially converge to some vector in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$ at rate $O((\sum_{k=1}^{n} w_k)^{-1})$, and by (4.14), the bound on the rate of subsequential convergence to stationary points of f over $\boldsymbol{\Theta}$ is slower at order $O((\sum_{k=1}^{n} w_k)^{-1/2})$. We can ensure that these two convergence rates are achieved over the same subsequence. Hence, the asymptotic rate of convergence is unchanged when we relax strongly convex surrogates to weakly convex or block multi-convex surrogates.

5. Applications and Experiments

5.1 Applications

In this subsection, we discuss some applications of our general results in the setting of online matrix and tensor factorization problems.

5.1.1 DOUBLE-AVERAGING PSGD AND ITS GENERALIZATION

In this section, we will apply our general framework of SRMM (2.7) to derive convergence results for variants of PSGD such as the 'double-averaging PSGD' due to Nesterov and Shikhman (Nesterov and Shikhman, 2015) as well as its generalization.

Suppose we have a prescribed weight sequence $(w_n)_{m \geq 1}$ and hyperparameters $L, \lambda \geq 0$. Consider the following iterates

$$
\begin{cases}
\overline{\nabla}_n & \leftarrow (1 - w_n)\overline{\nabla}_{n-1} + w_n \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}) \quad (\overline{\nabla}_0 := \mathbf{0}) \\
\bar{\boldsymbol{\theta}}_{n-1} & \leftarrow (1 - w_n)\bar{\boldsymbol{\theta}}_{n-2} + w_n \boldsymbol{\theta}_{n-1} \quad (\bar{\boldsymbol{\theta}}_0 := \boldsymbol{\theta}_0) \\
\widetilde{\boldsymbol{\theta}}_{n-1} & \leftarrow \frac{L}{L+\lambda}\bar{\boldsymbol{\theta}}_{n-1} + \frac{\lambda}{L+\lambda}\boldsymbol{\theta}_{n-1} \quad (\widetilde{\boldsymbol{\theta}}_0 := \boldsymbol{\theta}_0) \\
\boldsymbol{\theta}_n & \leftarrow \mathrm{Proj}_{\boldsymbol{\Theta}}\left(\widetilde{\boldsymbol{\theta}}_{n-1} - \frac{1}{L+\lambda}\overline{\nabla}_n\right).
\end{cases}
\tag{5.1}
$$

Setting $\lambda = 0$ reduces (5.1) to the following 'double-averaging PSGD', which was first investigated by Nesterov and Shikhman (Nesterov and Shikhman, 2015):

$$
\begin{cases}
\overline{\nabla}_n & \leftarrow (1 - w_n)\overline{\nabla}_{n-1} + w_n \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}) \quad (\overline{\nabla}_0 := \mathbf{0}) \\
\bar{\boldsymbol{\theta}}_{n-1} & \leftarrow (1 - w_n)\bar{\boldsymbol{\theta}}_{n-2} + w_n \boldsymbol{\theta}_{n-1} \quad (\bar{\boldsymbol{\theta}}_0 := \boldsymbol{\theta}_0) \\
\boldsymbol{\theta}_n & \leftarrow \mathrm{Proj}_{\boldsymbol{\Theta}}\left(\bar{\boldsymbol{\theta}}_{n-1} - \frac{1}{L}\overline{\nabla}_n\right).
\end{cases}
\tag{5.2}
$$

Compared to the standard PSGD update $\boldsymbol{\theta}_n \leftarrow \mathrm{Proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_{n-1} - \frac{1}{L}\nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_k, \boldsymbol{\theta}_{k-1})\right)$ with fixed step size $1/L$, the algorithm (5.2) uses the recursively averaged gradient $\overline{\nabla}_n$ as well as the recursively averaged iterate $\bar{\boldsymbol{\theta}}_{n-1}$. The update $\boldsymbol{\theta}_n \leftarrow \mathrm{Proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_{n-1} - \frac{1}{L}\overline{\nabla}_n\right)$, where one uses the averaged gradients in place of the stochastic gradient, is known as 'dual-averaging' Nesterov (2009); Xiao (2009). For convex objectives in the online setting with i.i.d. observations, Xiao (2009) obtained a bound on the regret of this method of order $O(\sqrt{t})$. Compared to such dual-averaging methods, the double-averaging method (5.2) uses additional inline averaging of the iterates, which is known to be equivalent to using a momentum (Defazio, 2020).

In Nesterov and Shikhman (2015), Nestrove and Shikhman showed that for convex (possibly non-smooth) objective functions in the offline setting, the double-averaging method (5.2) generates a sequence of iterates that decreases the optimality gap in the objective value as $O(t^{-1/2})$. Moreover, such a rate of convergence is justified for the *whole sequence* of test points for the first time for subgradient methods. However, to the author's best knowledge, the double-averaging method has not been analyzed in the stochastic non-convex setting. This is in contrast to the standard PSGD case, which has been analyzed for general non-smooth nonconvex objectives under i.i.d. (Davis and Drusvyatskiy, 2019) and Markovian (Alacaoglu and Lyu, 2023) data assumption. As an application of our general SRMM analysis, we will derive very general convergence and complexity results for the iterates (5.1) (and hence for its specialization (5.2) as well).

Assuming the per-sample loss function $f_n(\boldsymbol{\theta}) := \ell(\mathbf{x}, \boldsymbol{\theta})$ is L-smooth for each data point $\mathbf{x}$, the following prox-linear surrogate (see Ex. 3) is indeed a majorizing surrogate of f_n:

$$g_n(\boldsymbol{\theta}) := \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}) + \langle \nabla_{\boldsymbol{\theta}} \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2. \tag{5.3}$$

Now we claim that the generalized double-averaging scheme (5.1) is in fact equivalent to the following SRMM iterate

$$\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \right), \tag{5.4}$$

where $\bar{g}_n(\boldsymbol{\theta}) := (1 - w_n)\bar{g}_{n-1}(\boldsymbol{\theta}) + w_n g_n(\boldsymbol{\theta})$ denotes the recursive average of the surrogates as in (2.7). Indeed, letting $\widetilde{\boldsymbol{\theta}}_{n-1} := \frac{L}{L+\lambda}\bar{\boldsymbol{\theta}}_{n-1} + \frac{\lambda}{L+\lambda}\boldsymbol{\theta}_{n-1}$, one can easily see that (5.4) is equivalent to

$$\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \langle \overline{\nabla}_n, \boldsymbol{\theta} \rangle + \frac{L+\lambda}{2} \|\boldsymbol{\theta}\|^2 - (L+\lambda) \left\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}_{n-1} \right\rangle$$

$$= \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \boldsymbol{\theta} - (\widetilde{\boldsymbol{\theta}}_{n-1} - \frac{1}{L+\lambda}\overline{\nabla}_n) \right\|^2$$

$$= \text{Proj}_{\boldsymbol{\Theta}} \left(\widetilde{\boldsymbol{\theta}}_{n-1} - \frac{1}{L+\lambda}\overline{\nabla}_n \right),$$

verifying the claim. The above discussion also shows that SRMM with prox-linear surrogates (5.3) satisfies (A7), as one only needs to carry the averaged gradients $\overline{\Delta}_n$ and averaged iterates $\boldsymbol{\theta}_n$ for executing (5.1). In particular, SRMM in this case has bounded per-iteration cost, as other first-order methods (e.g., SGD, Adam) do. See 5.2.2 for more discussion.

Interestingly, the above derivation also tells us that the double-averaging scheme (5.2) is equivalent to the SMM update $\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \bar{g}_n(\boldsymbol{\theta})$ with the prox-linear surrogates in (5.3). Using the additional proximal regularization with parameter $\lambda \geq 0$ as in (5.4) has the effect of additional averaging of the parameters, giving a constant weight $\frac{\lambda}{L+\lambda}$ to the most recent iterate $\boldsymbol{\theta}_{n-1}$ (see (5.1)) instead of the possibly decaying weight w_n (see (5.2)).

Now that we have verified the generalized double-averaging PSGD (5.1) is a special instance of SRMM (2.7) with prox-linear surrogate and proximal regularizer, we obtain the following general convergence and complexity results for the (5.1) as well as (5.2).

Corollary 10 (Convergence and complexity of double-averaging PSGD) *Theorems 2, 3, and 4 hold for the double-averaging PSGD (5.2) (case C1 with prox-linear surrogates) and its generalization (5.1) (case C3 with prox-linear surrogates and proximal regularization).*

Next, our general framework enables us to consider a block coordinate version of the double-averaging PSGD (5.2), which will be an example of Algorithm 1 in case C2. For instance, instead of updating the entire parameter $\boldsymbol{\theta}_n$, we may only update its ith block $\theta_n^{(i)}$ where i is chosen uniformly at random independently at each iteration:

$$\begin{cases} i & \sim \text{Uniform}(\{1, \ldots, m\}) \\ \overline{\nabla}_n^{(i)} & \leftarrow (1 - w_n)\overline{\nabla}_{n-1}^{(i)} + w_n \nabla_{\theta^{(i)}} \ell(\mathbf{x}_n, \boldsymbol{\theta}_{n-1}) \quad (\overline{\nabla}_0^{(i)} := \mathbf{0}) \\ \bar{\theta}_{n-1}^{(i)} & \leftarrow (1 - w_n)\bar{\theta}_{n-2}^{(i)} + w_n \theta_{n-1}^{(i)} \quad (\bar{\theta}_0^{(i)} := \theta_0^{(i)}) \\ \theta_n^{(i)} & \leftarrow \text{Proj}_{\Theta^{(i)} \cap \{\theta : \|\theta - \theta_{n-1}^{(i)}\| \leq w_n\}} \left(\bar{\theta}_{n-1}^{(i)} - \frac{1}{L}\overline{\nabla}_n^{(i)} \right). \end{cases} \tag{5.5}$$

Note that in (5.5), we used the diminishing radius regularization to update $\theta_n^{(i)}$. Also notice that the above block version of (5.2) has the computational advantage that one only needs to compute the partial stochastic gradient $\nabla_{\theta^{(i)}}\ell(\mathbf{x}_k, \boldsymbol{\theta}_{k-1})$ instead of the full stochastic gradient $\nabla_{\boldsymbol{\theta}}\ell(\mathbf{x}_k, \boldsymbol{\theta}_{k-1})$, which could be significant when the parameter dimension p is large. For instance, here we can choose individual coordinates as single blocks so the per-iteration computational cost of (5.5) is of $O(1)$ plus the cost of sampling the new data point $\mathbf{x}_n$ and for the projection onto the convex set $\Theta^{(i)} \cap \{\theta : \|\theta - \theta_{n-1}^{(i)}\| \leq w_n\}$.

As before, our general result immediately implies the following convergence and complexity result for the randomized block variant of the double-averaging PSGD (5.5).

Corollary 11 (Convergence and complexity of double-averaging PSGD) *Theorems 2, 3, and 4 hold for the randomized block variant of the double-averaging PSGD (5.5) (case C2 with prox-linear surrogates and randomized coordinate descent)*

We remark that the above corollary holds if we used cyclic block coordinate descent in (5.5) instead of randomly choosing block coordinates to update.

5.1.2 PROXIMAL POINT EMPIRICAL LOSS MINIMIZATION

Consider the following iterates (5.6), where we recursively update the empirical loss function $\bar{f}_n$ and minimize is proximal point modification $\tilde{f}_n$ over the parameter space $\boldsymbol{\Theta}$.

$$\begin{cases} \bar{f}_n(\boldsymbol{\theta}) & \leftarrow (1 - w_n)\bar{f}_{n-1}(\boldsymbol{\theta}) + w_n\ell(\mathbf{x}_n, \boldsymbol{\theta}) \\ \boldsymbol{\theta}_n & \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\tilde{f}_n(\boldsymbol{\theta}) := \bar{f}_n(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2\right) \end{cases} \tag{5.6}$$

Under (A1), each $\bar{f}_n$ is L-smooth so if $\lambda > L$, then the proximal point modification $\tilde{f}_n$ is $(\lambda - L)$-strongly convex (see Lemma 36). The proximal point method has been used in empirical loss minimization problems with convex objectives Frostig et al. (2015); Lin et al. (2015).

It is easy to see that (5.6) is an instance of case C3, where we use the trivial surrogate $g_n(\cdot) = \ell(\mathbf{x}_n, \cdot)$. Hence our general results yield that the proximal point ERM (5.6) in the general non-convex constrained setting with dependent data.

Corollary 12 (Convergence and complexity of Proximal Point ERM) *Theorems 2, 3, and 4 hold for the proximal point ERM (5.6).*

5.1.3 ONLINE (NONNEGATIVE) MATRIX FACTORIZATION

Consider the matrix factorization loss $f(W, H) := \|X - WH\|^2 + \lambda\|H\|_1$, where $X \in \mathbb{R}^{p \times d}$ is a given data matrix to be factorized into the product of dictionary $W \in \mathbb{R}^{p \times r}$ and the code $H \in \mathbb{R}^{r \times d}$ with $\lambda \geq 0$ being the L_1-regularization parameter for H. Clearly f is two-block multi-convex and differentiable with respect to W with gradient $\nabla_W f(W, H) = (X - WH)H^T$. Fix compact and convex constraint sets $\Theta_1 \subseteq \mathbb{R}^{p \times r}$ and $\Theta_2 \subseteq \mathbb{R}^{r \times d}$. Then $W \mapsto \nabla f(W, H)$ and $H \mapsto \nabla f(W, H)$ are both Lipschitz over the compact and convex set $\Theta_1 \times \Theta_2$.

Suppose we have a sequence of data matrices $(X_n)_{n \geq 1}$ and a sequence of weights $(w_n)_{n \geq 1}$. For each $n \geq 1$, the function $f_n : W \mapsto \ell(X_n, W) = \inf_{H \in \Theta_2}\|X_n - WH\|^2 + \lambda\|H\|_1$ is

the minimum reconstruction error for factorizing X_n using the dictionary matrix W. The corresponding empirical loss minimization problem is

$$W_n \in \arg\min_{W \in \Theta_1} \bar{f}_n(W) := (1 - w_n)\bar{f}_{n-1}(W) + w_n \ell(X_n, W).$$

If we have a target distribution π for the data matrices X_n, then the corresponding expected loss minimization problem is

$$W \in \arg\min_{W \in \Theta_1} f(W) := \mathbb{E}_{X \sim \pi}\left[\ell(X, W)\right]. \tag{5.7}$$

The latter is known as the online matrix factorization problem introduced in Mairal et al. (2010). A well-known instance is the online nonnegative matrix factorization, which corresponds to (5.7) with nonnegativity constraints for W and H.

In order to apply Algorithm 1 in this setting, denote $H_n \in \arg\min_{H \in \Theta_2} \|X_n - W_{n-1}H\|_F^2 + \lambda\|H\|_1$, which is an optimal code for factorizing X_n using the previous dictionary W_{n-1}. Then the function $g_n(W) = \|X_n - WH_n\|_F^2 + \lambda\|H_n\|_1$ is a surrogate of $f_n : W \mapsto \ell(X_n, W)$ at W_{n-1} and belongs to $\mathtt{Srg}_{L'',0}(f_n)$ for some $L'' > 0$ (see Ex. 6.) A simple calculation shows that the resulting averaged surrogate function $\bar{g}_n$ becomes

$$\bar{g}_n(W) = \mathrm{tr}(WA_nW^T) - 2\mathrm{tr}(WB_n) + C_n,$$

where the matrices A_n, B_n, C_n are recursively defined as

$$A_n = (1 - w_n)A_{n-1} + w_n H_n H_n^T; \quad B_n = (1 - w_n)B_{n-1} + w_n H_n X_n^T;$$
$$C_n = (1 - w_n)\bar{X}_{n-1} + w_n X_n X_n^T.$$

Thus Algorithm 1 in this case reduces to

$$\textit{Upon arrival of } X_n: \quad \begin{cases} H_n = \arg\min_{H \in \Theta_2 \subseteq \mathbb{R}^{r \times d}} \|X_n - W_{n-1}H\|_F^2 + \lambda\|H\|_1 \\ A_n = (1 - w_n)A_{n-1} + w_n H_n H_n^T \\ B_n = (1 - w_n)B_{n-1} + w_n H_n X_n^T \\ W_n = \arg\min_{W \in \Theta_1 \subseteq \mathbb{R}^{p \times r}} \left(\mathrm{tr}(WA_nW^T) - 2\mathrm{tr}(WB_n)\right), \end{cases} \tag{5.8}$$

which is the online matrix factorization algorithm (OMF) proposed in Mairal et al. (2010) for i.i.d. data matrices. Later in Lyu et al. (2020), this algorithm was analyzed in a Markovian data setting. The following corollary is a direct consequence of our general results.

Corollary 13 (Rate of convergence of OMF) *Theorems 3 and 4 hold for the online matrix factorization algorithm (5.8) (case C1).*

To the author's best knowledge, such a result was not known for online matrix factorization even under the i.i.d. data setting.

5.1.4 SUBSAMPLED ONLINE MATRIX FACTORIZATION

While the OMF algorithm (5.8) is efficient in handling matrices with a large number of columns d, its computational cost with respect to the data dimension p (i.e., the number of rows) is not reduced. In Mensch et al. (2017), subsampled OMF was proposed to improve the computational efficiency with respect to p, by using only a random subsample of rows. For a preliminary version, consider the following variant of the online matrix factorization algorithm 5.8, which uses a random coordinate descent on subsampled rows of W_n:

Upon arrival of X_n:

$$
\begin{cases}
H_n = \arg\min_{H \in \Theta_2 \subseteq \mathbb{R}^{r \times d}} \|X_n - W_{n-1}H\|_F^2 + \lambda\|H\|_1 \\
A_n = (1 - w_n)A_{n-1} + w_n H_n H_n^T \\
B_n = (1 - w_n)B_{n-1} + w_n H_n X_n^T \\
J \leftarrow \text{Random subset of } \{1, \ldots, p\} \\
W_n = \arg\min_{W \in \Theta_1 \subseteq \mathbb{R}^{p \times r}} \left(\text{tr}(WA_nW^T) - 2\text{tr}(WB_n) \right) \\
\qquad \text{while freezing rows of } W_n \text{ not in } J \text{ to the rows of } W_{n-1}.
\end{cases}
\tag{5.9}
$$

For instance, J may be taken as the uniform subset of $\{1, \ldots, p\}$ of a fixed size $p' < p$ or it may contain each index $i \in \{1, \ldots, p\}$ independently with a fixed probability. All our main results in this paper (Theorems 2, 3, and 4) apply to the OMF algorithm in (5.9).

However, note that the code computation for H_n in (5.9) still involves solving a least squares problem with p dimensional data matrix X_n. In order to fully reduce the dependency on p, one may also use row subsampling to compute only p' rows of H_n. The approach used in Mensch et al. (2017) was to replace X_n with an averaged p'-dimenisonal matrix $\bar{X}_n^{(i)}$ assuming $X_n = X^{(i)}$, where a finite pool of data matrices $\mathfrak{X} = \{X^{(1)}, \ldots, X^{(l)}\}$ was assumed and $\bar{X}_n^{(i)}$ is a recursively defined matrix of size $p' \times d$ based on previous occurances of the matrix $X^{(i)}$. See (Mensch et al., 2017, Alg. 3) for more details.

An almost sure convergence to stationary points of subsampled OMF algorithm under i.i.d. data samples was shown in Mensch et al. (2017). The analysis there is based on a general convergence result on stochastic approximate majorization-minimization algorithm (2.6) using strongly convex ε-approximate surrogate functions (see (Mensch et al., 2017, Prop. 3)). A similar convergence result is retained by Theorem 2, although the assumptions are slightly different. In addition, Theorems 3-4 also provide a rate of convergence, as stated in the following corollary.

Corollary 14 (Rate of convergence of Subsampled OMF) *Theorems 3 and 4 hold for the subsampled online matrix factorization algorithm (5.9) (case C1 with randomized block coordinate descent).*

5.1.5 ONLINE CP-DICTIONARY LEARNING

Suppose we have $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}_{\geq 0}^{I_1 \times \cdots \times I_m}$, which are N observed m-mode tensor-valued signals. Consider the following tensor-valued dictionary learning problem

$$
[\mathbf{X}_1, \ldots, \mathbf{X}_b] \approx [\mathbf{D}_1, \ldots, \mathbf{D}_r] \times_{m+1} H,
$$

where $\times_{m+1}$ denotes the mode-$(m+1)$ tensor-matrix product and we impose the tensor dictionary atoms $\mathbf{D}_1, \ldots, \mathbf{D}_r \in \mathbb{R}^{I_1 \times \cdots \times I_m \times r}$ to be of rank-1 and $H \in \mathbb{R}^{R \times b}$ is called a code matrix. Equivalently, we assume that there exist *loading matrices* $[U^{(1)}, \ldots, U^{(n)}] \in \mathbb{R}_{\geq 0}^{I_1 \times r} \times \cdots \times \mathbb{R}_{\geq 0}^{I_n \times r}$ such that

$$[\mathbf{D}_1, \ldots, \mathbf{D}_r] = \texttt{Out}(U^{(1)}, \ldots, U^{(m)})$$
$$:= \left[\bigotimes_{k=1}^m U^{(k)}(:, 1), \bigotimes_{k=1}^m U^{(k)}(:, 2), \ldots, \bigotimes_{k=1}^m U^{(k)}(:, r) \right] \in \mathbb{R}_{\geq 0}^{I_1 \times \cdots \times I_m \times r},$$

where $U^{(k)}(:, j)$ denotes the ith column of the $I_k \times r$ matrix $U^{(k)}$ and $\otimes$ denotes the outer product. Since we impose a CANDECOMP/PARAFAC (CP) Tucker (1966); Harshman (1970); Carroll and Chang (1970) structure for the tensor-valued dictionary $[\mathbf{D}_1, \ldots, \mathbf{D}_r]$, the above is called a CP-dictionary learning problem introduced in Lyu et al. (2022). When $m = 1$, it reduces to the standard vector-valued dictionary learning problem.

In Lyu et al. (2022), the following online CP-dictionary learning problem was proposed and analyzed. Fix compact and convex constraint sets for code and loading matrices $\mathcal{C}^{\text{code}} \subseteq \mathbb{R}^{R \times b}$ and $\mathcal{C}^{(i)} \subseteq \mathbb{R}^{I_i \times r}$, $i = 1, \ldots, n$, respectively. Write $\boldsymbol{\Theta} := \mathcal{C}^{(1)} \times \cdots \times \mathcal{C}^{(n)}$. For each $\mathcal{X} \in \mathbb{R}_{\geq 0}^{I_1 \times \cdots \times I_n \times b}$, $\mathbf{U} := [U^{(1)}, \ldots, U^{(n)}] \in \mathbb{R}^{I_1 \times r} \times \cdots \times \mathbb{R}^{I_n \times r}$, $H \in \mathbb{R}^{R \times b}$, define

$$\ell(\mathcal{X}, \mathbf{U}, H) := \|\mathcal{X} - \texttt{Out}(\mathbf{U}) \times_{n+1} H\|_F^2 + \lambda \|H\|_1, \tag{5.10}$$
$$\ell(\mathcal{X}, \mathbf{U}) := \inf_{H \in \mathcal{C}^{\text{code}}} \ell(\mathcal{X}, \mathbf{U}, H),$$

where $\lambda \geq 0$ is a regularization parameter. Fix a sequence of non-increasing weights $(w_t)_{t \geq 0}$ in $(0, 1]$. Here $\mathcal{X}$ denotes a minibatch of b tensors in $\mathbb{R}^{I_1 \times \cdots \times I_n}$, so minimizing $\ell(\mathcal{X}, \mathbf{U})$ with respect to $\mathbf{U}$ amounts to fitting the CP-dictionary $\texttt{Out}(\mathbf{U})$ to the minibatch of b tensors in $\mathcal{X}$. The corresponding empirical loss minimization problem is

$$\mathbf{U}_n \in \arg\min_{\mathbf{U} \in \boldsymbol{\Theta}} \left(\bar{f}_n(\mathbf{U}) := (1 - w_n) \bar{f}_{n-1}(\mathbf{U}) + w_n \ell(\mathcal{X}_n, \mathbf{U}) \right). \tag{5.11}$$

If we have a target distribution π for the data tensors $\mathcal{X}_n$, then the corresponding expected loss minimization problem is

$$\mathbf{U} \in \arg\min_{\mathbf{U} \in \boldsymbol{\Theta}} \left(f(W) := \mathbb{E}_{X \sim \pi} [\ell(\mathcal{X}, \mathbf{U})] \right). \tag{5.12}$$

In order to solve (5.11) and (5.12), the following online CP-dictionary learning algorithm was proposed and analyzed in Lyu et al. (2022):

Upon arrival of $\mathcal{X}_n$: $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (5.13)

$$\begin{cases} H_n = \arg\min_{H \in \mathcal{C}^{\text{code}} \subseteq \mathbb{R}^{r \times b}} \ell(\mathcal{X}_n, \mathbf{U}_{n-1}, H) \\ \bar{g}_n(\mathbf{U}) = (1 - w_n) \bar{g}_{n-1}(\mathbf{U}) + w_n \ell(\mathcal{X}_n, \mathbf{U}, H_n) \\ \text{for } i = 1, \ldots, m : \\ \quad U_n^{(i)} \in \arg\min_{\substack{U \in \mathcal{C}^{(i)} \\ \|U - U_{n-1}^{(i)}\| \leq w_n}} \bar{g}_n(U_n^{(1)}, \ldots, U_n^{(i-1)}, U, U_{n-1}^{(i+1)}, \ldots, U_{n-1}^{(m)}). \end{cases}$$

An almost sure convergence to the stationary points of the above algorithm under the Markovian data setting was obtained in Lyu et al. (2022). In fact, the algorithm (5.13) is a special case of Algorithm 1 with m-block multi-convex variational surrogate corresponding (see Example 7) to the function ℓ in (5.10) with $\mathbb{J} = \{J_1, \ldots, J_m\}$, where J_i denotes the coordinate block corresponding to the ith loading matrix in $\mathbb{R}^{I_i \times r}$ with cyclic block coordinate descent in Algorithm 3.

Our generalized algorithm and refined analysis improve the results in Lyu et al. (2022) in multiple ways. First, Theorems 3 and 4 provide a rate of convergence results for the online CP-dictionary learning algorithm (5.13), which has not been known before.

Corollary 15 (Rate of convergence of OCPDL) *Theorems 3 and 4 hold for the online CP-dictionary learning algorithm* (5.13) *(case C2).*

Second, due to the flexibility of using approximate surrogate functions, the convergence results of (5.13) hold under inexact code or factor matrix computation, following a similar argument as in Mensch et al. (2017). Third, one can only optimize a small number of subsampled rows when updating the factor matrices $U_n^{(i)}$ in (5.13) by using a refined block coordinate structure (see Example 8). A similar idea of using subsampling for tensor CP-decomposition was also used recently in Kolda and Hong (2020).

Lastly, we remark that random row subsampling can be utilized for the code computation for H_n in (5.13), which essentially involves solving a least-squares problem with the mode-$(m+1)$ unfolding of $\mathcal{X}_n$, which has size $(I_1 \times \cdots \times I_m) \times b$. Since the number of rows of a such matrix $I_1 \times \cdots \times I_m$ can be very large, the computational gain in this approach should be significant. We believe it would be straightforward to adapt the approach in Mensch et al. (2017) for subsampled online matrix factorization for a subsampled online CP-dictionary learning setting. For our theoretical results to apply, one only needs to verify that the resulting inexact code computation for H_n yields ε_n-approximate surrogates $\bar{g}_n$ that verifies the assumption (A4). We do not proceed with this line of research in the present paper.

5.2 Experiments

In this section, we provide experimental results of SRMM on two tasks - Network Dictionary Learning **?** and Image classification with Deep Convolutional Neural Networks for the CIFAR-10 dataset Krizhevsky et al. (2009).

5.2.1 NETWORK DICTIONARY LEARNING

Network Dictionary Learning (NDL) Lyu et al. (2020); **?** is the task of learning a fixed number of 'latent motifs' from a large number of connected k-node subgraphs in a given large and possibly sparse networks. The learned latent motifs provide a concise description of a network's mesoscale structure and can be used for reconstructing and denoising networks (see **?** for more details). NDL is naturally formulated as a nonconvex and constrained stochastic optimization problem, where the input data is a stream of k-node subgraphs sampled by the MCMC motif-sampling algorithm Lyu et al. (2023).

The standard optimization algorithm for NDL is based on SMM for online nonnegative matrix factorization (see Sec. 5.1.3). We compare the performance of SMM (with adaptivity

weights $w_t = 1/t$) with that of various stochastic optimization algorithms – PSGD, PSGD with momentum (PSGD-HB) (with step size schedule $\alpha_t = 1/t$) and AdaGrad Ward et al. (2020).
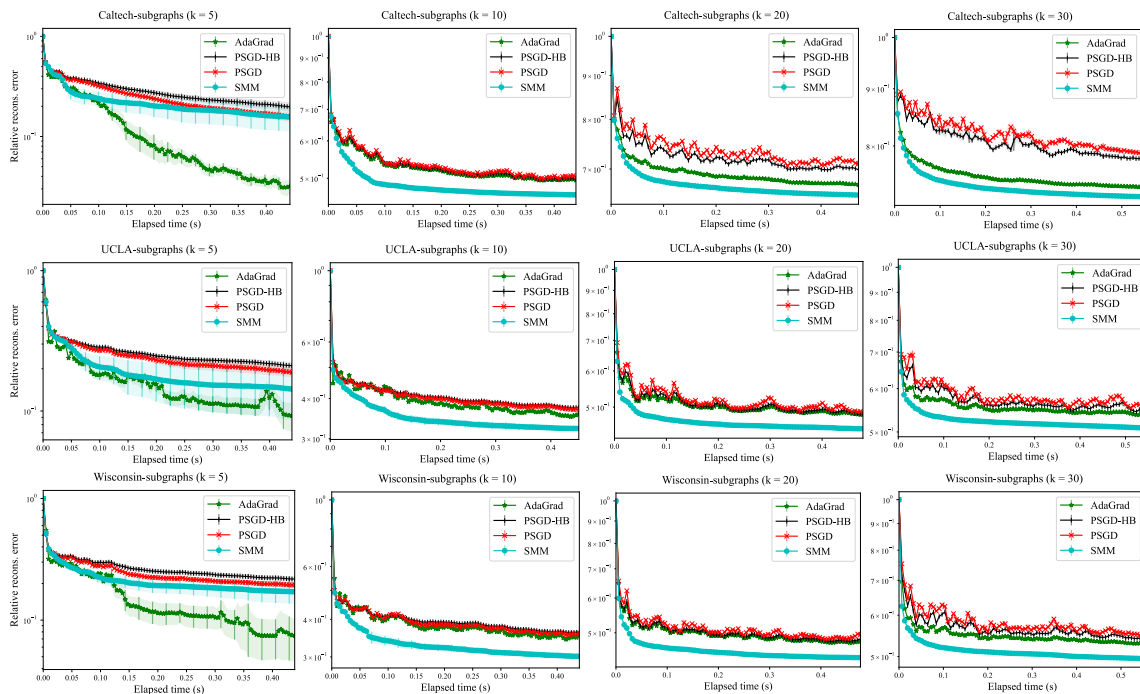
Figure 4: Plot of reconstruction error vs. elapsed time for four algorithms for online NMF: AdaGrad, PSGD-Heavy Ball, PSGD, and SRMM. The data stream is a sequence of 4-node subgraph adjacency matrices sampled by an MCMC motif-sampling algorithm in Lyu et al. (2023) from three college Facebook networks Traud et al. (2012). The shaded region represents one standard deviation from ten runs.

We consider three facebook networks of schools CALTECH, UCLA, and WISCONSIN from the Facebook100 dataset (Traud et al., 2012), following a similar setup to (Lyu et al., 2020). We then used the MCMC motif-sampling algorithm of (Lyu et al., 2023) to generate 300 correlated subgraphs from the networks. This gives a stream of 300 $k \times k$ binary subgraph adjacency matrices, from which the latent motifs need to be learned by solving an online nonnegative matrix factorization problem. We ran each stochastic optimization algorithm on this streaming data (in the order that each subgraph is sampled) exactly once (e.g., one epoch). We used four subgraph sizes: $k \in \{5, 10, 20, 30\}$.

In Fig. 4, we see the convergence of all the algorithms with respect to the normalized reconstruction error, which is in line with our theoretical results. We observe that SMM shows significantly faster convergence in all cases except the smallest subgraph size $k = 4$, in which case AdaGrad seems to converge faster than all the other methods. Overall SMM seems to be providing robust performance for NDL with various choices of subgraph sizes.

33
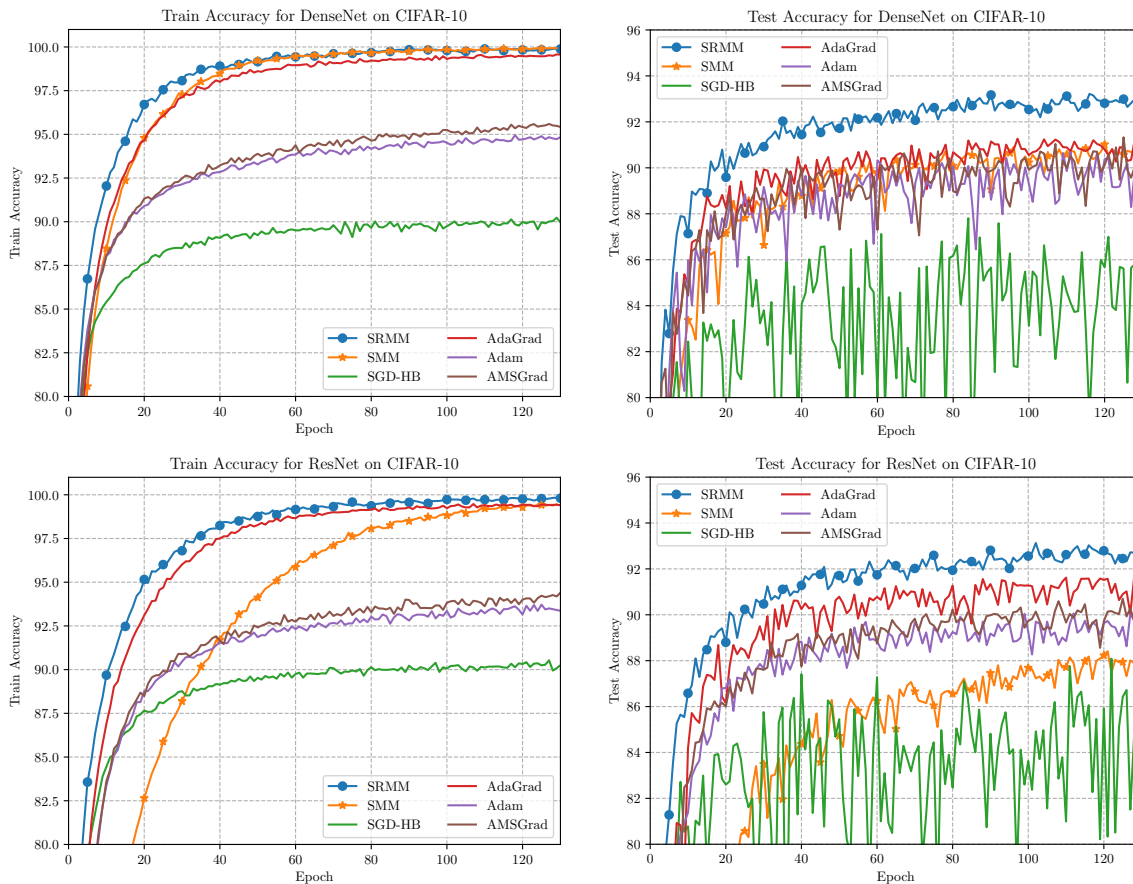
Figure 5: Plot of training and test accuracies for CIFAR-10 image classification using DenseNet-121 and ResNet-34 with various optimization algorithms – SRMM, SMM, SGD, AdaGrad, Adam, and AMSGrad.

5.2.2 Image classification using CNN

Using DenseNet-121 (Huang et al., 2017) and ResNet-34 (He et al., 2016), we then consider the task of image classification on the standard CIFAR-10 dataset (Krizhevsky et al., 2009). We compared six different optimization algorithms: (1) The generalized double-averaging scheme (5.1) (labeled as SRMM); (2) The double-averaging scheme (5.2) (labeled as SMM); (3) SGD-HB (SGD with momentum); (4) AdaGrad (Duchi et al., 2011); (5) Adam (Kingma and Ba, 2015); and (6) AMSGrad (Reddi et al., 2018). We use these algorithms for a total of 130 epochs and report the train and test accuracies in Figure 5. The particular implementation of SMM/SRMM with prox-linear surrogates that we use here ((5.2) and (5.1)) have bounded per-iteration cost and also satisfy (A7).

We see that SRMM achieves the highest train and test accuracies (around 93%) the most rapidly both for DenseNet and ResNet. SMM also shows a competitive performance against the benchmark methods for DenseNet, and outperforms SGD-HB for ResNet in test accuracy. In the experiments reported in Figure 5, the hyperparameters are chosen as follows. (1) SRMM: $w_n = n^{-1/2}$ for $n \geq 1$, $L \in \{10, 1\}$, and $\lambda \in \{10, 1, 10^{-1}, 10^{-2}, 10^{-3}\}$; (2) SMM:

$w_n = n^{-1/2}$ for $n \geq 1$ and $L \in \{10, 1\}$; (3) SGD-HB: step sizes in $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and momentum parameter of 0.9; (4) AdaGrad: stepsizes in $\{5e\text{-}4, 1e\text{-}3, 5e\text{-}3, 5e\text{-}2, 1e\text{-}1\}$ and accumulator value 0; (5)-(6) Adam and AMSGrad: step sizes in $\{1e\text{-}4, 5e\text{-}4, 1e\text{-}3, 5e\text{-}3, 1e\text{-}2\}$, $\beta_1 \in \{0.9, 0.99\}$, and $\beta_2 \in \{0.99, 0.999\}$. For SRMM and SMM, we reset the iteration counter to zero at the beginning of each epoch in order to avoid the adaptivity weight w_n becoming too small after many epochs.

6. Preliminary analysis

In this section, we give some preliminary lemmas that we will use in our analysis in the following sections. Recall the definition of the optimality gaps $\Delta_n^{(i)}$ in (A5). Also, we assume throughout that $g_n \in \mathtt{Srg}_{L,\rho}^{\mathbb{J}}(f_n, \boldsymbol{\theta}_{n-1}, \varepsilon_n)$ for $n \geq 1$ for some fixed set $\mathbb{J}$ of coordinate blocks, $L > 0$, and $\rho \in \mathbb{R}$.

We first observe some regularity properties of the surrogate gradients ∇g_n and $\nabla \bar{g}_n$.

Proposition 16 *Assume (A1) and (A3). Suppose that the data sequence $(\mathbf{x}_n)_{n \geq 1}$ is contained in some compact subset $\mathfrak{X}_0 \subseteq \mathfrak{X}$. Then the following hold for all $n \geq 1$:*

(i) *Both ∇g_n and $\nabla \bar{g}_n$ are L-Lipschitz over $\boldsymbol{\Theta}$;*

(ii) *$g_n, \bar{g}_n, \nabla g_n$, and $\nabla \bar{g}_n$ are uniformly bounded over $\boldsymbol{\Theta}$;*

(iii) *There exists a constant $L' > 0$ such that for all $n \geq 1$, $g_n, \bar{g}_n$ are L'-Lipschitz continuous over $\boldsymbol{\Theta}$ and $\|\nabla \bar{g}_{n+1}(\boldsymbol{\theta}) - \nabla \bar{g}_n(\boldsymbol{\theta})\| \leq L' w_{n+1}$.*

Proof Denote $f_n = \ell(\mathbf{x}_n, \cdot)$ for $n \geq 1$. Note that $\nabla \ell(\cdot, \cdot)$ is L-Lipscthiz continuous (see (A1)). Also, $\nabla(g_n - f_n)$ is L-Lipscthiz by Definition 1. It follows that ∇g_n is also L-Lipschitz over $\boldsymbol{\Theta}$. This holds for all $n \geq 1$, so by an induction, $\nabla \bar{g}_n$ is also L-Lipschitz over $\boldsymbol{\Theta}$ for $n \geq 1$. This shows **(i)**.

Note that by Definition 1, $\nabla(g_n - f_n)$ is L-Lipschitz continuous and has norm bounded by $\varepsilon_n \leq 1$ at $\boldsymbol{\theta}_{n-1}$. Since the parameter space $\boldsymbol{\Theta}$ is compact by (A3), it follws that $\nabla(g_n - f_n)$ is uniformly bounded by some constant $L' > 0$ over $\boldsymbol{\Theta}$. Also, by the assumption, the continuous map $(\mathbf{x}, \boldsymbol{\theta}) \mapsto \|\nabla \ell(\mathbf{x}, \boldsymbol{\theta})\|$ over the compact domain $\mathfrak{X}_0 \times \boldsymbol{\Theta}$ assumes bounded values. Hence $\nabla \ell$ is uniformly bounded over $\boldsymbol{\Theta}$ by (A1) and (A3). It follows that ∇f_n is uniformly bounded, so ∇g_n is also uniformly bonded. Then by induction, it follows that $\nabla \bar{g}_n$ is also uniformly bounded. Similarly, as $\|g_n(\boldsymbol{\theta}_{n-1}) - f_n(\boldsymbol{\theta}_{n-1})\| \leq \varepsilon_n \leq 1$ and f_n is uniformly bounded, it follows that g_n is bounded over $\boldsymbol{\Theta}$. By induction, it follows that $\bar{g}_n$ is also uniformly bounded over $\boldsymbol{\Theta}$.

To show **(iii)**, we first let $L' > 0$ be a uniform bound on $\|\nabla g_n\|$ and $\|\nabla \bar{g}_n\|$ over $\boldsymbol{\Theta}$. Fix $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ and consider the linear curve $t \mapsto \gamma(t) := t\boldsymbol{\theta}' + (1-t)\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. Then

$$\|g_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta}')\| = \left| \int_0^1 \langle \nabla g_n(\gamma(t)), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \, dt \right| \leq L' \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

This shows that g_n is L'-Lipscthiz for all $n \geq 1$. From this, one also concludes that $\bar{g}_n$ is L'-Lipscthiz for all $n \geq 1$.

Lastly, note that $\nabla \bar{g}_{n+1} = (1 - w_{n+1}) \nabla \bar{g}_n + w_{n+1} \nabla g_n$. Hence by **(ii)**, there exists $L' \in (0, \infty)$ such that

$$\|\nabla \bar{g}_{n+1}(\boldsymbol{\theta}) - \nabla \bar{g}_n(\boldsymbol{\theta})\| \leq w_{n+1} \|\nabla \bar{g}_n(\boldsymbol{\theta}) - \nabla g_n(\boldsymbol{\theta})\| \leq L' w_{n+1}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This shows **(iii)**. ∎

Define a sequence $(\bar{\varepsilon}_n)_{n \geq 1}$ recursively as

$$\bar{\varepsilon}_n = (1 - w_n) \bar{\varepsilon}_{n-1} + w_n \varepsilon_n, \qquad \bar{\varepsilon}_0 = 0,$$

where $(\varepsilon_n)_{n \geq 1}$ is the sequence of surrogate error tolerance in Algorithm 1. Then

$$\bar{\varepsilon}_n = \sum_{k=1}^{n} w_k^n \varepsilon_k \leq w_n \sum_{k=1}^{n} \varepsilon_k, \tag{6.1}$$

where w_k^n is defined in (2.2) and the inequality above uses $w_1^n \leq w_2^n \leq \cdots \leq w_n^n = w_n$ under (A4). Since $g_n - f_n \geq -\varepsilon_n$ for all $n \geq 1$, it follows that

$$\bar{g}_n - \bar{f}_n \geq -\bar{\varepsilon}_n \geq -w_n \sum_{k=1}^{n} \varepsilon_k. \tag{6.2}$$

Lemma 17 *Let $(\boldsymbol{\theta}_n)_{n \geq 1}$ be an output of Algorithm 1. Assume (A5). Then for all $n \geq 1$, the following hold:*

(i) *(Forward Monotonicity)* $\quad \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \geq \Psi_n(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|) - \Delta_n;$

(ii) *(Stability of Estimates I)* $\quad$ *Under case C1 in Theorem 2, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n);$*

(iii) *(Stability of Estimates II)* $\quad$ *Under case C2 in Theorem 2, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq w_n.$*

Proof First suppose Algorithm 2 is used for (3.1). By the definition of the optimality gap Δ_n in (A5),

$$\bar{g}_n(\boldsymbol{\theta}_n) + \Psi_n(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|) \leq \bar{g}_n(\boldsymbol{\theta}_{n-1}) + \Delta_n.$$

On the other hand, suppose Algorithm 3 is used for (3.1). Consider the computation of $\boldsymbol{\theta}_n$ in Algorithm 3. Let $J_1, \ldots, J_m \in \mathbb{J}$ denote the coordinate blocks used in order, and let $\boldsymbol{\theta}_n^{(1)}, \ldots, \boldsymbol{\theta}_n^{(m)} (= \boldsymbol{\theta}_n)$ denote the outputs of (3.4) after each block minimization in Algorithm 3. Denote $\boldsymbol{\theta}_n^{(0)} = \boldsymbol{\theta}_{n-1}$. Note that $\boldsymbol{\theta}_n^{(i)}$ is an approximate minimizer of the convex function $\boldsymbol{\theta} \mapsto \bar{g}_n(\boldsymbol{\theta})$ over the convex set $\boldsymbol{\Theta}^{J_i}$. Also note that $\boldsymbol{\theta}_n^{(i-1)} \in \boldsymbol{\Theta}^{J_i}$, by definition of the optimality gap $\Delta_n^{(i)}$ in (A5), for $1 \leq i \leq m$,

$$\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) + \Psi_n(\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|) \leq \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}) + \Psi_n(\|\boldsymbol{\theta}_n^{(i-1)} - \boldsymbol{\theta}_{n-1}\|) + \Delta_n^{(i)},$$

Summing over all $i = 1, \ldots, m$ then gives $\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \geq -\Delta_n$. Recall that in this case we take the regularizer $\Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$ equals zero if $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq w_n/m$ and ∞ otherwise.

Hence this shows **(i)**. Note that **(iii)** is trivial by the search radius restriction in Algorithm 1.

It remains to show **(ii)**. We show the assertion under zero surrogate optimality gap $\Delta_n \equiv 0$ (see (A5)) so that each $\boldsymbol{\theta}_n$ is the exact minimizer of the ρ-strongly convex function $\bar{g}_n$ over $\boldsymbol{\Theta}$. Indeed, by the second-order growth property (Lemma 37) and using L'-Lipschitz continuity of g_n (see Lemma 16), almost surely,

$$
\begin{aligned}
\frac{\rho}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &\leq \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \\
&= (1 - w_n)\left(\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_{n-1}(\boldsymbol{\theta}_n)\right) + w_n\left(g_n(\boldsymbol{\theta}_{n-1}) - g_n(\boldsymbol{\theta}_n)\right) \\
&\leq w_n\left(g_n(\boldsymbol{\theta}_{n-1}) - g_n(\boldsymbol{\theta}_n)\right) \leq w_n L'\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|.
\end{aligned}
$$

This shows $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$, as desired. The proof for the general nonzero surrogate optimality gap can be found in Appendix C (see Lemma 39). ∎

Proposition 18 *Let* $(\boldsymbol{\theta}_n)_{n\geq 1}$ *be an output of Algorithm 1. Assume (A4). Then for each* $n \geq 0$, *the following hold:*

(i) $\bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_n(\boldsymbol{\theta}_n) \leq w_{n+1}\left(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right) + w_n^2\left(\sum_{k=1}^{n+1} \varepsilon_k\right) + \sum_{i=1}^m \Delta_{n+1}^{(i)}.$

(ii) $0 \leq w_{n+1}\left(\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right) \leq w_{n+1}\left(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right) + \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) + w_n^2\left(\sum_{k=1}^{n+1} \varepsilon_k\right).$

Proof We begin by observing that

$$
\bar{g}_{n+1}(\boldsymbol{\theta}_n) = (1 - w_{n+1})\bar{g}_n(\boldsymbol{\theta}_n) + w_{n+1}\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) + w_{n+1}(g_{n+1}(\boldsymbol{\theta}_n) - \ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n))
$$

for all $t \geq 0$. Hence

$$
\begin{aligned}
\bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) &- \bar{g}_n(\boldsymbol{\theta}_n) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (6.3)\\
&= \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_{n+1}(\boldsymbol{\theta}_n) + \bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n) \\
&= \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_{n+1}(\boldsymbol{\theta}_n) + (1 - w_{n+1})\bar{g}_n(\boldsymbol{\theta}_n) + w_{n+1}\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n) \\
&= \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_{n+1}(\boldsymbol{\theta}_n) + w_{n+1}(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)) \\
&\qquad + w_{n+1}(\bar{f}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n)) + w_{n+1}(g_{n+1}(\boldsymbol{\theta}_n) - \ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n)).
\end{aligned}
$$

Now note that $\bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_{n+1}(\boldsymbol{\theta}_n) \leq \sum_{i=1}^m \Delta_{n+1}^{(i)}$ by Lemma 17 **(i)**. Also, recall that $\bar{g}_n - \bar{f}_n \geq -w_n \sum_{k=1}^n \varepsilon_k$ from (6.2) and $0 \leq g_{n+1}(\boldsymbol{\theta}_n) - \ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) \leq \varepsilon_{n+1}$ since g_{n+1} is a multi-convex ε_n-approximate surrogate of $f_{n+1} = \ell(\mathbf{x}_{n+1}, \cdot)$ at $\boldsymbol{\theta}_n$. Thus the inequalities in both **(i)** and **(ii)** follow using $w_{n+1} \leq w_n$. ∎

Lemma 19 *Let* $(a_n)_{n\geq 0}$ *and* $(b_n)_{n\geq 0}$ *be sequences of nonnegative real numbers such that* $\sum_{n=0}^\infty a_n b_n < \infty$. *Then the following hold.*

(i) $\min\limits_{1 \leq k \leq n} b_k \leq \dfrac{\sum_{k=0}^{\infty} a_k b_k}{\sum_{k=1}^{n} a_k} = O\left(\left(\sum_{k=1}^{n} a_k\right)^{-1}\right).$

(ii) *Further assume $\sum_{n=0}^{\infty} a_n = \infty$ and $|b_{n+1} - b_n| = O(a_n)$. Then $\lim_{n \to \infty} b_n = 0$.*

Proof **(i)** follows from noting that

$$\left(\sum_{k=1}^{n} a_k\right) \min_{1 \leq k \leq n} b_k \leq \sum_{k=1}^{n} a_k b_k \leq \sum_{k=1}^{\infty} a_k b_k < \infty.$$

The proof of **(ii)** is omitted and can be found in (Mairal, 2013b, Lem. A.5). ∎

7. Key lemmas and proofs of main results

In this section, we state all key lemmas without proofs and derive the main results (Theorems 2, 3, and 4) assuming them. The key lemmas stated in this section will be proved in the subsequent sections.

7.1 Key Lemmas

In this subsection, we state all key lemmas that are sufficient to derive the main results in this paper.

First, Lemma 20 states some general concentration inequalities of recursively defined functions similar to the empirical and the surrogate loss functions in (2.1) and (2.7). One may regard them as the classical Glivenko-Cantelli theorem for a general weighting scheme. We state the lemma in a self-contained manner so that it may be more convenient to be used for other purposes.

Lemma 20 (Uniform concentration of parameterized empirical observables) *Fix compact subsets $\mathcal{X} \subseteq \mathbb{R}^q$, $\Theta \subseteq \mathbb{R}^p$ and a bounded Borel measurable function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^r$. Let $(\mathbf{x}_n)_{n \geq 1}$ denote a sequence of points in $\mathcal{X}$ such that $\mathbf{x}_n = \varphi(X_n)$ for $n \geq 1$, where $(X_n)_{n \geq 1}$ is a Markov chain on a state space Ω and $\varphi : \Omega \to \mathcal{X}$ is a measurable function. Fix a sequence of weights $w_n \in (0, 1]$, $n \geq 0$ and define functions $\bar{\psi}(\cdot) := \mathbb{E}_{\mathbf{x} \sim \pi}[\psi(\mathbf{x}, \cdot)]$ and $\bar{\psi}_n : \Theta \to \mathbb{R}^r$ recursively as $\bar{\psi}_0 \equiv \mathbf{0}$ and*

$$\bar{\psi}_n(\cdot) = (1 - w_n)\bar{\psi}_{n-1}(\cdot) + w_n \psi(\mathbf{x}_n, \cdot).$$

Assume the following:

(a1) *The Markov chain $(X_n)_{n \geq 1}$ mixes exponentially fast to its unique stationary distribution and the stochastic process $(\mathbf{x}_n)_{n \geq 1}$ on $\mathcal{X}$ has a unique stationary distribution π.*

(a2) *w_n is non-increasing in n and $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all sufficiently large $n \geq 1$.*

Then there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \bar{\psi}(\boldsymbol{\theta}) - \mathbb{E}[\bar{\psi}_n(\boldsymbol{\theta})] \right\| \leq C w_n, \quad \mathbb{E} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta}) \right\| \right] \leq C w_n \sqrt{n}. \qquad (7.1)$$

Furthermore, if $w_n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$, then $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta}) \right\| \to 0$ as $t \to \infty$ almost surely.

We remark that a similar result for the i.i.d. case appeared in (Mairal, 2013b, Lem B.7), which states the second inequality in (7.1) as well as the last almost sure convergence statement under the stronger condition of $\sum_{n=1}^{\infty} w_n = \infty$ and $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$. We will refer to the latter condition as the 'strong square-summability' condition. These conditions were necessary in order to use Lemma 19 **(ii)** to deduce the almost sure convergence. We give a direct argument using Borel-Cantelli lemma without this additional condition, for which the weaker condition of $w_n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$ is sufficient.

Next, Lemma 21 states a series of finite variation statements that provide a basis for the forthcoming arguments. Most of the statements also appeared in the literature (see, e.g., Mairal et al. (2010); Mairal (2013b); Lyu et al. (2020, 2022)) in some special cases, where the strong square summability condition was used. We give an improved argument only using the square summability condition except for the last item.

Lemma 21 (Variation of functions) *Let $(\boldsymbol{\theta}_n)_{n \geq 1}$ be an output of Algorithm 1 under any of the three cases C1-C3 in Theorem 2. Suppose (A1)-(A4). The following hold:*

(i) $\displaystyle\sum_{n=1}^{\infty} \mathbb{E} \left[w_{n+1} \left(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \right) \right]^+ < \infty;$

(ii) $\displaystyle\sum_{n=0}^{\infty} \left(\mathbb{E} \left[\bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) - \bar{g}_n(\boldsymbol{\theta}_n) \right] \right)^+ < \infty;$

(iii) $\displaystyle\mathbb{E} \left[\sum_{n=1}^{\infty} w_n \left| \bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \right| \right] < \infty;$

(iv) $\displaystyle\sum_{n=0}^{\infty} w_{n+1} \left| \mathbb{E} \left[f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \right] \right| < \infty;$

(v) *Suppose the optional condition in (A4) holds. Then*

$$\mathbb{E} \left[\sum_{n=0}^{\infty} w_{n+1} \left| f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \right| \right] < \infty.$$

(vi) $\mathbb{E} \left[\sum_{n=1}^{\infty} \| \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \|^2 \right] < \infty.$

The following lemma is one of the key innovations in the present work, which allows us to obtain convergence rate bounds stated in Theorems 3 and 4. Roughly speaking, it gives gradient versions of the finite variation statements in Lemma 21.

Lemma 22 (Variation of gradients) *Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be the output of Algorithm 1 under any of the three cases C1-C3 in Theorem 2. Suppose (A1)-(A4). The following hold:*

(i) $\mathbb{E}\left[\sum_{n=1}^{\infty} w_n \left(\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|^2 + \left\|\mathbb{E}\left[\nabla \bar{f}_n(\boldsymbol{\theta}_n)\right] - \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_n)\right]\right\|^2\right)\right] < \infty.$

(ii) *If in addition the optional condition in (A4) holds, then*

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1} \left(\begin{array}{c}\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|^2 + \|\nabla \bar{f}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|^2 \\ +\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|^2\end{array}\right)\right] < \infty.$$

Based on Lemmas 20, 21, and 22, one can deduce Theorem 3 as well as Theorem 2 for case C1. A nice property of Algorithm 1 with convex surrogates g_n with identically zero regularization ($\Psi_n \equiv 0$ in (3.1)) and optimality gaps (see (A5)) is that the iterates $\boldsymbol{\theta}_n$ are exact minimizers of the convex averaged surrogates $\bar{g}_n$ over $\boldsymbol{\Theta}$. Hence $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ lies in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$ for each $n \geq 1$. Thus, in order for asymptotic stationarity with respect to the empirical loss $\bar{f}_n$ or the expected loss f, one only needs to show that every convergent subsequence of $\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla h_n(\boldsymbol{\theta}_n)\|$ vanish almost surely, where $h_n \in \{\bar{f}_n, f\}$ for $n \geq 1$.

However, when $\bar{g}_n$ is (approximately) minimized with proximal regularization (Algorithm 2) or within a diminishing radius (Algorithm 3), $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ may be close to but not within the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$. For instance, even though each convex sub-problem in Algorithm 3 is exactly solved, the technique of radius restriction in Algorithm 3 introduces additional radius constraints so that $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ may be normal to the trust region boundary, which is the sphere of distance w_n/m centered at $\boldsymbol{\theta}_{n-1}$; Similarly, using proximal regularization in Algorithm 2 tilts the surrogate gradient. In order to handle these issues, in Lemma 23, we will show that the sequence $\boldsymbol{\theta}_n$, $n \geq 1$ still verifies stationarity for $\bar{g}_n$ in an asymptotic sense. For this, we need to take 'convergent subsequences' of the averaged surrogate functions $\bar{g}_n$, which can be easily done under the compact parameterization assumption in (A7).

Lemma 23 (Asymptotic Surrogate Stationarity) *Assume (A1)-(A4) and (A7) and let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1 under any of the three cases C1-C3 in Theorem 2. Let $(t_k)_{k\geq 1}$ be any sequence such that $\boldsymbol{\theta}_{t_k}$ and $\bar{g}_{t_k}$ converges almost surely. Then $\boldsymbol{\theta}_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{t_k}$ is almost surely a stationary point of $\bar{g}_\infty := \lim_{k\to\infty} \bar{g}_{t_k}$ over $\boldsymbol{\Theta}$.*

Finally, Lemma 24 is the last ingredient for proving Theorem 4, which states that the optimality for the averaged surrogates $\bar{g}_n$ is summable with weights w_n.

Lemma 24 (Finite variation of surrogate optimality) *Let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1 under any of the three cases C1-C3 in Theorem 2. Assume (A1)-(A4). Then for each initialization $\boldsymbol{\theta}_0$, we have*

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1} \left|\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle\right|\right] < \infty.$$

7.2 Proof of Theorem 3

In this subsection, we will assume Lemmas 20, 21, and 22 and prove Theorem 3.

Proof of Theorem 3. Suppose (A1)-(A4) and any of the cases C1-C3 stated in Theorem 2. By Lemma 21 **(iii)** and Lemma 22 **(i)**, we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left(\left|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right| + \|\nabla\bar{g}_n(\boldsymbol{\theta}_n) - \nabla\bar{f}_n(\boldsymbol{\theta}_n)\|^2\right)\right] < \infty. \tag{7.2}$$

Recall that $\mathbb{E}[|X|] < \infty$ implies $|X| < \infty$ almost surely for any random variable X. It follows that the summation above is almost surely finite. Then (4.7) follows from Lemma 19.

Next, we show (4.9). First, we deduce some intermediate bounds. By Lemma 21 **(iii)**-**(iv)** and Jensen's inequality, we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_n\left|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[f(\boldsymbol{\theta}_n)]\right|\right] \tag{7.3}$$

$$\leq \mathbb{E}\left[\sum_{n=1}^{\infty} w_n\left(\mathbb{E}\left[\left|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right|\right] + \left|\mathbb{E}[\bar{f}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[f(\boldsymbol{\theta}_n)]\right|\right)\right] < \infty.$$

Similarly, using Lemma 22 **(i)**, we can deduce

$$\sum_{n=1}^{\infty} w_n \left\|\mathbb{E}\left[\nabla\bar{g}_n(\boldsymbol{\theta}_n)\right] - \mathbb{E}\left[\nabla f(\boldsymbol{\theta}_n)\right]\right\|^2 < \infty. \tag{7.4}$$

Now recall that by Lemma 20, there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|f(\boldsymbol{\theta}) - \mathbb{E}[\bar{f}_n(\boldsymbol{\theta})]\right| \leq Cw_n, \qquad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\nabla f(\boldsymbol{\theta}) - \mathbb{E}[\nabla\bar{f}_n(\boldsymbol{\theta})]\right\| \leq Cw_n. \tag{7.5}$$

Using (7.3) and the first bound in (7.5), we get

$$|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - f(\boldsymbol{\theta}_n)|^2 \leq \left(|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[\bar{f}_n(\boldsymbol{\theta}_n)]| + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathbb{E}[\bar{f}_n(\boldsymbol{\theta})] - f(\boldsymbol{\theta})|\right)^2$$

$$\leq \left|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[\bar{f}_n(\boldsymbol{\theta}_n)]\right|^2 + Cw_{n+1}\left|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[\bar{f}_n(\boldsymbol{\theta}_n)]\right| + C^2 w_{n+1}^2.$$

Recalling that $\bar{g}_n$ and $\bar{f}_n$ are uniformly bounded over $\boldsymbol{\Theta}$ by (A1), using (7.3) we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - f(\boldsymbol{\theta}_n)\right|^2\right] < \infty. \tag{7.6}$$

An identical argument using (7.4) as well as the second bound in (7.5) shows

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left\|\mathbb{E}[\nabla\bar{g}_n(\boldsymbol{\theta}_n)] - \nabla f(\boldsymbol{\theta}_n)\right\|^2\right] < \infty. \tag{7.7}$$

Combining the above two bounds and using Lemma 19 give (4.9).

It remains to show (4.8) and (4.10). Suppose the optional condition in (A4) holds. Then by Lemma 21 **(v)** and Lemma 22 **(ii)**, we deduce

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left(|\bar{g}_n(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_n)| + \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|^2\right)\right] < \infty. \tag{7.8}$$

Then (4.10) follows from Lemma 19.

Lastly, recall that by Lemma 20 and see (A4), there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta} |f(\boldsymbol{\theta}) - \bar{f}_n(\boldsymbol{\theta})|\right] \leq C w_n \sqrt{n} \leq 2C, \quad \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla f(\boldsymbol{\theta}) - \nabla \bar{f}_n(\boldsymbol{\theta})\|\right] \leq C w_n \sqrt{n} \leq 2C. \tag{7.9}$$

Then note that

$$\mathbb{E}\left[|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \bar{f}_n(\boldsymbol{\theta}_n)|^2\right] \leq \mathbb{E}\left[\left(|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - f(\boldsymbol{\theta}_n)| + \sup_{\boldsymbol{\theta}\in\Theta} |\bar{f}_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})|\right)^2\right]$$

$$\leq \mathbb{E}\left[|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - f(\boldsymbol{\theta}_n)|^2 + 2C\,|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \mathbb{E}[f(\boldsymbol{\theta}_n)]|\right] + C^2(n+1)w_{n+1}^2.$$

Then multiply w_n on both sides sum over all $n \geq 1$. The first two terms in the resulting infinite sum in the righthand side are finite by (7.6) and (7.3). For the last term, note that since $w_n\sqrt{n} = o(1)$ (see (A4)) and using the optional condition in (A4),

$$\sum_{n=1}^{\infty}(n+1)w_{n+1}^3 = \sum_{n=1}^{\infty}\left(w_{n+1}\sqrt{n+1}\right)w_{n+1}^2\sqrt{n+1} \leq C'\sum_{n=1}^{\infty}w_{n+1}^2\sqrt{n+1} < \infty$$

for some constant $C' > 0$. It follows that

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left|\mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] - \bar{f}_n(\boldsymbol{\theta}_n)\right|^2\right] < \infty.$$

An identical argument using the second bound in (7.9) as well as (7.4) shows

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left\|\mathbb{E}[\nabla\bar{g}_n(\boldsymbol{\theta}_n)] - \nabla\bar{f}_n(\boldsymbol{\theta}_n)\right\|^2\right] < \infty.$$

Combining the above two bounds gives (4.8). This completes the proof. ∎

7.3 Proof of Theorem 2

Next, we prove Theorem 2 assuming Lemmas 20, 21, 22, and 23.

Proof of Theorem 2. Suppose (A1)-(A4). First assume cases C1 or C2 in Theorem 2. We will first show the assertion except for the asymptotic stationarity statements.

We first show the first part of **(i)**. By Lemma 21 **(iii)** and Lemma 22 **(i)**, we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_{n+1}\left(\left|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right| + \|\nabla\bar{g}_n(\boldsymbol{\theta}_n) - \nabla\bar{f}_n(\boldsymbol{\theta}_n)\|^2\right)\right] < \infty. \tag{7.10}$$

Denote $h_n^{(1)} := \bar{g}_n - \bar{f}_n$ and $h_n^{(2)} := \|\nabla \bar{g}_n - \nabla \bar{f}_n\|^2$. We claim that $|h_{n+1}^{(i)}(\boldsymbol{\theta}_{n+1}) - h_n^{(i)}(\boldsymbol{\theta}_n)| = O(w_{n+1})$ for $i = 1, 2$. Then by Lemma 19, it holds that $|h_n^{(i)}(\boldsymbol{\theta}_n)| \to 0$ almost surely as $n \to \infty$ for $i = 1, 2$.

To show the claim, recall the recursive definitions of $\bar{g}_n$ and $\bar{f}_n$ and that $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is R-Lipschitz continuous by (A1) for each $\mathbf{x} \in \mathfrak{X}$. Also, $h_n^{(1)}(\boldsymbol{\theta}) = (1 - w_n)h_{n-1}^{(1)}(\boldsymbol{\theta}) + w_n(g_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}))$, where $f_n(\cdot) = \ell(\mathbf{x}_n, \cdot)$ and $g_n \in \text{Srg}_{L,\rho}^{\mathbb{L}}(f_n, \boldsymbol{\theta}_{n-1})$ (see Definition 1) and $\nabla(g_n - f_n)$ is L-Lipschitz continuous and has norm $\leq \varepsilon_n$ at $\boldsymbol{\theta}_{n-1}$. Since the parameter space $\boldsymbol{\Theta}$ is compact by (A3), it follws that $\nabla(g_n - f_n)$ is uniformly bounded by some constant $L' > 0$ over $\boldsymbol{\Theta}$. Then $g_n - f_n$ is L'-Lipschitz, and by induction, $h_n^{(1)}$ is also L-Lipschitz for $n \geq 1$. Now note that

$$
\begin{aligned}
|h_{n+1}^{(1)}(\boldsymbol{\theta}_{n+1}) - h_n^{(1)}(\boldsymbol{\theta}_n)| &\leq |h_{n+1}^{(1)}(\boldsymbol{\theta}_{n+1}) - h_{n+1}^{(1)}(\boldsymbol{\theta}_n)| + |h_{n+1}^{(1)}(\boldsymbol{\theta}_n) - h_n^{(1)}(\boldsymbol{\theta}_n)| \\
&\leq L'\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| + \left|(\bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n)) - (\bar{f}_{n+1}(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n))\right| \\
\\
&\leq L'w_n + w_{n+1}\left|g_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_{n+1}) - f_{n+1}(\boldsymbol{\theta}_n) + \bar{f}_n(\boldsymbol{\theta}_n)\right| \\
&= L'\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| + w_{n+1}\left|\bar{g}_n(\boldsymbol{\theta}_{n+1}) - \bar{f}_n(\boldsymbol{\theta}_n)\right|,
\end{aligned}
$$

where the third inequality uses Lemma 17 (iii) as well as the recursive definitions of $\bar{g}_{n+1}$ and $\bar{f}_{n+1}$. Then note that $\ell : \mathcal{X} \times \boldsymbol{\Theta} \to \mathbb{R}$ is bounded by (A1), so $\bar{g}_n$ and $\bar{f}_n$ are also uniformly bounded in n. This shows $|h_{n+1}^{(1)}(\boldsymbol{\theta}_{n+1}) - h_n^{(1)}(\boldsymbol{\theta}_n)| = O(w_{n+1})$.

Next, we verify $|h_n^{(2)}(\boldsymbol{\theta}_n) - h_n^{(2)}(\boldsymbol{\theta}_{n-1})| = O(w_{n+1})$. Note that by linearity of gradients, $\nabla \bar{g}_n$ and $\nabla \bar{f}_n$ satisfy the same recursion as $\bar{g}_n$ and $\bar{f}_n$. Moreover, $\nabla h_n^{(2)}(\boldsymbol{\theta}) = (1 - w_n)\nabla h_{n-1}^{(2)}(\boldsymbol{\theta}) + w_n \nabla(g_n - f_n)(\boldsymbol{\theta})$ and since $\nabla(g_n - f_n)$ is L-Lipschitz, so is $\nabla h_n^{(2)}$ for all $n \geq 1$. Then by using a similar argument as above,

$$
\begin{aligned}
\left|\sqrt{h_{n+1}^{(2)}(\boldsymbol{\theta}_{n+1})} - \sqrt{h_n^{(2)}(\boldsymbol{\theta}_n)}\right| &\\
&= \|\nabla h_{n+1}^{(2)}(\boldsymbol{\theta}_{n+1}) - \nabla h_{n+1}^{(2)}(\boldsymbol{\theta}_n)\| + \|\nabla h_{n+1}^{(2)}(\boldsymbol{\theta}_n) - \nabla h_n^{(2)}(\boldsymbol{\theta}_n)\| \\
&\leq L\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| + w_{n+1}\|\nabla(g_{n+1} - f_{n+1})(\boldsymbol{\theta}_n) + \nabla\bar{f}_n(\boldsymbol{\theta}_n) - \nabla\bar{g}_n(\boldsymbol{\theta}_n)\| \\
&\leq Lw_{n+1} + w_{n+1}\|\nabla h_n^{(2)}(\boldsymbol{\theta}_n)\|.
\end{aligned}
$$

But recall that $\nabla h_n^{(2)}$ is uniformly bounded by L' over $\boldsymbol{\Theta}$. To conclude, write

$$
|h_n^{(2)}(\boldsymbol{\theta}_n) - h_n^{(2)}(\boldsymbol{\theta}_{n-1})| = \left|\sqrt{h_n^{(2)}(\boldsymbol{\theta}_n)} - \sqrt{h_{n-1}^{(2)}(\boldsymbol{\theta}_{n-1})}\right| \cdot \left|\sqrt{h_n^{(2)}(\boldsymbol{\theta}_n)} + \sqrt{h_{n-1}^{(2)}(\boldsymbol{\theta}_{n-1})}\right|.
$$

Noting that $h_n^{(2)}$ is uniformly bounded for all $n \geq 1$, we conclude that $|h_n^{(2)}(\boldsymbol{\theta}_n) - h_n^{(2)}(\boldsymbol{\theta}_{n-1})| = O(w_{n+1})$, as desired.

Next, we show the first part of (ii). Recall that under (A4) (without the optional condition), by Lemma 20 (se also Lemma 42), we have $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\bar{f}_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \to 0$ and $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla\bar{f}_n(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\| \to 0$ almost surely as $n \to \infty$. Then by (i), we have

$$
|\bar{g}_n(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_n)| \leq |\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)| + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\bar{f}_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \to 0 \quad \text{a.s. as } n \to \infty. \tag{7.11}
$$

Similarly, by triangle inequality and **(i)**,

$$\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\| \tag{7.12}$$
$$\leq \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\| + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla \bar{f}_n(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\| \to 0 \quad \text{a.s. as } n \to \infty.$$

Now we show the second parts of **(i)** and **(ii)**, the asymptotic stationarity. Assume case C1. Then each $\boldsymbol{\theta}_n$ is an exact minimizer of $\bar{g}_n$ over $\boldsymbol{\Theta}$, so $-\nabla \bar{g}_n(\boldsymbol{\theta}_n)$ is in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$ for each $n \geq 1$. Since we have shown that $\|\nabla \bar{f}_n(\boldsymbol{\theta}_n) - \nabla \bar{g}_n(\boldsymbol{\theta}_n)\|$ and $\|\nabla f(\boldsymbol{\theta}_n) - \nabla \bar{g}_n(\boldsymbol{\theta}_n)\|$ both converges to zero almost surely as $n \to \infty$, it follows that both $-\nabla \bar{f}_n(\boldsymbol{\theta}_n)$ and $-\nabla f(\boldsymbol{\theta}_n)$ belong to the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_n$ asymptotically almost surely as $n \to \infty$.

Assume case C2. Further assume (A7). Let $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$ be an arbitrary limit point of the sequence $(\boldsymbol{\theta}_t)_{t \geq 1}$ and let $(t_k)_{k \geq 1}$ be a (random) sequence such that $\boldsymbol{\theta}_{t_k} \to \boldsymbol{\theta}_\infty$ almost surely as $n \to \infty$. Since $\boldsymbol{\Theta} \times \mathcal{K}$ is compact, we may choose a further sequence of $(t_k)_{k \geq 1}$, which we will denote the same, so that κ_{t_k} converges to some element $\kappa_\infty \in \mathcal{K}$. Hence $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{t_k}$ is well-defined almost surely. It is important to note that $\boldsymbol{\theta}_\infty$ is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$ by Lemma 23. Hence $-\nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty)$ is in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}_\infty$. But since we have shown that $h^{(i)} \to 0$ almost suresly as $n \to \infty$ for $i = 2, 4$, we must have

$$\lim_{k \to \infty} \nabla \bar{f}_{t_k}(\boldsymbol{\theta}_\infty) = \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty) = \nabla f(\boldsymbol{\theta}_\infty).$$

Since $\boldsymbol{\theta}_\infty$ was an arbitrary limit point of $(\boldsymbol{\theta}_n)_{n \geq 1}$, this completes the proof of **(i)**-**(ii)**.

Lastly, assume C3 in Theorem 2. In this case, we do not have iterate stability $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ so we cannot use Lemma 19 **(ii)** to deduce the whole sequence convergence as we did before. However, we can still deduce subsequential convergence. Indeed, by Lemma 21 **(iii)** and Lemma 22 **(i)**, we have (7.10). Then by Lemma 19 **(i)** and noting that $\sum_{n=1}^{\infty} w_n = \infty$, we conclude that there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k \geq 1}$ of $(\boldsymbol{\theta}_n)_{n \geq 1}$ such that almost surely, $\left|\bar{g}_{n_k}(\boldsymbol{\theta}_{n_k}) - \bar{f}_{n_k}(\boldsymbol{\theta}_{n_k})\right| + \|\nabla \bar{g}_n(\boldsymbol{\theta}_{n_k}) - \nabla \bar{f}_{n_k}(\boldsymbol{\theta}_{n_k})\|^2 \to 0$ as $k \to \infty$. Also, using a similar argument as before in (7.11) and (7.12), we can deduce that almost surely along the same subsequence, $|\bar{g}_{n_k}(\boldsymbol{\theta}_{n_k}) - f(\boldsymbol{\theta}_{n_k})| + \|\nabla \bar{g}_n(\boldsymbol{\theta}_{n_k}) - \nabla f(\boldsymbol{\theta}_{n_k})\|^2 \to 0$ as $k \to \infty$. Lastly, asymptotic stationarity for $\bar{g}_n$ is given by Lemma 23. ∎

Remark 25 Without appealing to Lemma 20 and using the convergence results for the empirical loss minimization stated in Theorem 2, one can directly prove Theorem 2 **(ii)** using a similar argument as in the proof of **(i)**. This amounts to use the finite sum (7.8) in place of (7.2) and showing the bounds $|h_{n+1}^{(i)}(\boldsymbol{\theta}_{n+1}) - h_n^{(i)}(\boldsymbol{\theta}_n)| = O(w_{n+1})$ for $i = 3, 4$, where $h^{(3)} = \bar{g}_{n+1} - f$ and $h_n^{(4)} = \|\nabla \bar{g}_n - \nabla f\|^2$. However, (7.8) holds under the additional optional condition in (A4), which essentially says $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$. While this condition is standard in the literature (see, e.g., (Mairal, 2013b, (E))), (Lyu et al., 2020, (M2)), and (Lyu et al., 2022, (A3)), our proof above avoids it and only relies on the more standard square summability condition $\sum_{n=1}^{\infty}(\log n)w^2 < \infty$ (see (A4)). However, the stronger condition $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$ as well as the finite sum (7.8) is crucial in our proof of Theorem 4.

7.4 Proof of Theorem 4

Next, we prove Theorem 4 assuming Lemmas 21, 22, and 24.

Proof of Theorem 4. We first show **(i)**. Note that (4.11) follows immediately from Lemmas 24 and 19. Next, we show (4.12). By Cauchy-Schwarz inequality, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\left| \langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle - \langle \nabla \bar{f}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \leq \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|.$$

It follows that for all $n \geq 1$,

$$\left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{f}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \leq \left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \quad (7.13)$$
$$+ \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|.$$

On the other hand, by Lemmas 24 and 22 **(i)**, we have

$$\sum_{n=1}^{\infty} w_{n+1} \left[\left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| + \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|^2 \right] < \infty.$$

Then by Lemma 19, we have

$$\min_{1 \leq k \leq n} \left[\left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| \leq 1} \langle -\nabla \bar{g}_k(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle \right| + \|\nabla \bar{g}_k(\boldsymbol{\theta}_k) - \nabla \bar{f}_k(\boldsymbol{\theta}_k)\|^2 \right]$$
$$= O\left(\left(\sum_{k=1}^{n} w_k \right)^{-1} \right).$$

Let $t_n \in \{1, \ldots, n\}$ for $n \geq 1$ be such that the minimum above is achieved. Namely, denoting the term in the minimum above by A_k, we have $A_{t_n} = O\left(\left(\sum_{k=1}^{n} w_k \right)^{-1} \right)$. Since all terms in A_k are nonnegative, it follows that there exists a constant $c_1, c_2 > 0$ such that for all $n \geq 1$, alsmot surely,

$$\left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t_n}\| \leq 1} \langle -\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t_n} \rangle \right| \leq \frac{c_1}{\sum_{k=1}^{n} w_k},$$
$$\|\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n}) - \nabla f(\boldsymbol{\theta}_{t_n})\| \leq \frac{c_2}{\sqrt{\sum_{k=1}^{n} w_k}}.$$

Hence from (7.13), it follows that there exists some constant $c_3 > 0$ such that for all $n \geq 1$,

$$\left| \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t_n}\| \leq 1} \langle -\nabla \bar{f}_{t_n}(\boldsymbol{\theta}_{t_n}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t_n} \rangle \right| \leq \frac{c_3}{\sqrt{\sum_{k=1}^{n} w_k}}.$$

This completes the proof of **(i)**.

Next, we show (4.13) in **(ii)**. Recall that by Lemma 20 and see (A4), there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla f(\boldsymbol{\theta}) - \nabla \bar{f}_n(\boldsymbol{\theta})\| \right] \leq C w_n \sqrt{n}. \quad (7.14)$$

Then first observe that by Cauchy-Schwarz and triangle inequalities,

$$|\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle - \langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle|$$
$$\leq \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|$$
$$\leq \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\| + \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla \bar{f}_n(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|.$$

It follows that for all $n \geq 1$,

$$\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \leq \left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \quad (7.15)$$
$$+ \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|$$
$$+ \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla \bar{f}_n(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|.$$

Note that by Lemmas 24, 22 **(i)**, we have

$$\sum_{n=1}^{\infty} w_n \left(\mathbb{E} \left[\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \right] + \mathbb{E} \left[\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)\|^2 \right] \right) < \infty.$$

By Lemma 19, similarly as before, we can take a sequence $t_n \in \{1, \ldots, n\}$ for $n \geq 1$ such that for some constants $c_3, c_4 > 0$ and for all $n \geq 1$,

$$\mathbb{E} \left[\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t_n}\| \leq 1} \langle -\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t_n} \rangle \right| \right] \leq \frac{c_3}{\sum_{k=1}^{n} w_k},$$
$$\mathbb{E} \left[\|\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n}) - \nabla f(\boldsymbol{\theta}_{t_n})\| \right] \leq \frac{c_4}{\sqrt{\sum_{k=1}^{n} w_k}}.$$

Hence taking expectation on (7.15) and using this sequence t_n with (7.14), we get

$$\mathbb{E} \left[\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t_n}\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}_{t_n}), \boldsymbol{\theta} - \boldsymbol{\theta}_{t_n} \rangle \right| \right] \leq \frac{c_5}{\sqrt{\sum_{k=1}^{n} w_k}} + w_n \sqrt{n}.$$

for all $n \geq 1$ for some constant $c_5 > 0$. This shows (4.13).

It remains to show (4.14) in **(ii)**. Assume the optional condition in (A4) holds. Using a similar argument as before, for all $n \geq 1$,

$$\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| \leq \left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right|$$
$$+ \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|.$$

Also, by Lemmas 24 and 22 **(ii)**, almost surely,

$$\sum_{n=1}^{\infty} w_{n+1} \left[\left| \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right| + \|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)\|^2 \right] < \infty.$$

The rest of the argument is identical to the proof of **(i)**. ∎

46

7.5 Proofs of Corollaries 5 and 6

Lastly in this section, we prove Corollaries 5 and 6.

Proof of Corollary 5. Suppose that there exists $\varepsilon > 0$ such that the iterates $\boldsymbol{\theta}_n$ are all in the ε-interior of $\boldsymbol{\Theta}$, that is, the open ε-ball centered at $\boldsymbol{\theta}_n$ lies in $\boldsymbol{\Theta}$, for all $n \geq 1$. Then in Theorem 4 (4.11), we may choose $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in a way that $\boldsymbol{\theta} - \boldsymbol{\theta}_n = \varepsilon \nabla \bar{g}_n(\boldsymbol{\theta}_n)$. Then (4.11) reduces to the first bound in (4.15). The same argument shows the second bound in (4.15). For the last bound in (4.15), we use (7.7) we derived in the proof of Theorem 3. Combining it with Lemma 22 and using Fubini's theorem, we get

$$\sum_{n=1}^{\infty} w_{n+1} \left(\mathbb{E} \left[\|\nabla \bar{g}_n(\boldsymbol{\theta}_n)\| \right] + \left\| \mathbb{E}[\nabla \bar{g}_n(\boldsymbol{\theta}_n)] - \nabla \bar{f}_n(\boldsymbol{\theta}_n) \right\|^2 \right) < \infty.$$

Then by Lemma 19, we have

$$\min_{1 \leq k \leq n} \left[\mathbb{E} \left[\|\nabla \bar{g}_n(\boldsymbol{\theta}_n)\| \right] + \|\nabla \mathbb{E}[\bar{g}_k(\boldsymbol{\theta}_k)] - \nabla f(\boldsymbol{\theta}_k)\|^2 \right] = O \left(\left(\sum_{k=1}^{n} w_k \right)^{-1} \right)$$

Choose a subsequence t_n along which the above minimum is achieved for all $n \geq 1$. Then we have

$$\mathbb{E} \left[\|\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n})\| \right] = O \left(\left(\sum_{k=1}^{t_n} w_k \right)^{-1} \right),$$

$$\|\nabla \mathbb{E}[\bar{g}_{t_n}(\boldsymbol{\theta}_{t_n})] - \nabla f(\boldsymbol{\theta}_{t_n})\| = O \left(\left(\sum_{k=1}^{t_n} w_k \right)^{-1/2} \right).$$

Therefore by using Jensen's inequality and the above bound, we deduce

$$\|\nabla f(\boldsymbol{\theta}_{t_n})\| \leq \mathbb{E} \left[\|\nabla \bar{g}_{t_n}(\boldsymbol{\theta}_{t_n})\| \right] + \|\nabla \mathbb{E}[\bar{g}_{t_n}(\boldsymbol{\theta}_{t_n})] - \nabla f(\boldsymbol{\theta}_{t_n})\| = O \left(\left(\sum_{k=1}^{t_n} w_k \right)^{-1/2} \right).$$

This completes the proof. ∎

Proof of Corollary 6. First, suppose $w_n = n^{-1/2} (\log n)^{\delta}$ for some $\delta > 1$. Then the upper bound on the rate of convergence in Theorem 4 is of order $O(n^{-1/2}(\log n)^{\delta})$. Then one can conclude by using the fact that $n \geq \varepsilon^{-2}(3 \log \varepsilon^{-1})^{2\delta}$ implies $n^{-1/2}(\log n)^{\delta} \leq \varepsilon$ for all sufficiently small $\varepsilon > 0$. Indeed, assuming $n \geq \varepsilon^{-2}(3 \log \varepsilon^{-1})^{2\delta}$,

$$n^{-1/2}(\log n)^{\delta} \leq \frac{\varepsilon}{(3 \log \varepsilon^{-1})^{\delta}} (2 \log \varepsilon^{-1} + 2\delta \log(3 \log \varepsilon^{-1}))^{\delta}$$

and the last expression is at most ε for all sufficiently small $\varepsilon > 0$. This shows **(i)**. A similar argument using Theorem 4 (ii) as well as Corollary 5 shows **(ii)**. ∎

8. Proof of Lemmas 20, 21, and 22

Recall that for each $n \geq 0$, $\mathcal{F}_n$ denotes the σ-algebra generated by the history of underlying Markov chain $Y_0, Y_1, \ldots, Y_n$ as well as the possible randomness in Algorithm 1 up to iteration n. Note that $\mathbf{x}_n = \varphi(Y_n)$ for $n \geq 1$ by (A2).

We first prove Lemma 20 below. Our proof uses an auxiliary lemma, Lemma 42 in the appendix.

Proof of Lemma 20. Recall that $\mathbf{x}_k = \varphi(Y_k)$ under (A2). Let π_k denote the distribution of Y_k. Let $M := \sup_{\mathbf{x}, \boldsymbol{\theta}} \|\psi(\mathbf{x}, \boldsymbol{\theta})\| < \infty$. Note that, by a change of measure,

$$
\begin{aligned}
\mathbb{E}\left[\psi(\mathbf{x}_n, \boldsymbol{\theta})\right] &= \sum_{\mathbf{y} \in \Omega} \psi(\varphi(\mathbf{y}), \boldsymbol{\theta}) \, P^n(Y_0, \mathbf{y}) \\
&= \sum_{\mathbf{y} \in \Omega} \psi(\varphi(\mathbf{y}), \boldsymbol{\theta}) \, \pi(\mathbf{y}) + \sum_{\mathbf{y} \in \Omega} \psi(\varphi(Y_0), \boldsymbol{\theta})(P^n(Y_0, \mathbf{y}) - \pi(\mathbf{y})) \\
&= \bar{\psi}(\boldsymbol{\theta}) + \sum_{\mathbf{y} \in \Omega} \psi(\varphi(Y_0), \boldsymbol{\theta}) \, (P^n(Y_0, \mathbf{y}) - \pi(\mathbf{y})).
\end{aligned}
$$

By the triangle inequality, it follows that

$$
\begin{aligned}
\left\|\mathbb{E}\left[\psi(\mathbf{x}_n, \boldsymbol{\theta})\right] - \bar{\psi}(\boldsymbol{\theta})\right\| &\leq \sum_{\mathbf{y} \in \Omega} \|\psi(\varphi(Y_0), \boldsymbol{\theta})\| \, |P^n(Y_0, \mathbf{y}) - \pi(\mathbf{y})| \\
&\leq M \sum_{y \in \Omega} |P^n(Y_0, \mathbf{y}) - \pi(\mathbf{y})| \\
&\leq 2M \|P^n(Y_0, \cdot) - \pi\|_{TV},
\end{aligned}
$$

where the last inequality follows from the relation $2\|\mu - \nu\|_{TV} = \sum_x |\mu(x) - \nu(x)|$ where μ, ν are probability distributions on the same sample space (see (Levin and Peres, 2017, Prop. 4.2)).

Now recall that under the hypothesis in (A4), $w_k^n = w_k \prod_{i=k+1}^n (1 - w_i)$ is non-decreasing in $k \in \{1, \ldots, n\}$. So, $w_1^n \leq \cdots \leq w_n^n = w_n$. Then using (A2), we have

$$
\begin{aligned}
\left\|\bar{\psi}(\boldsymbol{\theta}) - \mathbb{E}\left[\bar{\psi}_n(\mathbf{x}, \boldsymbol{\theta})\right]\right\| &\leq \left\|\sum_{k=1}^n \left(\bar{\psi}(\boldsymbol{\theta}) - \mathbb{E}[\psi(\mathbf{x}_k, \boldsymbol{\theta})]\right) w_k^n\right\| \\
&\leq \sum_{k=1}^n \left\|\bar{\psi}(\boldsymbol{\theta}) - \mathbb{E}[\psi(\mathbf{x}_k, \boldsymbol{\theta})]\right\| w_k^n \\
&\leq 2M \sum_{k=1}^n \|P^k(Y_0, \cdot) - \pi\|_{TV} w_k^n \\
&\leq 2M w_n \sum_{k=1}^n \sup_{\mathbf{y} \in \mathfrak{X}} \|P^k(\mathbf{y}, \cdot) - \pi\|_{TV} \leq \frac{2L\lambda w_n}{1 - \lambda}.
\end{aligned}
$$

This shows the first bound in (7.1).

Finally, for $d = 1$, the second inequality in (7.1), as well as the last part of the statement, are direct consequences of Lemma 42, nothing that $w_1^n \leq \cdots \leq w_n^n = w_n$. Applying this

to each of the d coordinates of ψ and applying triangle inequality will imply the assertion. Namely, write $\psi = (\psi^{(1)}, \ldots, \psi^{(d)})^T$. Then noting that $\psi, \bar{\psi}_n$ are uniformly bounded by some constant $M > 0$, we have

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta}) \right\|^2 \leq \sum_{i=1}^{d} \sup_{\boldsymbol{\theta} \in \Theta} \left| \bar{\psi}^{(i)}(\boldsymbol{\theta}) - \bar{\psi}_n^{(i)}(\boldsymbol{\theta}) \right|^2 \leq 2M \sum_{i=1}^{d} \sup_{\boldsymbol{\theta} \in \Theta} \left| \bar{\psi}^{(i)}(\boldsymbol{\theta}) - \bar{\psi}_n^{(i)}(\boldsymbol{\theta}) \right|.$$

Note that by Lemma 42, the expectation of each summand in the last expression is of order $O(w_n \sqrt{n})$. Hence the second inequality in (7.1) follows. Lastly, the summands in the last expression above converge to zero almost surely if $w_n^n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$ by Lemma 42. Note that $w_n^n = w_n$. Hence the left-hand side above converges to zero almost surely $w_n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$. This completes the proof. ∎

To handle the issue of dependence in signals, we adopt the strategy developed in Lyu et al. (2020) in order to handle a similar issue for vector-valued signals (or matrix factorization). The key insight in Lyu et al. (2020) is that, while the 1-step conditional distribution $P(Y_{t-1}, \cdot)$ may be far from the stationary distribution π, the N-step conditional distribution $P^N(Y_{t-N}, \cdot)$ is exponentially close to π under mild conditions. Hence we can condition much early on – at time $t - N$ for some suitable $N = N(t)$. Then the Markov chain runs $N + 1$ steps up to time $t + 1$, so if N is large enough for the chain to mix to its stationary distribution π, then the distribution of Y_{t+1} conditional on $\mathcal{F}_{t-N}$ is close to π. The error of approximating the stationary distribution by the $N + 1$ step distribution can be controlled using total variation distance and Markov chain mixing bound. We refine the analysis for (Lyu et al., 2020, Lem. 12) and obtain a more improved version below.

Proposition 26 *Assumptions (A1)-(A4) hold. Let $(\boldsymbol{\theta}_n)_{n \geq 1}$ be an output of Algorithm and denote $L = \|\ell(\cdot, \cdot)\|_\infty$. Then there exists a constant $C > 0$ such that for all $0 \leq N \leq n$, it holds that*

$$\mathbb{E}\left[\mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_{n-N}) - \bar{f}_n(\boldsymbol{\theta}_{n-N}) \,\Big|\, \mathcal{F}_{n-N} \right]^+ \right] \leq C w_{n-N} + 2L\lambda^{N+1},$$

where $\lambda \in [0, 1)$ is the exponential mixing rate of the underlying Markov chain in (A2). Furthermore, if $\mathbf{x}_n$'s are i.i.d. from π, then the assertion holds with $N = 0$ and $\lambda = 0$.

Proof Fix $\mathbf{y} \in \Omega$ and suppose $Y_{n-N} = \mathbf{y}$. By the Markov property, the distribution of Y_{n+1} conditional on $\mathcal{F}_{n-N}$ equals $P^{N+1}(Y_{n-N}, \cdot)$, where P denotes the transition kernel of the chain $(Y_n)_{n \in \mathbb{N}}$. Then by the Markov property, the distribution of Y_{n+1} conditional on $\mathcal{F}_{n-N}$ equals $P^{N+1}(\mathbf{y}, \cdot)$, where P denotes the transition kernel of the chain $(Y_n)_{t \in \mathbb{N}}$. Denote $L := \|\ell(\cdot, \cdot)\|_\infty$. Denote $\boldsymbol{\theta} := \boldsymbol{\theta}_{n-N}$, which is deterministic with respect to $\mathcal{F}_{n-N}$.

We first claim that

$$\left| \mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) \,\Big|\, \mathcal{F}_{n-N} \right] - f(\boldsymbol{\theta}) \right| \leq 2L\|P^{N+1}(\mathbf{y}, \cdot) - \pi\|_{TV} \leq 2L\lambda^{N+1}. \tag{8.1}$$

This would yield

$$\mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) \,\Big|\, \mathcal{F}_{n-N} \right] \leq f(\boldsymbol{\theta}) + \left| \mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) \,\Big|\, \mathcal{F}_{n-N} \right] - f(\boldsymbol{\theta}) \right| \leq f(\boldsymbol{\theta}) + 2L\lambda^{N+1}.$$

To justify (8.1), note that by a change of measure,

$$
\mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) \,\Big|\, \mathcal{F}_{n-N}\right] = \sum_{\mathbf{y}' \in \Omega} \ell(\varphi(\mathbf{y}'), \boldsymbol{\theta}) \, P^{N+1}(\mathbf{y}, \mathbf{y}')
$$
$$
= \sum_{\mathbf{y}' \in \Omega} \ell(\varphi(\mathbf{y}'), \boldsymbol{\theta}) \, \pi(\mathbf{y}') + \sum_{\mathbf{y}' \in \Omega} \ell(\varphi(\mathbf{y}'), \boldsymbol{\theta})(P^{N+1}(\mathbf{y}, \mathbf{y}') - \pi(\mathbf{y}'))
$$
$$
= f(\boldsymbol{\theta}) + \sum_{\mathbf{y}' \in \Omega} \ell(\varphi(\mathbf{y}'), \boldsymbol{\theta})(P^{N+1}(\mathbf{y}, \mathbf{y}') - \pi(\mathbf{y}')).
$$

Then the claim follows by bouding $\ell(\varphi(\mathbf{y}'), \boldsymbol{\theta}) \le \|\cdot, \boldsymbol{\theta}\|_\infty$ for all $\mathbf{y}' \in \Omega$ and

$$
\sum_{\mathbf{y}' \in \Omega} \left|P^{N+1}(\mathbf{y}, \mathbf{y}') - \pi(\mathbf{y}')\right| = 2\|P^{N+1}(\mathbf{y}, \cdot) - \pi\|_{TV}.
$$

Next, we analyze $\mathbb{E}\left[-\bar{f}_n(\boldsymbol{\theta}) \,|\, \mathcal{F}_{n-N}\right]$ using a different approach. Namely, we decompose the times interval $[0, n]$ into two intervals $[1, n-N]$ and $[n-N+1, t]$ and consider the trajectory $(\mathbf{x}_k)_{1 \le k \le n}$ restricted onto each of these intervals. Since we are conditioning on $\mathcal{F}_{n-N}$, the trajectory $(\mathbf{x}_k)_{1 \le k \le n-N}$ is fully observed and may be far from the average behavior. The distribution of $\mathbf{x}_k$ then becomes close to the stationary distribution π during the second interval of length at an exponential mixing rate of λ (see (A2)). We first compare the trajectory on the second interval using the Markov chain mixing, and then compare the mean of the trajectory of the first interval with the stationary trajectory by integrating out the conditioning on $\mathcal{F}_{n-N}$. See Figure 6 for illustration.
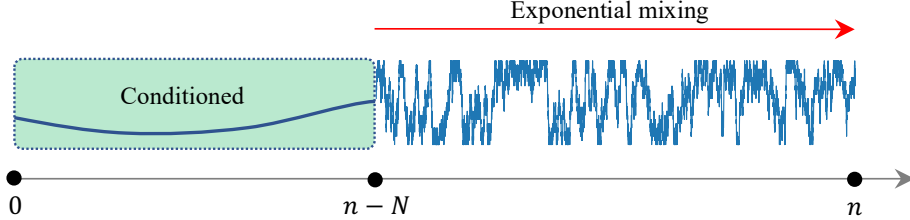
Figure 6: Illustration of the decomposition of the process used to bound the positive variation.

We proceed with the sketch given above. Denote $\hat{f}_{n:N} := \sum_{k=n-N+1}^n \ell(\mathbf{x}_k, \cdot) w_k^n$. Write

$$
\mathbb{E}\left[-\bar{f}_n(\boldsymbol{\theta}) \,\Big|\, \mathcal{F}_{n-N}\right] = -\mathbb{E}\left[\sum_{k=1}^n \ell(\mathbf{x}_k, \boldsymbol{\theta}) w_k^n \,\Big|\, \mathcal{F}_{n-N}\right]
$$
$$
= -\left(\sum_{k=1}^{n-N} \ell(\mathbf{x}_k, \boldsymbol{\theta}) w_k^n\right) - \mathbb{E}\left[\hat{f}_{n:N} \,\Big|\, \mathcal{F}_{n-N}\right]
$$

Next, denote $C_N = \sum_{k=n-N+1}^{n} w_k^n$. Using Markov property, we have

$$\left| C_N f(\boldsymbol{\theta}) - \mathbb{E}\left[\hat{f}_{n:N}(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{n-N}\right]\right| \leq \sum_{k=t-N+1}^{n} \left| f(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{y} \sim \pi_{k-N}}[\ell(\varphi(\mathbf{y}), \boldsymbol{\theta})]\right| w_k^n$$

$$\leq 2L \sum_{k=1}^{n-N} \|P^k(Y_{n-N}, \cdot) - \pi\|_{TV} w_k^n$$

$$\leq 2L w_n \sum_{k=1}^{n} \sup_{\mathbf{y} \in \mathfrak{X}} \|P^k(\mathbf{y}, \cdot) - \pi\|_{TV} \leq \frac{2L w_n}{1 - \lambda},$$

by using $w_1^n \leq \cdots \leq w_n^n = w_n \leq 1$ under (A4) and also (A2) to bound the total variation distance terms. Hence by triangle inequality,

$$\mathbb{E}\left[-\bar{f}_n(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{n-N}\right] \leq -\sum_{k=1}^{n-N} \ell(\mathbf{x}_k, \boldsymbol{\theta}) w_k^n - C_N f(\boldsymbol{\theta}) + \mathbb{E}\left[\left|C_N f(\boldsymbol{\theta}) - \hat{f}_{n:N}(\boldsymbol{\theta})\right| \,\middle|\, \mathcal{F}_{n-N}\right] \quad (8.2)$$

$$\leq -\sum_{k=1}^{n-N} \ell(\mathbf{x}_k, \boldsymbol{\theta}) w_k^n - \sum_{k=n-N+1}^{n} f(\boldsymbol{\theta}) w_k^n + \frac{2L}{1 - \lambda}.$$

Then combining the above bounds with (8.1) and a triangle inequality gives

$$\left(\mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) - \bar{f}_n(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{n-N}\right]\right)^+$$

$$\leq f(\boldsymbol{\theta}) + \left|\mathbb{E}\left[\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{n-N}\right] - f(\boldsymbol{\theta})\right| + \mathbb{E}\left[-\bar{f}_n(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{n-N}\right]$$

$$\leq \left(\sum_{k=1}^{n-N} \left(f(\boldsymbol{\theta}) - \ell(\mathbf{x}_k, \boldsymbol{\theta})\right) w_k^n\right) + \frac{2L w_n}{1 - \lambda} + 2L\lambda^{N+1}$$

$$\leq \left(\sum_{k=1}^{n-N} \left(f(\boldsymbol{\theta}) - \ell(\mathbf{x}_k, \boldsymbol{\theta})\right) w_k^{n-N+M}\right) + \frac{2L w_{n-N}}{1 - \lambda} + 2L\lambda^{N+1},$$

where we have used that w_k is non-increasing in k and $w_k^n = w_k \prod_{i=k+1}^{n}(1 - w_i) \leq w_k \prod_{i=k+1}^{m}(1 - w_i) = w_k^m$ for $k \leq m \leq n$. Then by Lemma 20, we have

$$\mathbb{E}\left[\sum_{k=1}^{n-N+M} \left(f(\boldsymbol{\theta}) - \ell(\mathbf{x}_k, \boldsymbol{\theta})\right) w_k^{n-N+M}\right] \leq C w_{n-N} \quad (8.3)$$

Therefore, the assertion follows from (8.2) by integrating it with respect to $\mathcal{F}_{n-N}$ and using (8.3).

Lastly, suppose the data sequence $\mathbf{x}_n$ is i.i.d. from the stationary distribution π. We can then take $Y_n = \mathbf{x}_n$, so Y_n is a Markov chain with mixing rate $\lambda = 1$, since for any $\mathbf{y} \in \Omega$, the one-step conditional distribution $P(\mathbf{y}, \cdot)$ exactly equals the stationary distribution π. This shows the assertion. ∎

Next, we give proof of Lemma 21.

Proof of Lemma 21. We will first show **(i)-(v)** under cases C1-C2 in Theorem 2. We will then show these statements for case C3 in Theorem 2. After that, we will finally prove **(vi)**.

Denote $V(\mathbf{x}, \boldsymbol{\theta}) := \ell(\mathbf{x}, \boldsymbol{\theta}) - \bar{f}_n(\boldsymbol{\theta})$. By (A1), V is Lipschitz in $\boldsymbol{\theta}$ for some $R > 0$. Under cases C1-C2, by Lemma 17, we have $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| \leq c_0 w_k$ for all $k \geq 1$ for some constant $c_0 > 0$. Since w_k is non-increasing in k, by triangle inequality we get $\|\boldsymbol{\theta}_{n-N} - \boldsymbol{\theta}_n\| \leq c_0 \sum_{k=n-N}^{N} w_k \leq c_0 N w_{n-N}$. This yields

$$\mathbb{E}\left[V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) \,|\, \mathcal{F}_{n-N}\right]$$
$$= \mathbb{E}\left[V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_{n-N}) \,|\, \mathcal{F}_{n-N}\right] + \mathbb{E}\left[V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_{n-N}) \,|\, \mathcal{F}_{n-N}\right]$$
$$\leq \mathbb{E}\left[V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_{n-N}) \,|\, \mathcal{F}_{n-N}\right] + R\,\mathbb{E}\left[\|\boldsymbol{\theta}_{n-N} - \boldsymbol{\theta}_n\| \,|\, \mathcal{F}_{n-N}\right]$$
$$\leq \mathbb{E}\left[V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_{n-N}) \,|\, \mathcal{F}_{n-N}\right] + c_0 R N w_{n-N}.$$

Note that $\boldsymbol{\theta}_{n-N}$ is deterministic with respect to $\mathcal{F}_{n-N}$. Hence multiplying by $w_{n+1} \leq w_n$, taking positive parts, and using Proposition 26, we get

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[w_{n+1} V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) \,|\, \mathcal{F}_{n-a_n}\right]^+\right] \leq \sum_{n=1}^{\infty} C w_n w_{n-a_n} + 2L w_n \lambda^{a_n+1} + c_0 R a_n w_n w_{n-a_n}. \quad (8.4)$$

Now we show **(i)**. Denote $Z_n = w_{n+1} V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n)$. Then by iterated expectation and Jensen's inequality, it follows that

$$\sum_{n=1}^{\infty} \mathbb{E}[Z_n]^+ = \sum_{n=1}^{\infty} \left(\mathbb{E}\left[\mathbb{E}\left[Z_n \,\Big|\, \mathcal{F}_{n-a_t}\right]\right]\right)^+ \leq \sum_{t=1}^{\infty} \mathbb{E}\left[\left(\mathbb{E}\left[Z_t \,\Big|\, \mathcal{F}_{t-a_t}\right]\right)^+\right]. \quad (8.5)$$

Notice that the first expression above is the summation in **(i)** that we want to bound, and the last expression above equals the left-hand side of (8.4). This shows **(i)**.

Next, we show **(ii)**. Note that the left hand side in **(ii)** is bounded by the left hand side of **(i)** plus $\sum_{n=1}^{\infty} w_n^2 \left(\sum_{k=1}^{n+1} \mathbb{E}[\varepsilon_k]\right) < \infty$ by Proposition 18 and (A4). So **(ii)** follows from **(i)**.

For **(iii)**, observe from Proposition 18 and (6.2),

$$\sum_{n=1}^{\infty} w_{n+1}\left(\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right) \leq \bar{g}_1(\boldsymbol{\theta}_1) + \sum_{n=1}^{\infty} w_{n+1}\left(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right) + \sum_{n=1}^{\infty} w_n^2 \left(\sum_{k=1}^{n+1} \varepsilon_k\right).$$

Then **(iii)** follows from taking expectation and using **(i)** with (A4).

Next, we show **(iv)**. The argument is similar. Since the loss function ℓ is R-Lipschitz by (A1), so is $f - \bar{f}_n$. Then, as before, we can $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq c_0 w_n$ and w_n non-increasing to show

$$\left|\mathbb{E}\left[f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \,\Big|\, \mathcal{F}_{n-N}\right]\right| \leq \left|\mathbb{E}\left[f(\boldsymbol{\theta}_{n-N}) - \bar{f}_{n-N}(\boldsymbol{\theta}_n) \,\Big|\, \mathcal{F}_{n-N}\right]\right| + c_0 R N w_{n-N}.$$

Note that $\boldsymbol{\theta}_{n-N}$ is deterministic with respect to $\mathcal{F}_{n-N}$. Hence using Proposition 26 with $N = a_n$ where a_n as in (A4), it follows that

$$\left|\mathbb{E}\left[f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \,\Big|\, \mathcal{F}_{n-a_n}\right]\right| \leq \frac{2L w_{n-a_n}}{1 - \lambda} + c_0 R a_n w_{n-a_n}.$$

Multiplying by w_{n+1} and integrating with respect to $\mathcal{F}_{n-a_n}$, we get

$$w_{n+1}\mathbb{E}\left[\left|\mathbb{E}\left[f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\,\Big|\,\mathcal{F}_{n-a_n}\right]\right|\right] \leq \frac{2Lw_nw_{n-a_n}}{1-\lambda} + c_0Ra_nw_nw_{n-a_n}. \qquad (8.6)$$

By (A4), the right-hand side is summable. Then the assertion follows from Jensen's inequality similarly as in the proof of **(i)**.

We turn to prove **(v)**. Suppose the optional condition in (A4) holds, that is,

$$\sum_{n=1}^{\infty} w_nw_{n-a_n}\sqrt{n} < \infty,$$

where a_n is the sequence in (A4). As in the proof of **(iv)**, and using the second inequality in Lemma 20,

$$\mathbb{E}\left[\left|f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right|\,\Big|\,\mathcal{F}_{n-N}\right] \leq \mathbb{E}\left[\left|f(\boldsymbol{\theta}_{n-N}) - \bar{f}_{n-N}(\boldsymbol{\theta}_n)\right|\,\Big|\,\mathcal{F}_{n-N}\right] + c_0RNw_{n-N}$$
$$\leq Cw_{n-N}\sqrt{n-N} + c_0RNw_{n-N},$$

where $C > 0$ is a constant in Lemma 20. Then letting $N = a_n$ in (A4) integrating out $\mathcal{F}_{n-N}$, multiplying by $w_{n+1} \leq w_n$ and summing over all $n \geq 1$ gives

$$\sum_{n=0}^{\infty} w_{n+1}\mathbb{E}\left[\left|f(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)\right|\right] \leq \sum_{n=1}^{\infty} Cw_nw_{n-a_n}\sqrt{n-a_n} + c_0Ra_nw_nw_{n-a_n} < \infty.$$

Now, we establish **(i)**-**(v)** under case C3 in Theorem 2. This case must be dealt with separately since the stability of estimates $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ is not readily available (see Lemma 17). In this case, we rely on 'instant mixing' ($\lambda = 0$) of the i.i.d. data sequence $\mathbf{x}_n$ stated in the second part of Proposition 26. Namely, we condition on $\mathcal{F}_n$ and use $\lambda = 0$ while avoiding using $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$. For **(i)**, we first use (8.4) with $a_n = 0 = N$ to get

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[w_{n+1}V(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n)\,|\,\mathcal{F}_n\right]^+\right] \leq \sum_{n=1}^{\infty} Cw_n^2 < \infty.$$

The rest follows from (8.5). Then **(ii)**-**(iii)** can be deduced similarly as in the previous case using **(i)**. Also, **(iv)** follows from (8.6) using $\lambda = 0$ and $N = 0$. **(v)** can be deduced similarly.

Lastly, we show **(vi)**. For cases C1-C2 in Theorem 2, this is trivial since we have $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ and $\sum_{n=1}^{\infty} w_n^2 < \infty$ by (A4). Suppose C3. By Lemma 17 **(i)**, we have

$$\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \geq \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 - \Delta_n.$$

Moreover, using (6.3) and (6.2), we get

$$-\Delta_{n+1} \leq \bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1})$$
$$\leq \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) + w_{n+1}(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)) + w_n^2\sum_{k=1}^{n}\varepsilon_k.$$

Taking expectation and summing over $n \geq 1$ and using **(i)** as well as (A4),

$$\left| \sum_{n=1}^{\infty} \mathbb{E}\left[\bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) \right] \right| < \infty.$$

Using (A5), this implies

$$\mathbb{E}\left[\sum_{n=1}^{N} \frac{\hat{\rho}}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \right] \leq \sum_{n=1}^{\infty} \mathbb{E}\left[\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \right] + \mathbb{E}\left[\sum_{n=1}^{\infty} \Delta_n \right] < \infty,$$

as desired. ■

Lastly in this section, we prove Lemma 22.

Proof of Lemma 22. Denote $\alpha_n = \nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n)$ and $\beta_n = \nabla \bar{f}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n)$. Fix $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\varepsilon > 0$. With (A1) and a recursion, $\bar{f}_n$ and g_n are also R-Lipschitz continuous. By Lemma 34, we can write

$$|\bar{g}_n(\boldsymbol{\theta}_n + \varepsilon\boldsymbol{\theta}) - \bar{g}_n(\boldsymbol{\theta}_n) - \langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \varepsilon\boldsymbol{\theta} \rangle| \leq \frac{L\varepsilon^2}{2} \|\boldsymbol{\theta}\|^2,$$

$$\left| \bar{f}_n(\boldsymbol{\theta}_n + \varepsilon\boldsymbol{\theta}) - \bar{f}_n(\boldsymbol{\theta}_n) - \langle \nabla \bar{f}_n(\boldsymbol{\theta}_n), \varepsilon\boldsymbol{\theta} \rangle \right| \leq \frac{L\varepsilon^2}{2} \|\boldsymbol{\theta}\|^2,$$

for some uniform constant $L > 0$ that does not depend on $\boldsymbol{\theta}, \varepsilon$ and n. Recall that $\bar{g}_n \geq \bar{f}_n - \bar{\varepsilon}_n$ for all $n \geq 1$. Hence

$$-\frac{L\varepsilon^2}{2} \|\boldsymbol{\theta}\|^2 + \bar{f}_n(\boldsymbol{\theta}_n) + \langle \nabla \bar{f}_n(\boldsymbol{\theta}_n), \varepsilon\boldsymbol{\theta} \rangle \leq \bar{f}_n(\boldsymbol{\theta}_n + \varepsilon\boldsymbol{\theta})$$

$$\leq \bar{g}_n(\boldsymbol{\theta}_n + \varepsilon\boldsymbol{\theta}) + \bar{\varepsilon}_n$$

$$\leq \bar{g}_n(\boldsymbol{\theta}_n) + \langle \nabla \bar{g}_n(\boldsymbol{\theta}_n, \varepsilon\boldsymbol{\theta} \rangle + \frac{L\varepsilon^2}{2} \|\boldsymbol{\theta}\|^2 + \bar{\varepsilon}_n,$$

so, we obtain the following key inequality

$$\langle \alpha_n, \varepsilon\boldsymbol{\theta} \rangle = \mathrm{tr}\left(\left(\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{f}_n(\boldsymbol{\theta}_n) \right)^T (\varepsilon\boldsymbol{\theta}) \right) \geq \left(\bar{f}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n) - \bar{\varepsilon}_n \right) - L\varepsilon^2 \|\boldsymbol{\theta}\|^2.$$

Choosing $\boldsymbol{\theta} = -\alpha_n$, we get

$$-\varepsilon \|\alpha_n\|^2 \geq \left(\bar{f}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n) - \bar{\varepsilon}_n \right) - c\varepsilon^2 \|\alpha_n\|^2.$$

Rearranging, we get

$$(\varepsilon - c\varepsilon^2) \|\alpha_n\|^2 \leq \left(\bar{g}_n(\boldsymbol{\theta}_n) + \bar{\varepsilon}_n - \bar{f}_n(\boldsymbol{\theta}_n) \right).$$

Recall that this holds for all $\varepsilon > 0$ and $n \geq 1$. Taking expectation, multiplying by w_n, and summing up in n,

$$(\varepsilon - c\varepsilon^2) \sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\alpha_n\|^2 \right] \leq \sum_{n=1}^{\infty} w_n \mathbb{E}\left[\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n) \right] + \sum_{n=1}^{\infty} w_n \mathbb{E}[\bar{\varepsilon}_n] < \infty,$$

where the finiteness above uses Lemma 21 and (A4) with (6.1). Since $\varepsilon > 0$ was arbitrary, this shows

$$\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\alpha_n\|^2\right] < \infty. \tag{8.7}$$

Next, we will analyze the sums $\sum_{n=1}^{\infty} w_n \|\mathbb{E}[\beta_n]\|^2$ and $\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\beta_n\|^2\right]$. To this end, note that

$$\nabla f(\boldsymbol{\theta}) = \nabla \mathbb{E}_{\mathbf{x} \sim \pi}\left[\ell(\mathbf{x}, \boldsymbol{\theta})\right] = \mathbb{E}_{\mathbf{x} \sim \pi}\left[\nabla_2 \ell(\mathbf{x}, \boldsymbol{\theta})\right].$$

Define $\psi : (\mathbf{x}, \boldsymbol{\theta}) \mapsto \nabla_2 \ell(\mathbf{x}, \boldsymbol{\theta})$. Then by the linearity of gradients and induction, we can show

$$\bar{\psi}_n(\boldsymbol{\theta}) = \nabla \bar{f}_n(\boldsymbol{\theta}).$$

Thus by Lemma 20, there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\|\mathbb{E}\left[\nabla f(\boldsymbol{\theta})\right] - \mathbb{E}\left[\nabla \bar{f}_n(\boldsymbol{\theta})\right]\right\|^2 \leq C w_n, \quad \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \Theta} \left\|\nabla f(\boldsymbol{\theta}) - \nabla \bar{f}_n(\boldsymbol{\theta})\right\|^2\right] \leq C w_n \sqrt{n}.$$

By (A4), this implies

$$\sum_{n=1}^{\infty} w_n \|\mathbb{E}[\beta_n]\|^2 \leq \sum_{n=1}^{\infty} w_n \sup_{\boldsymbol{\theta} \in \Theta} \left\|\mathbb{E}\left[\nabla f(\boldsymbol{\theta})\right] - \mathbb{E}\left[\nabla \bar{f}_n(\boldsymbol{\theta})\right]\right\|^2 \leq C \sum_{n=1}^{\infty} w_n^2 < \infty.$$

Similarly, we also get

$$\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\beta_n\|^2\right] \leq \sum_{n=1}^{\infty} w_n \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta})\| - \mathbb{E}\left[\nabla \bar{f}_n(\boldsymbol{\theta})\right\|^2\right] \leq C \sum_{n=1}^{\infty} w_n^2 \sqrt{n}. \tag{8.8}$$

Now we conclude the assertions. For **(i)**, by Jensen's inequality, we have

$$\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\alpha_n\|^2\right] \leq \sum_{n=1}^{\infty} w_n \|\mathbb{E}[\alpha_n]\|^2 < \infty.$$

Hence by Fubini's theorem, we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_n \left(\|\alpha_n\|^2 + \|\mathbb{E}[\beta_n]\|^2\right)\right] = \sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\alpha_n\|^2\right] + \sum_{n=1}^{\infty} w_n \|\mathbb{E}[\beta_n]\|^2 < \infty.$$

This shows **(i)**.

Next, assume the optional condition in (A4) holds. since w_n is non-increasing in n, it follows that $\sum_{n=1}^{\infty} w_n^2 \sqrt{n} < \infty$. Then (8.8) shows $\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\beta_n\|^2\right] < \infty$. We have shown that $\sum_{n=1}^{\infty} w_n \mathbb{E}\left[\|\alpha_n\|^2\right] < \infty$ in (8.7). Then noting that $\|\alpha_n + \beta_n\|^2 \leq 2\|\alpha_n\|^2 + 2\|\beta_n\|^2$, using Fubini's theorem, we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty} w_n \left(\|\alpha_n\|^2 + \|\beta_n\|^2 + \|\alpha_n + \beta_n\|^2\right)\right] < \infty.$$

Since $\bar{g}_n(\boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n) = \alpha_n + \beta_n$, this is the desired conclusion. ∎

9. Proof of Lemma 24

The goal of this section is to prove Lemma 24. Throughout this and the following sections, let $(\boldsymbol{\theta}_n)_{n\geq 1}$ be an output of Algorithm 1, where each update $\boldsymbol{\theta}_{n-1} \to \boldsymbol{\theta}_n$ uses either Algorithm 2 (full coordinate descent with proximal regularization) or Algorithm 3 (block coordinate descent with diminishing radius). For the latter case, decompose the update $\boldsymbol{\theta}_{n-1} \to \boldsymbol{\theta}_n$ into $\boldsymbol{\theta}_{n-1} = \boldsymbol{\theta}_n^{(0)} \to \boldsymbol{\theta}_n^{(1)} \to \cdots \to \boldsymbol{\theta}_n^{(m)} = \boldsymbol{\theta}_n$, where each $\boldsymbol{\theta}_n^{(i)}$ is the output of (3.4) at sub-iteration i of Algorithm 3. Namely, let $\mathbb{J}$ denote the set of coordinate blocks used in Algorithm 1 and let $J_1(n), \ldots, J_m(n) \in \mathbb{J}$ denote the coordinate blocks used in the ith iteration of Algorithm 3 for the update $\boldsymbol{\theta}_{n-1} \to \boldsymbol{\theta}_n$. Then $\boldsymbol{\theta}_n^{(i)}$ is an approximate minimizer of $\bar{g}_n(\boldsymbol{\theta}) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$ over the convex set $\boldsymbol{\Theta}_n^{J_i(n)}$ (see (3.3)) with optimality gap $\Delta_n^{(i)}$ (see (A5)), where $\Psi_n(x)$ takes value 0 if $x \leq w_n/m$ and ∞ otherwise (radius restriction). Let $\boldsymbol{\theta}_n^{(i\star)}$ denote an exact minimizer of $\bar{g}_n(\boldsymbol{\theta}) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$ over $\boldsymbol{\Theta}_n^{J_i(n)}$.

For a unified treatment, we also use the same notation when Algorithm 2 is used instead of Algorithm 3. In this case, we take $m = 1$ (single update), $\mathbb{J} = \{\{1,\ldots,p\}\}$, $J_n^{(1)} = \{1,\ldots,p\}$, and $\boldsymbol{\Theta}_n^{J_1(n)} = \boldsymbol{\Theta}$ (no frozen coordinates), and $\Psi_n(x) \equiv \frac{\hat{\rho}}{2}x^2$ (proximal regularization). In this case, the optimality gap is denoted as $\Delta_n^{(1)} = \Delta_n$ (see (A5)).

The starting point of the analysis in this section is the following proposition, which states that the linear variation of $\bar{g}_n$ over the sequence $(\boldsymbol{\theta}_{n;i})_{n\geq 1, 1\leq i\leq m}$ is finite.

Proposition 27 *Assume (A1)-(A4). Under any of the three cases C1-C3 in Theorem 2, we have*

$$\mathbb{E}\left[\sum_{n=1}^\infty |\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle|\right] < \infty. \tag{9.1}$$

Furthermore, almost surely,

$$\sum_{n=1}^\infty \mathbb{E}_{J_1(n),\ldots,J_m(n)}\left[|\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle| \,\Big|\, \mathcal{F}_{n-1}\right] < \infty.$$

Proof Fix $i = 1,\ldots,m$. By Proposition 16, $\nabla \bar{g}_n$ over $\boldsymbol{\Theta}$ is L-Lipschitz for all $n \geq 1$. Hence by Lemma 34, for all $t \geq 1$,

$$\left|\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) - \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\right\rangle\right| \leq \frac{L}{2}\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|^2.$$

Note that for $1 \leq i \leq m$,

$$\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) + \Psi_n(\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|) \leq \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}) + \Psi_n(\|\boldsymbol{\theta}_n^{(i-1)} - \boldsymbol{\theta}_{n-1}\|) + \Delta_n^{(i)}.$$

By induction, we get

$$\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) \leq \bar{g}_n(\boldsymbol{\theta}_n^{(i)}) + \Psi_n(\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|) \leq \bar{g}_n(\boldsymbol{\theta}_{n-1}) + \sum_{j=1}^i \Delta_n^{(j)}.$$

Hence

$$|\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) - \bar{g}_n(\boldsymbol{\theta}_{n-1})| \leq \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n^{(i)}) + 2\sum_{J=1}^{i}\Delta_n^{(j)},$$

so it follows that

$$\left|\left\langle \nabla\bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\right\rangle\right| \leq \frac{L}{2}\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|^2 + \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n^{(i)}) + 2\sum_{j=1}^{i}\Delta_n^{(j)}$$

$$\leq \frac{L}{2}\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|^2 + \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) + 3\sum_{j=i}^{m}\Delta_n^{(j)}.$$

On the other hand, using (6.3) and $\bar{g}_n \geq \bar{f}_n$, note that

$$0 \leq \bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) \leq \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) + w_{n+1}(\ell(\mathbf{x}_{n+1}, \boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)).$$

Hence using Lemma 21, we have

$$\sum_{n=1}^{\infty}\mathbb{E}\left[\bar{g}_{n+1}(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1})\right] < \infty.$$

Next, by Lemma 21, we have

$$\sum_{n=1}^{\infty}\mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2\right] < \infty.$$

Moreover, we have $\mathbb{E}\left[\sum_{n=1}^{\infty}\Delta_n\right] < \infty$ by (A5) for cases C2 and C3 and by Lemma 39 for case C1. Thus the above inequalities yield

$$\sum_{n=1}^{\infty}\mathbb{E}\left[\left|\left\langle \nabla\bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\right\rangle\right|\right] < \infty.$$

Taking $i = m$ and using Fubini's theorem, we obtain the first assertion (9.1). Furthermore, one can rewrite (9.1) as

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\mathbb{E}_{J_1(n),...,J_m(n)}\left[|\langle\nabla\bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle|\,\bigg|\,\mathcal{F}_{n-1}\right]\right] < \infty.$$

Since the random variable inside the expectation above is nonnegative, it has to be finite almost surely. This shows the second part of the assertion. ■

Next, we show that the block coordinate descent we use to obtain $\boldsymbol{\theta}_n$ should always give the optimal first-order descent up to a small additive error.

Lemma 28 (Approximate first-order optimality) *Assume (A1)-(A4). Assume cases C1-C3 in Theorem 2. Then there exists constants $c_1, \ldots, c_4 > 0$ such that almost surely, for any sequence $(b_n)_{n \geq 1}$, $0 < b_n \leq w_n/m$,*

$$b_n \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq 1} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle$$

$$\leq \mathbb{E}\left[\langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle \,\Big|\, \mathcal{F}_{n-1} \right] + \mathbb{E}\left[\Delta_n \,\Big|\, \mathcal{F}_{n-1} \right]$$

$$+ c_1 b_n w_n + c_2 b_n^2 + c_3 \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \,\Big|\, \mathcal{F}_{n-1} \right] + c_4 w_n^2.$$

Furthermore, the above holds for all $b_n > 0$ for cases C1 and C3.

Proof We first show the assertion for cases C1 and C3. Fix arbitrary $\boldsymbol{\theta} = \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$. Recall that $\boldsymbol{\theta}_n$ is an inexact minimizer of $\bar{g}_n(\boldsymbol{\theta}) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$ over $\boldsymbol{\Theta}$. Let $\boldsymbol{\theta}_n^\star$ denote an exact minimizer of the same problem. Then

$$\begin{aligned}
\bar{g}_n(\boldsymbol{\theta}_n) - \Delta_n &\leq \bar{g}_n(\boldsymbol{\theta}_n) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|) - \Delta_n \\
&\leq \bar{g}_n(\boldsymbol{\theta}_n^\star) + \Psi_n(\|\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}\|) \\
&\leq \bar{g}_n\left(\boldsymbol{\theta}_{n-1} + (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})\right) + \Psi_n(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|).
\end{aligned}$$

Note that by Proposition 16, $\nabla \bar{g}_n$ is L-Lipschitz and $\|\nabla g_{n+1} - \nabla g_n\| \leq L' w_{n+1}$ for some constant $L' > 0$. Hence subtracting $\bar{g}_n(\boldsymbol{\theta}_{n-1})$ from both side, we get

$$\begin{aligned}
\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle &- \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 - \Delta_n \\
&\leq \langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{(L + \hat{\rho})}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \\
&\leq \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{(L + \hat{\rho})}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 + L' w_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|,
\end{aligned}$$

where we have used $\Psi_n(x) \equiv \frac{\hat{\rho}}{2} x^2$ for the first inequality. Rearranging, we get

$$\begin{aligned}
\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle \leq \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle &+ \frac{(L + \hat{\rho})}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \\
&+ L' w_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| + \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 + \Delta_n.
\end{aligned}$$

It follows that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle \tag{9.2}$$

$$\leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle + \frac{(L + \hat{\rho}) b_n^2}{2}$$

$$+ L' b_n w_n + \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 + \Delta_n.$$

Now observe that by the convexity of $\boldsymbol{\Theta}$, if $\boldsymbol{\theta}_{n-1} + av \in \boldsymbol{\Theta}$ for a scalar $a \geq 1$ and $v \in \mathbb{R}^p$, then $\boldsymbol{\theta}_n + v \in \boldsymbol{\Theta}$. Hence we get

$$
\sup_{\|u\| \leq 1,\, \boldsymbol{\theta}_{n-1}+u\in\boldsymbol{\Theta}} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}),\, u \rangle = \sup_{\|v\| \leq b_n,\, \boldsymbol{\theta}_{n-1}+b_n^{-1}v\in\boldsymbol{\Theta}} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}),\, b_n^{-1}v \rangle
$$

$$
\leq b_n^{-1} \sup_{\|v\| \leq b_n,\, \boldsymbol{\theta}_{n-1}+v\in\boldsymbol{\Theta}} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}),\, v \rangle. \qquad (9.3)
$$

Then the assertion follows from combining (9.3) and (9.2) then taking the conditional expectation with respect to $\mathcal{F}_{n-1}$.

Next, we show the assertion for case C2. The argument is similar but a bit more complicated due to the use of randomized block coordinate descent in Algorithm 3. As before, fix arbitrary $\boldsymbol{\theta} = \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$. We will suppress the dependency of random block coordinate $J_i(n)$ on n below and denote it J_i. Let $\boldsymbol{\Theta}_n^{J_i}$ be as in (3.3), which is a convex subset of $\boldsymbol{\Theta}$ that contains both $\boldsymbol{\theta}_n^{(i-1)}$ and $\boldsymbol{\theta}_n^{(i-1)} + \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_n^{(i-1)})^j e_j$, where e_j denote the jth standard basis vector of $\mathbb{R}^p$. Recalling that $\boldsymbol{\theta}_n^{(i)}$ is an approximate minimizer of $\bar{g}_n(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}_n^{J_i}$ with gap $\Delta_n^{(i)}$ (see (A5)),

$$
\bar{g}_n(\boldsymbol{\theta}_n^{(i)}) - \Delta_n^{(i)} \leq \bar{g}_n(\boldsymbol{\theta}_n^{(i\star)}) \leq \bar{g}_n\left(\boldsymbol{\theta}_n^{(i-1)} + \left(\sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_n^{(i-1)})^j e_j \right) \right),
$$

where $\boldsymbol{\theta}_n^{(i\star)}$ denotes an exact minimizer of $\bar{g}_n$ over $\boldsymbol{\Theta}_n^{J_i}$. Then subtract $\bar{g}_n(\boldsymbol{\theta}_n^{(i-1)})$ from both sides and use Lemma 34. Note that by Proposition 16, $\nabla \bar{g}_n$ is L-Lipschitz and $\|\nabla g_{n+1} - \nabla g_n\| \leq L' w_{n+1}$ for some constant $L' > 0$. Also, note that the coordinate blocks $J_1, \ldots, J_m$ are disjoint so that each coordinate of $\boldsymbol{\theta}_{n-1}$ is updated at most once, see (A6). This implies

$$
\sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_n^{(i-1)})^j e_j = \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j,
$$

$$
\|\boldsymbol{\theta}_n^{(i-1)} - \boldsymbol{\theta}_{n-1}\|,\, \|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)}\| \leq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|.
$$

So it follows that

$$
\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}),\, \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle - \frac{L}{2}\|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)}\|^2 - \Delta_n^{(i)}
$$

$$
\leq \left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}),\, \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2
$$

$$
\leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_n^{(i-1)}),\, \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 + L' w_n\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|
$$

$$
\leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}),\, \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + \frac{L}{2}b_n^2 + L' w_n b_n + L b_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|.
$$

59

Also, by Cauchy-Schwarz inequality and Proposition 16,

$$
\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}), \, \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle
$$
$$
\geq \left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle - L\|\boldsymbol{\theta}_n^{(i-1)} - \boldsymbol{\theta}_{n-1}\| \, \|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)}\|
$$
$$
\geq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle - L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 - L' w_n^2.
$$

Thus combining the above inequalities, we obtain

$$
\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle
$$
$$
\leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \, \sum_{j \in J_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle
$$
$$
+ c_1 \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 + c_2 b_n^2 + c_3 w_n b_n + c_4 w_n^2 + \Delta_n^{(i)},
$$

where $c_1, \cdots, c_4 > 0$ are constants. Hence summing over $i = 1, \ldots, m$,

$$
\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle \tag{9.4}
$$
$$
\leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \, \sum_{i=1}^m \sum_{j \in J_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle
$$
$$
+ c_1' \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 + c_2' L b_n^2 + c_3' w_n b_n + c_4' w_n^2 + \Delta_n,
$$

where we denoted $c_i' = mc_i$ for $i = 1, 2, 3$. Note that by (A6), for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$
\mathbb{E}_{J_1(n), \ldots, J_m(n)} \left[\sum_{i=1}^m \sum_{j \in J_i(n)} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \,\middle|\, \mathcal{F}_{n-1} \right] = \bar{c}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}), \tag{9.5}
$$

where $\bar{c}$ denotes the common expected number of each coordinate $j \in \{1, \ldots, m\}$ appearing in the coordinate blocks $J_1(n), \ldots, J_m(n)$. Thus taking the conditional expectation with respect to $\mathcal{F}_{n-1}$ on both sides of (10.3), and taking supremum for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$, we get

$$
\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \bar{c} \left\langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \right\rangle
$$
$$
\leq \mathbb{E}\left[\langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle \,\middle|\, \mathcal{F}_{n-1} \right] + c_1' \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \,\middle|\, \mathcal{F}_{n-1} \right]
$$
$$
+ c_2' L b_n^2 + c_3' w_n b_n + c_4' w_n^2 + \mathbb{E}\left[\Delta_n \,\middle|\, \mathcal{F}_{n-1} \right].
$$

Then combining the above with (9.3) finishes the proof. ∎

Now we show Lemma 24.

Proof of Lemma 24. By using Lemma 28 with deterministic $b_n = w_n$ and taking the conditional expectation with respect to $\mathcal{F}_n$, we have

$$w_{n+1} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1} \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \right] \leq \mathbb{E} \left[\langle \nabla \bar{g}_{n+1}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \, \Big| \, \mathcal{F}_n \right] \qquad (9.6)$$

$$+ c w_{n+1}^2 + \mathbb{E}\left[\Delta_{n+1}\right] + \mathbb{E} \left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \, \Big| \, \mathcal{F}_n \right]$$

for some constant $c > 0$ for all $n \geq 1$. By Proposition 27 and (A4), the right-hand side above except for the last term is summable. For the last term, note that by Lemma 21 **(vi)** and Fubini's theorem, we have

$$\mathbb{E} \left[\sum_{n=1}^{\infty} \mathbb{E} \left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \, \Big| \, \mathcal{F}_n \right] \right] = \mathbb{E} \left[\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \right] < \infty.$$

Hence $\sum_{n=1}^{\infty} \mathbb{E} \left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \, \Big| \, \mathcal{F}_n \right] < \infty$ almost surely. Thus the right-hand side of (9.6) is summable almost surely. ∎

10. Proof of Lemma 23

In this section, we prove the only remaining key lemma, Lemma 23, on asymptotic stationarity of the iterates with respect to the averaged surrogate loss functions. We proceed by first proving the assertion for cases C1 and C3 and then for the case C2 in Theorem 2.

Throughout this section, we will make use of the parameterized surrogate assumption (A7). From this, there exists a compact set $\mathcal{K}$ and a function $\hat{g} : \mathcal{K} \times \boldsymbol{\Theta} \to [0, \infty)$ such that for all $n \geq 1$, $\bar{g}_n(\boldsymbol{\theta}) = \hat{g}(\kappa_n, \boldsymbol{\theta})$ for some $\kappa_n \in \mathcal{K}$. Furthermore, $\hat{g}$ is Lipschitz in the first coordinate. Hence by taking a subsequence of $(t_k)_{k \geq 1}$, we may assume that $\kappa_{\infty} := \lim_{k \to \infty} \kappa_{t_k}$. Hence the function $\bar{g}_{\infty} := \lim_{k \to \infty} \bar{g}_{t_k} = \hat{g}(\kappa_{\infty}, \cdot)$ is well-defined. By continuity of $\nabla \bar{g}_n$ (see Lemma 34), we also have $\nabla g_{t_k} \to \nabla g_{\infty}$ almost surely.

10.1 Proof of Lemma 23 for cases C1 and C3 in Theroem 2.

In this subsection, we prove Lemma 23 assuming cases C1 and C3 in Theroem 2. The argument for case C3 may require a bit of care since we cannot use per-iteration stability $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$ (see Lemma 17) and rely on the total stability $\mathbb{E}\left[\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\right] < \infty$ (see Lemma 21 **(vi)**). Nontheless, Lemma 23 for both cases C1 and C3 can be shown using a simple argument that only uses $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = o(1)$.

We begin with a simple observation.

Proposition 29 *Assume the cases **C1** or **C3** in Theorem 2. Take $\hat{\rho} = 0$ for case C1 and $\hat{\rho} > -\rho$ for case C2. For each $n \geq 1$, let $\boldsymbol{\theta}_n^{\star}$ be the minimizer of the strongly convex function $\boldsymbol{\theta} \mapsto \bar{G}_n(\boldsymbol{\theta}) := \bar{g}_n(\boldsymbol{\theta}) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2$ over the convex set $\boldsymbol{\Theta}$. Then there exists a constant $\alpha > 0$ such that for all $n \geq 1$,*

$$\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^{\star}\|^2 \leq \alpha \Delta_n.$$

Proof Note that $\bar{G}_n$ is γ-strongly convex with modulus of convexity $\gamma = \rho > 0$ for case C1 and $\gamma = (\hat{\rho} - |\rho|)$ for case C2. Then the assertion follows from

$$\frac{\gamma}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^\star\|^2 \leq \bar{G}_n(\boldsymbol{\theta}_n) - \bar{G}_n(\boldsymbol{\theta}_n^\star) \leq \Delta_n$$

for $n \geq 1$. Indeed, the first inequality follows from the second-order growth property (see Lemma 37) since $\bar{G}_n^{(n)}$ is a γ-weakly convex function minimized at $\boldsymbol{\theta}_n$ over $\boldsymbol{\Theta}$, and the second inequality follows from the definition of optimality gap Δ_n in (A5). ■

We are now ready to give proof of Lemma 23 for no regularization (case C1) and proximal regularization (case C3).

Proof of Lemma 23 for cases C1 and C3 in Theorem 2. Assume (A1)-(A4), (A7), and one of the cases C1 or C3 in Theorem 2. Fix an arbitrary convergent subsequence $(\boldsymbol{\theta}_{t_k})_{k \geq 1}$ of $(\boldsymbol{\theta}_n)_{n \geq 1}$ denote $\boldsymbol{\theta}_\infty := \lim_{k \to \infty} \boldsymbol{\theta}_{t_k}$. By taking a further subsequence, we may assume that the following limits $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{t_k}$ and $\nabla \bar{g}_\infty := \lim_{k \to \infty} \nabla \bar{g}_{t_k}$ exist almost surely. It suffies to show that $\boldsymbol{\theta}_\infty$ is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$.

We keep the same notation for $\hat{\rho}$ as in Proposition 29. By using the first-order optimality of $\boldsymbol{\theta}_n^\star \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \right)$, we have

$$\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^\star) + \hat{\rho}(\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle \geq 0 \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

On the other hand, using Cauchy-Schwarz inequality and Proposition 29,

$$\begin{aligned}
|\langle \nabla &\bar{g}_n(\boldsymbol{\theta}_n) + \hat{\rho}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle - \langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^\star) + \hat{\rho}(\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle| \\
&\leq |\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{g}_n(\boldsymbol{\theta}_n^\star) + \hat{\rho}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^\star), \, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle| \\
&\leq (\|\nabla \bar{g}_n(\boldsymbol{\theta}_n) - \nabla \bar{g}_n(\boldsymbol{\theta}_n^\star)\| + \hat{\rho}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^\star\|) \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \\
&\leq (L + \hat{\rho})\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \, \|\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_n\| \\
&\leq (L + \hat{\rho})\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \, \sqrt{\alpha \Delta_n}.
\end{aligned}$$

Similarly, also note that

$$|\langle \hat{\rho}(\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle| \leq \hat{\rho}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \, \|\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_n\| \leq \hat{\rho}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \, \sqrt{\alpha \Delta_n}.$$

Note that from (A5), $\mathbb{E}[\sum_{n=1}^\infty \Delta_n] < \infty$, and by Fubini's theorem, this implies that $\sum_{n=1}^\infty \Delta_n < \infty$ almost surely. Hence $\Delta_n = o(1)$. Furthermore, since $\boldsymbol{\theta}_{t_k}$ converges, the sequence $(\boldsymbol{\theta}_{t_k})_{k \geq 1}$ is bounded. So the above inequalities imply

$$\liminf_{k \to \infty} \langle \nabla \bar{g}_{t_k}(\boldsymbol{\theta}_{t_k}), \, \boldsymbol{\theta} - \boldsymbol{\theta}_{t_k} \rangle \geq 0 \qquad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Since $\nabla \bar{g}_{t_k} \to \nabla \bar{g}_\infty$ as $k \to \infty$, we conclude that $\langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \, \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which means that $\boldsymbol{\theta}_\infty$ is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$. ■

10.2 Proof of Lemma 23 for cases C2 in Theroem 2.

A key property of the iterates $(\boldsymbol{\theta}_n)_{n\geq 1}$ for cases C1-C2 in Theroem 2 is their stability, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$, which is given in Lemma 17. For C2, which we consider in this subsection, we have an extra complication due to the use of an additional diminishing radius condition. Roughly speaking, for each limit point $\boldsymbol{\theta}_\infty$ of the iterates $(\boldsymbol{\theta}_n)_{n\geq 1}$, we will need to show that $\boldsymbol{\theta}_\infty$ is not touching any trust region boundary we used along the way and it is indeed a stationary point of some limiting function of the averaged surrogates $\bar{g}_n$. The argument we provide here is adapted from Lyu and Li (2023), which was first developed for (deterministic) block coordinate descent with diminishing radius.

Recall that during the update $\boldsymbol{\theta}_{n-1} = \boldsymbol{\theta}_n^{(0)} \to \boldsymbol{\theta}_n^{(1)} \to \cdots \to \boldsymbol{\theta}_n^{(m)} = \boldsymbol{\theta}_n$ in Algorithm 3, each block coordinate changes by at most w_n/m in the Euclidean norm. Each $\boldsymbol{\theta}_n^{(i)}$ is found by approximately solving (3.3), while an exact minimizer is denoted as $\boldsymbol{\theta}_n^{(i\star)}$ (if there are multiple exact minimizers, then we choose an arbitrary one). We say $\boldsymbol{\theta}_n^\star := \boldsymbol{\theta}_n^{(m\star)}$ is a *short point* if $\|\boldsymbol{\theta}_n^{(i-1)} - \boldsymbol{\theta}_n^{(i\star)}\| < w_n/m$ for $i = 1, \ldots, m$ and *long point* otherwise. Observe that if $\boldsymbol{\theta}_n^\star$ is a short point, then imposing the search radius restriction in (3.4) has no effect, and $\boldsymbol{\theta}_n^\star$ is obtained from $\boldsymbol{\theta}_{n-1}$ by a single cycle of exact block minimization of $\bar{g}_n$ over $\boldsymbol{\Theta}$ with no trust-region constraints.

Lemma 30 below is a slight modification of Lemma 28 where we compare the optimality gap (the supremum in the left-hand side of (10.1)) to the expected first-order variation with respect to the optimal parameter change $\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}$ instead of the approximate change $\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}$.

Lemma 30 (Approximate first-order optimality for inexact trust-region) *Assume (A1)-(A4). Assume case C2 in Theorem 2. Then there exists a constant $c > 0$ such that almost surely,*

$$w_n \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\|\leq 1} \langle -\nabla\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\rangle \tag{10.1}$$

$$\leq \mathbb{E}\left[\langle -\nabla\bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}\rangle \,\Big|\, \mathcal{F}_{n-1}\right] + cw_n^2.$$

Proof The proof is similar to the one for Lemma 28. Fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\| \leq w_n/m$. Then we have

$$\bar{g}_n(\boldsymbol{\theta}_n^{(i\star)}) \leq \bar{g}_n\left(\boldsymbol{\theta}_n^{(i-1)} + \left(\sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_n^{(i-1)})^j e_j\right)\right).$$

Subtract $\bar{g}_n(\boldsymbol{\theta}_n^{(i-1)})$ from both sides and use Lemma 34. Note that by Proposition 16, $\nabla\bar{g}_n$ is L-Lipschitz and $\|\nabla g_{n+1} - \nabla g_n\| \leq L' w_{n+1}$ for some constant $L' > 0$. Also, note that the coordinate blocks $J_1, \ldots, J_m$ are disjoint so that each coordinate of $\boldsymbol{\theta}_{n-1}$ is updated at most once, see (A6). This implies

$$\sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_n^{(i-1)})^j e_j = \sum_{j\in J_i}(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j,$$

$$\sum_{j\in J_i}(\boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)})^j e_j = \sum_{j\in J_i}(\boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_{n-1})^j e_j. \tag{10.2}$$

So it follows that

$$\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}), \boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle - \frac{L}{2} \|\boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)}\|^2$$

$$\leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \sum_{j \in J_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + \frac{L}{2} b_n^2 + L' w_n b_n + L b_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|.$$

Also, by Cauchy-Schwarz inequality and L-Lipschitz continuity of $\nabla \bar{g}_n$, we have

$$\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n^{(i-1)}), \boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle \geq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle - c_1 w_n^2$$

for some constant $c_1 > 0$. Thus combining the above inequalities, we obtain

$$\left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)} \right\rangle \leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \sum_{j \in J_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + c_1 w_n^2.$$

Note that (10.2) implies

$$\sum_{i=1}^m \boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_n^{(i-1)} = \sum_{i=1}^m \sum_{j \in J_i(n)} (\boldsymbol{\theta}_n^{(i\star)} - \boldsymbol{\theta}_{n-1})^j e_j = \boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}.$$

Hence summing over $i = 1, \ldots, m$ and taking the conditional expectation with respect to $\mathcal{F}_{n-1}$,

$$\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1} \rangle \leq \left\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \sum_{i=1}^m \sum_{j \in J_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1})^j e_j \right\rangle + m c_1 w_n^2.$$

By using (9.5), taking the conditional expectation with respect to $\mathcal{F}_{n-1}$ on both sides of (10.3), and taking supremum for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \bar{c} \left\langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \right\rangle \tag{10.3}$$

$$\leq \mathbb{E}\left[\langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1} \rangle \,\Big|\, \mathcal{F}_{n-1} \right] + c_1' w_n^2.$$

Then combining the above with (9.3) finishes the proof. ∎

We first show a sufficient condition for a subsequence $(\boldsymbol{\theta}_{t_k})_{k \geq 1}$ to converge to a stationary point of a limiting averaged surrogate function.

Proposition 31 *Assume (A1)-(A4), (A7), and the case C2 in Theorem 2. Assume $r_{n+1}/r_n = O(1)$. Suppose there exists a sequence $(t_k)_{k \geq 1}$ in $\mathbb{N}$ such that almost surely, either*

$$\sum_{k=1}^\infty \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| = \infty \text{ or } \liminf_{k \to \infty} \mathbb{E}\left[\left| \left\langle \nabla \bar{g}_{t_k+1}(\boldsymbol{\theta}_{t_k}), \frac{\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}}{\|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\|} \right\rangle \right| \,\Big|\, \mathcal{F}_{t_k} \right] = 0. \tag{10.4}$$

Then there exists a further subsequence $(s_k)_{k \geq 1}$ of $(t_k)_{k \geq 1}$ such that $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{s_k}$ and $\boldsymbol{\theta}_\infty := \lim_{k \to \infty} \boldsymbol{\theta}_{s_k}$ exist almost surely and $\boldsymbol{\theta}_\infty$ and is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$.

Proof By Proposition 27, we have that for all $i = 1, \ldots, m$, almost surely,

$$\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| \, \mathbb{E}\left[\sum_{i=1}^{m} \left|\left\langle \nabla \bar{g}_{t_k+1}(\boldsymbol{\theta}_{t_k}), \frac{\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}}{\|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\|} \right\rangle\right| \, \Big| \, \mathcal{F}_{t_k}\right] < \infty.$$

Thus the former condition in the assertion implies the latter almost surely. Thus it suffices to show that this latter condition implies the assertion. Assume this condition, and let $(s_k)_{k \geq 1}$ be a subsequence of $(t_k)_{k \geq 1}$ for which the liminf in (10.4) is achieved. Using (A7), we may take a subsequence along which the following limits $\boldsymbol{\theta}_\infty = \lim_{k \to \infty} \boldsymbol{\theta}_{s_k}$, $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{s_k}$, and $\nabla \bar{g}_\infty := \lim_{k \to \infty} \nabla \bar{g}_{s_k}$ exist.

Now suppose for contradiction that $\boldsymbol{\theta}_\infty$ is not a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$. Then there exists $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ and $\delta > 0$ such that

$$\left\langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}' - \boldsymbol{\theta}_\infty \right\rangle < -\delta < 0.$$

Denote $\boldsymbol{\theta}^t := t\boldsymbol{\theta}' + (1-t)\boldsymbol{\theta}_\infty$ for $t \in [0, 1]$. Then

$$\langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^t - \boldsymbol{\theta}_\infty \rangle = t \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}' - \boldsymbol{\theta}_\infty \rangle < t\, \delta < 0 \quad \text{for all } t \in [0, 1].$$

Choose t sufficiently small so that $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\infty\| < 1/2$ and denote $\boldsymbol{\theta}^*$ for such $\boldsymbol{\theta}^t$. Note that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ by convexity of $\boldsymbol{\Theta}$. Necessarily $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_\infty$. By the triangle inequality,

$$\|\langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{s_k} \rangle - \langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_\infty \rangle\|$$
$$\leq \|\nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}) - \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty)\| \cdot \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_{s_k}\| + \|\nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty)\| \cdot \|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{s_k}\|.$$

By the choice of the subsequence s_k, we see that the right-hand side goes to zero as $k \to \infty$. Hence for all sufficiently large $k \geq 1$, we have

$$\langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{s_k} \rangle < -\delta/2.$$

Note that by Lemma 17, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n)$. Hence by Lemma 30, there is a constant $c > 0$ such that for al $n \geq 1$,

$$\frac{\|\boldsymbol{\theta}^\star_{s_k+1} - \boldsymbol{\theta}_{s_k}\|}{w_{s_k}} \mathbb{E}\left[\left\langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}^\star_{s_k+1}), \frac{\boldsymbol{\theta}^\star_{s_k+1} - \boldsymbol{\theta}_{s_k}}{\|\boldsymbol{\theta}^\star_{s_k+1} - \boldsymbol{\theta}_{s_k}\|} \right\rangle \Big| \mathcal{F}_{s_k}\right]$$
$$\leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{s_k}\| \leq 1} \langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{s_k} \rangle + c\frac{r^2_{s_k+1}}{r_{s_k}}.$$

Since $\|\boldsymbol{\theta}^\star_{n+1} - \boldsymbol{\theta}_n\| \leq w_{n+1}$, by using the hypothesis, the left-hand side above converges to zero almost surely as $k \to \infty$. All terms on the right-hand side above except the first one vanish as $k \to \infty$. Also note that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{s_k}\| \leq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_\infty\| + \|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{s_k}\| < 1$ for all k sufficiently large. Hence we obtain

$$0 \leq \liminf_{k \to \infty} \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{s_k}\| \leq 1} \langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{s_k} \rangle \leq \liminf_{k \to \infty} \langle \nabla \bar{g}_{s_k+1}(\boldsymbol{\theta}_{s_k}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{s_k} \rangle$$
$$= \langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^* - \boldsymbol{\theta}_\infty \rangle < 0,$$

which is a contradiction. This shows the assertion. ∎

Proposition 32 *Assume (A1)-(A4), (A7), and the case C2 in Theorem 2. If $(t_k)_{k \geq 1}$ is such that $\boldsymbol{\theta}_{t_k}$ is a short point for all $k \geq 1$, $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{t_k}$ and $\boldsymbol{\theta}_\infty := \lim_{k \to \infty} \boldsymbol{\theta}_{t_k}$ exist almost surely, then $\boldsymbol{\theta}_\infty$ is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$.*

Proof The argument is similar to that of (Bertsekas, 1997, Prop. 2.7.1). However, here we do not need to assume the uniqueness of solutions to minimization problems of $\bar{g}_t$ in each block coordinate due to the added search radius restriction, and also note that we are considering a random choice of block coordinates in each application of Algorithm 3.

Recall that $J_1(n), \ldots, J_m(n)$ denotes the coordinate blocks chosen by Algorithm 3 at iteration n of Algorithm 1, and by (A6), their joint distribution does not depend on n and also they are independent of everything else. Fix a coordinate block $J \in \mathbb{J}$ that has a positive probability of being chosen by $J_1(n)$. Then $J_1(t_k) = J$ for infinitely many k's almost surely by a Borel-Cantelli lemma. We may refine the subsequence $(t_k)_{k \geq 1}$ so that $J_1(t_k) = J$ for all $k \geq 1$ almost surely. For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we write $\boldsymbol{\theta} = [\theta^J, \theta^{J^c}]$, where $\boldsymbol{\theta}^J$ is the projection of $\boldsymbol{\theta}$ onto the coordinate block J and similarly for the complementary coordinate block J^c. In this fashion, we denote $\boldsymbol{\theta}_n = [\theta_n^J, \theta_n^{J^c}]$ and $\boldsymbol{\theta}_\infty = [\theta_\infty^J, \theta_\infty^{J^c}]$. Recall the block update $\boldsymbol{\theta}_{t_k-1} \to \boldsymbol{\theta}_{t_k}^{(1)} \to \cdots \to \boldsymbol{\theta}_{t_k}^{(m)} = \boldsymbol{\theta}_{t_k}$ and denote $\boldsymbol{\theta}_{t_k;1} := \boldsymbol{\theta}_{t_k}^{(1)}$. Then since $\boldsymbol{\theta}_{t_k}$ is a short point,

$$\bar{g}_{t_k}\left(\theta_{t_k;1}^J, \theta_{t_k;1}^{J^c}\right) \leq \bar{g}_{t_k}\left(\theta, \theta_{t_k;1}^{J^c}\right)$$

for every $\theta \in \boldsymbol{\Theta}^J = \text{Proj}_{\mathbb{R}^J}(\boldsymbol{\Theta})$. Indeed, in this case, his is because $\theta_{t_k;1}^J$ is a minimizer of $\theta \mapsto \bar{g}_{t_k+1}(\theta, \theta_{t_k;1}^c)$ over $\boldsymbol{\Theta}^J$. Noting that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = O(w_n) = o(1)$, it follows that

$$\bar{g}_{t_k}\left(\theta_\infty^J, \theta_\infty^{J^c}\right) \leq \bar{g}_{t_k}\left(\theta, \theta_\infty^{J^c}\right)$$

for all $\theta \in \boldsymbol{\Theta}^J$. Since $\boldsymbol{\theta}^J$ is convex, it follows that

$$\left\langle \nabla_J \bar{g}_\infty(\boldsymbol{\theta}_\infty), \theta - \theta_\infty^J \right\rangle \geq 0 \qquad \text{for all } \theta \in \boldsymbol{\Theta}^J.$$

By using a similar argument for the subsequent coordinate blocks $J_2(n), \ldots, J_m(n)$, one can show that the above holds for every coordinate block J that has a positive chance of being chosen by some of the blocks among $J_1(n), \ldots, J_m(n)$. The union of all such J covers the entire coordinates $\{1, \ldots, p\}$ by (A6). Hence we deduce that $\langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This shows the assertion. ∎

The following proposition gives a key property of a hypothetical non-stationary limit point of the iterates $(\boldsymbol{\theta}_n)_{n \geq 1}$, if exists.

Proposition 33 *Assume (A1)-(A4), (A7), and the case C2 in Theorem 2. Suppose there exists a sequence $(n_k)_{k \geq 1}$ such that $\bar{g}_\infty := \lim_{k \to \infty} \bar{g}_{n_k}$ and $\boldsymbol{\theta}_\infty := \lim_{k \to \infty} \boldsymbol{\theta}_{n_k}$ exist almost surely and $\boldsymbol{\theta}_\infty$ is not a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$. Then there exists $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| < \varepsilon\}$ with the following properties:*

(a) *$B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any stationary points of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$;*

(b) *There exists infinitely many $\boldsymbol{\theta}_t$'s outside of $B_\varepsilon(\boldsymbol{\theta}_\infty)$.*

Proof We will first show that there exists an ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty)$ of $\boldsymbol{\theta}_\infty$ that does not contain any short points of Λ. Suppose for contradiction that for each $\varepsilon > 0$, there exists a short point Λ in $B_\varepsilon(\boldsymbol{\theta}_\infty)$. Then one can construct a sequence of short points converging to $\boldsymbol{\theta}_\infty$. But then by Proposition 32, $\boldsymbol{\theta}_\infty$ is a stationary point, a contradiction.

Next, we show that there exists $\varepsilon > 0$ such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies **(a)**. Suppose for contradiction that there exists no such $\varepsilon > 0$. Then we have a sequence $(\boldsymbol{\theta}_{\infty;k})_{k\geq 1}$ of stationary points of Λ that converges to $\boldsymbol{\theta}_\infty$. Denote the limiting surrogate loss function associated with $\boldsymbol{\theta}_{\infty;k}$ by $\bar{g}_{\infty;k}$. Recall that each $\bar{g}_{\infty;k}$ is parameterized by elements in a compact set according to (A7). Hence by choosing a subsequence, we may assume that $\bar{g}_\infty := \lim_{k\to\infty} \bar{g}_{\infty;k}$ is well-defined. Fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ note that by Cauchy-Schwarz inequality,

$$\langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq -\|\nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty) - \nabla \bar{g}_{\infty;k}(\boldsymbol{\theta}_{\infty;k})\| \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\|$$
$$- \|\nabla \bar{g}_{\infty;k}(\boldsymbol{\theta}_{\infty;k})\| \cdot \|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{\infty;k}\| + \langle \nabla \bar{g}_{\infty;k}(\boldsymbol{\theta}_{\infty;k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{\infty;k} \rangle.$$

Note that $\langle \nabla \bar{g}_{\infty;k}(\boldsymbol{\theta}_{\infty;k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{\infty;k} \rangle \geq 0$ since $\boldsymbol{\theta}_{\infty;k}$ is a stationary point of $\bar{g}_{\infty;k}$ over $\boldsymbol{\Theta}$. Hence by taking $k \to \infty$, this shows $\langle \nabla \bar{g}_\infty(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$. Since $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ was arbitrary, this shows that $\boldsymbol{\theta}_\infty$ is a stationary point of $\bar{g}_\infty$ over $\boldsymbol{\Theta}$, a contradiction.

Lastly, from the earlier results, we can choose $\varepsilon > 0$ such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ has no short points of Λ and also satisfies **(b)**. We will show that $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies **(b)**. Then $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies **(a)**-**(b)**, as desired. Suppose for contradiction there are only finitely many $\boldsymbol{\theta}_t$'s outside of $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$. Then there exists an integer $M \geq 1$ such that $\boldsymbol{\theta}_t \in B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ for all $t \geq M$. Then each $\boldsymbol{\theta}_n$ for $n \geq M$ is a short point of Λ. By definition, it follows that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \geq w_n$ for all $t \geq M$. This implies $\alpha_n = \sqrt{\sum_{i=1}^n \|\boldsymbol{\theta}_n^{(i)} - \boldsymbol{\theta}_{n-1}\|^2} \geq \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \geq w_{n+1}$ for all $n \geq M$. Then by Proposition 31, since $\sum_{n=1}^\infty w_n = \infty$, there exists a subsequence $(s_k)_{k\geq 1}$ such that $\boldsymbol{\theta}'_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{t_k}$ exists and is stationary. But since $\boldsymbol{\theta}'_\infty \in B_\varepsilon(\boldsymbol{\theta})$, this contradicts **(a)** for $B_\varepsilon(\boldsymbol{\theta})$. This completes the proof. ∎

We are now ready to give proof of Lemma 23 for the diminishing radius case.

Proof of Lemma 23 for cases C1-C2 in Theorem 2. Assume (A1)-(A4), (A7), and the case C2 in Theorem 2. (As mentioned at the beginning of this subsection, C1 is a special case of C2.) Suppose there is a non-stationary limit point $\boldsymbol{\theta}_\infty$ of Λ. By Proposition 33, we may choose $\varepsilon > 0$ such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies the conditions **(a)**-**(b)** of Proposition 33. Choose $M \geq 1$ large enough so that $w_t < \varepsilon/4$ whenever $t \geq M$. We call an integer interval $I := [\ell, \ell']$ a *crossing* if $\boldsymbol{\theta}_\ell \in B_{\varepsilon/3}(\boldsymbol{\theta}_\infty)$, $\boldsymbol{\theta}_{\ell'} \notin B_{2\varepsilon/3}(\boldsymbol{\theta}_\infty)$, and no proper subset of I satisfies both of these conditions. By definition, two distinct crossings have empty intersections. Fix a crossing $I = [\ell, \ell']$, it follows that by triangle inequality,

$$\sum_{t=\ell}^{\ell'-1} \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\| \geq \|\boldsymbol{\theta}_{\ell'} - \boldsymbol{\theta}_\ell\| \geq \varepsilon/3. \tag{10.5}$$

Note that since $\boldsymbol{\theta}_\infty$ is a limit point of Λ, $\boldsymbol{\theta}_t$ visits $B_{\varepsilon/3}(\boldsymbol{\theta}_\infty)$ infinitely often. Moreover, by condition **(a)** of Proposition 33, $\boldsymbol{\theta}_t$ also exits $B_\varepsilon(\boldsymbol{\theta}_\infty)$ infinitely often. It follows that there are infinitely many crossings. Let t_k denote the kth smallest integer that appears in some

crossing. Then $t_k \to \infty$ as $k \to \infty$, and by (10.5),

$$\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| \geq (\# \text{ of crossings}) \frac{\varepsilon}{3} = \infty.$$

Then by Proposition 31, there is a further subsequence $(s_k)_{k \geq 1}$ of $(t_k)_{k \geq 1}$ such that $\boldsymbol{\theta}'_{\infty} := \lim_{k \to \infty} \boldsymbol{\theta}_{s_k}$ exists and is stationary. However, since $\boldsymbol{\theta}_{t_k} \in B_{2\varepsilon/3}(\boldsymbol{\theta}_{\infty})$, we have $\boldsymbol{\theta}'_{\infty} \in B_{\varepsilon}(\boldsymbol{\theta}_{\infty})$. This contradicts the condition **(b)** of Proposition 33 for $B_{\varepsilon}(\boldsymbol{\theta}_{\infty})$ that it cannot contain any stationary point of Λ. This shows the assertion. ∎

Acknowledgements

References

Ahmet Alacaoglu and Hanbaek Lyu. Convergence of stochastic first-order methods with dependent data for constrained nonconvex optimization. *International Conference on Machine Learning*, 2023.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

Léon Bottou. Online algorithms and stochastic approxima-p tions. *Online learning and neural networks*, 1998.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Aris Daniilidis and Jérôme Malick. Filling the gap between lower-c1 and lower-c2 functions. *Journal of Convex Analysis*, 12(2):315–329, 2005.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Aaron Defazio. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis. *arXiv preprint arXiv:2010.00406*, 2020.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Roy Frostig, Rong Ge, Sham Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning*, pages 2540–2548. PMLR, 2015.

Sara A Geer, Sara van de Geer, and D Williams. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Charles J Geyer. On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010, 1994.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. 1970.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163(1):85–114, 2017.

Reiner Horst and Nguyen V Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Bjorn Johansson, Maben Rabi, and Mikael Johansson. A simple peer-to-peer algorithm for distributed optimization in sensor networks. In *2007 46th IEEE Conference on Decision and Control*, pages 4705–4710. IEEE, 2007.

Björn Johansson, Maben Rabi, and Mikael Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20 (3):1157–1170, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Tamara G Kolda and David Hong. Stochastic gradients for large-scale tensor decomposition. *SIAM Journal on Mathematics of Data Science*, 2(4):1066–1095, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2007.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015.

Hanbaek Lyu and Yuchen Li. Block majorization-minimization with diminishing radius for constrained nonconvex optimization. *arXiv preprint arXiv:2012.03503*, 2023.

Hanbaek Lyu, Deanna Needell, and Laura Balzano. Online matrix factorization for markovian data and applications to network dictionary learning. *Journal of Machine Learning Research*, 21(251): 1–49, 2020.

Hanbaek Lyu, Christopher Strohmeier, and Deanna Needell. Online nonnegative cp-dictionary learning for markovian data. *The Journal of Machine Learning Research*, 23(1):6630–6679, 2022.

Hanbaek Lyu, Facundo Memoli, and David Sivakoff. Sampling random graph homomorphisms and applications to network data analysis. *Journal of Machine Learning Research*, 24:1–79, 2023.

Hanbaek Lyu, Yacoub H Kureh, Joshua Vendrow, and Mason A Porter. Learning low-rank latent mesoscale structures in networks. *Nature Communications*, 15(1):224, 2024.

Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013a.

Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013b.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

Kurt Marti. *Stochastic optimization methods*, volume 2. Springer, 2005.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2017.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Yu Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1 (3):127–239, 2014.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Johannes Schneider and Scott Kirkpatrick. *Stochastic optimization*. Springer Science & Business Media, 2007.

Leonard A Stefanski and Dennis D Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. In *Advances in Neural Information Processing Systems*, pages 9896–9905, 2018.

Tao Sun, Yuejiao Sun, Yangyang Xu, and Wotao Yin. Markov chain block coordinate descent. *Computational Optimization and Applications*, 75(1):35–61, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of Facebook networks. *Physica A*, 391(16):4165–4180, 2012.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966.

Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR, 2017.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR, 2019.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Yi Xu, Zhuoning Yuan, Sen Yang, Rong Jin, and Tianbao Yang. On the convergence of (stochastic) gradient descent with extrapolation for non-convex optimization. *arXiv preprint arXiv:1901.10682*, 2019.

Ya-xiang Yuan. Recent advances in trust region algorithms. *Mathematical Programming*, 151(1): 249–281, 2015.

Renbo Zhao, Vincent Tan, and Huan Xu. Online nonnegative matrix factorization with general divergences. In *Artificial Intelligence and Statistics*, pages 37–45, 2017.

Appendix A. Background on Markov chains and MCMC

A.1 Markov chains

Here, we provide a brief overview of Markov chains on a countable state space (refer to, for example, Levin and Peres (2017)). Let Ω be a countable set. A function $P : \Omega^2 \to [0, \infty)$ is called a *Markov transition matrix* if each row of P sums to 1. A sequence of random variables $(X_t)_{t \geq 0}$ taking values in Ω is called a *Markov chain* with transition matrix P if, for any $x_0, x_1, \ldots, x_n \in \Omega$,

$$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) = P(x_{n-1}, x_n). \quad \text{(A.1)}$$

A probability distribution π on Ω is called a *stationary distribution* for the Markov chain $(X_t)_{t \geq 0}$ if $\pi = \pi P$, meaning

$$\pi(x) = \sum_{y \in \Omega} \pi(y) P(y, x).$$

The chain $(X_t)_{t \geq 0}$ is called *irreducible* if, for any $x, y \in \Omega$, there exists an integer $t \geq 0$ such that $P^t(x, y) > 0$. For each $x \in \Omega$, let $\mathcal{T}(x) = \{t \geq 1 \mid P^t(x, x) > 0\}$ denote the set of times at which the chain can return to x. The *period* of x is the greatest common divisor of $\mathcal{T}(x)$, and the chain is called *aperiodic* if all states have period 1. A chain is *positive recurrent* if, for some $x \in \Omega$, the expected time to return to x (starting from x) is finite. An irreducible, aperiodic Markov chain has a unique stationary distribution if and only if it is positive recurrent (Levin and Peres, 2017, Thm 21.21).

Given two probability distributions μ and ν on Ω, the *total variation distance* between them is defined as

$$\|\mu - \nu\|_{TV} = \sup_{A \subseteq \Omega} |\mu(A) - \nu(A)|.$$

If a Markov chain $(X_t)_{t \geq 0}$ starts at $x_0 \in \Omega$, then by (A.1), the distribution of X_t is $P^t(x_0, \cdot)$. If the chain is irreducible and aperiodic, and has a stationary distribution π, the convergence theorem (see, e.g., (Levin and Peres, 2017, Thm 21.14)) states that as $t \to \infty$, the distribution of X_t converges to π in total variation distance:

$$\sup_{x_0 \in \Omega} \|P^t(x_0, \cdot) - \pi\|_{TV} \to 0.$$

In the case of a finite state space, this convergence is exponential in t (see (Levin and Peres, 2017, Thm 4.9)), meaning there are constants $\lambda \in (0, 1)$ and $C > 0$ such that for all $t \geq 0$,

$$\max_{x_0 \in \Omega} \|P^t(x_0, \cdot) - \pi\|_{\text{TV}} \leq C \lambda^t. \quad \text{(A.2)}$$

A.2 Markov chain Monte Carlo Sampling

Suppose we have a finite sample space Ω and probability distribution π on it. We would like to sample a random element $\omega \in \Omega$ according to the distribution π. *Markov chain Monte Carlo (MCMC)* is a sampling algorithm that leverages the properties of Markov chains we mentioned in Subsection A.1. Namely, suppose that we have found a Markov chain $(X_t)_{t \geq 0}$ on state space Ω that is irreducible, aperiodic[2], and has π as its unique stationary distribution. Denote its transition matrix as P. Then

---

2. Aperiodicity can be easily obtained by making a given Markov chain lazy, that is, adding a small probability ε of staying at the current state. Note that this is the same as replacing the transition matrix P by $P_\varepsilon := (1 - \varepsilon)P + \varepsilon I$ for some $\varepsilon > 0$. This 'lazyfication' does not change stationary distributions, as $\pi P = \pi$ implies $\pi P_\varepsilon = \pi$.

by (A.2), for any $\varepsilon > 0$, one can find a constant $\tau = \tau(\varepsilon) = O(\log \varepsilon^{-1})$ such that the conditional distribution of $X_{t+\tau}$ given X_t is within total variation distance ε from π regardless of the distribution of X_t. Recall such $\tau = \tau(\varepsilon)$ is called the *mixing time* of the Markov chain $(X_t)_{t \geq 1}$. Then if one samples a long Markov chain trajectory $(X_t)_{t \geq 1}$, the subsequence $(X_{k\tau})_{k \geq 1}$ gives approximate i.i.d. samples from π.

We can further compute how far the thinned sequence $(X_{k\tau})_{k \geq 1}$ is away from being independent. Namely, observe that for any two nonempty subsets $A, B \subseteq \Omega$,

$$
\begin{aligned}
|\mathbb{P}(X_{k\tau} &\in A, \, X_\tau \in B) - \mathbb{P}(X_{k\tau} \in A)\mathbb{P}(X_\tau \in B)| \\
&= |\mathbb{P}(X_{k\tau} \in A) - \mathbb{P}(X_{k\tau} \in A \,|\, X_\tau \in B)| \, |\mathbb{P}(X_\tau \in B)| \\
&\leq |\mathbb{P}(X_{k\tau} \in A) - \mathbb{P}(X_{k\tau} \in A \,|\, X_\tau \in B)| \\
&\leq |\mathbb{P}(X_{k\tau} \in A) - \pi(A)| + |\pi(A) - \mathbb{P}(X_{k\tau} \in A \,|\, X_\tau \in B)| \leq \lambda^{k\tau} + \lambda^{(k-1)\tau}.
\end{aligned}
$$

Hence the correlation between $X_{k\tau}$ and X_τ is $O(\lambda^{(k-1)\tau})$.

For the lower bound, let us assume that X_t is *reversible* with respect to π, that is, $\pi(x)P(x,y) = \pi(y)P(y,x)$ for $x, y \in \Omega$ (e.g., random walk on graphs). Then $\tau(\varepsilon) = \Theta(\log \varepsilon^{-1})$ (see (Levin and Peres, 2017, Thm. 12.5)), which yields $\sup_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV} = \Theta(\lambda^t)$. Also, $\mathbb{P}(X_\tau \in B) > \delta > 0$ for some $\delta > 0$ whenever τ is large enough under the hypothesis. Hence

$$
\begin{aligned}
|\mathbb{P}(X_{k\tau} &\in A, \, X_\tau \in B) - \mathbb{P}(X_{k\tau} \in A)\mathbb{P}(X_\tau \in B)| \\
&\geq \delta^{-1} \, |\mathbb{P}(X_{k\tau} \in A) - \mathbb{P}(X_{k\tau} \in A \,|\, X_\tau \in B)| \\
&\geq \big| \, |\mathbb{P}(X_{k\tau} \in A) - \pi(A)| - |\pi(A) - \mathbb{P}(X_{k\tau} \in A \,|\, X_\tau \in B)| \, \big| \geq c\lambda^{(k-1)\tau}
\end{aligned}
$$

for some constant $c > 0$. Hence the correlation between $X_{k\tau}$ and X_τ is $\Theta(\lambda^{(k-1)\tau})$. In particular, the correlation between two consecutive terms in $(X_{k\tau})_{k \geq 1}$ is of $\Theta(\lambda^\tau) = \Theta(\varepsilon)$. Thus, we can make the thinned sequence $(X_{k\tau})_{k \geq 1}$ arbitrarily close to being i.i.d. for π, but if X_t is reversible with respect to π, the correlation within the thinned sequence is always nonzero.

In practice, one may not know how to estimate the mixing time $\tau = \tau(\varepsilon)$. In order to empirically assess that the Markov chain has mixed to the stationary distribution, multiple chains are run for diverse mode exploration, and their empirical distribution is compared to the stationary distribution (a.k.a. multi-start heuristic). See Brooks et al. (2011) for more details on MCMC sampling.

Appendix B. Example of Surrogate functions

In this section, we list some examples of block-convex surrogate functions, which include the usual convex surrogate functions in the literature (see, Mairal (2013a,b)).

Example 1 (Proximal surrogates for L-smooth functions) *Suppose f is L-smooth, that is, ∇f is L-Lipscthiz continuous. Then f is L-weakly convex, that is, $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta}\|^2$ is convex (see Lemma 35). Hence for each $\rho \geq L$, the following function g belongs to $\mathtt{Srg}_{L+\rho, \rho-L}(f, \boldsymbol{\theta}^\star)$:*

$$
g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2.
$$

Indeed, $g \geq f$, $g(\boldsymbol{\theta}^\star) = f(\boldsymbol{\theta}^\star)$, $\nabla g(\boldsymbol{\theta}^\star) = \nabla f(\boldsymbol{\theta}^\star)$, and ∇g is $(L+\rho)$-Lipschtiz. Moreover, g is convex being some of two convex functions:

$$
f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 = \left(f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta}\|^2 \right) + \left(-L\langle \boldsymbol{\theta}, \boldsymbol{\theta}^\star \rangle + \frac{L}{2}\|\boldsymbol{\theta}^\star\|^2 \right).
$$

Minimizing the above function over $\boldsymbol{\Theta}$ is equivalent to applying a proximal mapping of f, where the resulting estimate is denoted as $\mathrm{prox}_{f/\rho}(\boldsymbol{\theta}^\star)$ (see Parikh and Boyd (2014); Davis and Drusvyatskiy (2019)). ▲

74

Example 2 (Proximal modification of weakly convex surrogates) *One can also use proximal mapping to convert block weakly convex surrogates to multi-convex surrogates. Namely, suppose $g \in \mathtt{Srg}_{L,-\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^\star)$ is a surrogate of f at $\boldsymbol{\theta}^\star$ that is ρ-weakly convex in each block $J \in \mathbb{J}$. That is, $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta}\|$ is convex in $\boldsymbol{\Theta}^J$ for each $J \in \mathbb{J}$. In this case, g is nonconvex in each block coordinate so it cannot be directly block-minimized. In this case, we can add a quadratic term to make it a multi-convex surrogate. That is, for each $\rho' \geq \rho$, the following function $\tilde{g}$ belongs to $\mathtt{Srg}_{L+\rho',\rho'-\rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^\star)$:*

$$\tilde{g} : \boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}) + \frac{\rho'}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2,$$

which can be easily verified by expanding out the quadratic term. In the slightly more general case when the modulus of weak convexity of g depends on the coordinate block J, we may add block-dependent quadratic terms to reduce the amount of proximal modification. ▲

Example 3 (Prox-linear surrogates) If f is L-smooth, then the following quadratic function g belongs to $\mathtt{Srg}_{2L,L}(f, \boldsymbol{\theta}^\star)$:

$$g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}^\star) + \langle \nabla f(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2.$$

(See Lemma 34 in Appendix C.) If in addition f is convex, then $g \in \mathtt{Srg}_{L,L}(f, \boldsymbol{\theta}^\star)$; If f is μ-strongly convex, then $g \in \mathtt{Srg}_{L-\mu,L}(f, \boldsymbol{\theta}^\star)$. ▲

Example 4 (Prox-linear surrogates) *Suppose $f = f_1 + f_2$ where f_1 is differentiable with L-Lipschitz gradient and f_2 is convex over $\boldsymbol{\Theta}$. Then the following function g belongs to $\mathtt{Srg}_{2L,L}(f, \boldsymbol{\theta}^\star)$:*

$$g : \boldsymbol{\theta} \mapsto f_1(\boldsymbol{\theta}^\star) + \langle \nabla f_1(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 + f_2(\boldsymbol{\theta}).$$

Minimizing g over $\boldsymbol{\Theta}$ amounts to performing a proximal gradient step in Beck and Teboulle (2009); Nesterov (2013). ▲

Example 5 (DC programming surrogates) *Suppose $f = f_1 + f_2$ where f_1 is convex and f_2 is concave and differentiable with L_2-Lipschitz gradient over $\boldsymbol{\Theta}$. One can also write $f = f_1 - (-f_2)$ which is the difference of convex (DC) functions f_1 and $-f_2$. Then the following function g belongs to $\mathtt{Srg}_{2L,L}(f, \boldsymbol{\theta}^\star)$:*

$$g : \boldsymbol{\theta} \mapsto f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\theta}^\star) + \langle \nabla f_2(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle.$$

Such surrogates are important in DC programming (see, e.g., Horst and Thoai (1999)).

In fact, our method allows f_1 to be only multi-convex with respect to the coordinate blocks in $\mathbb{J}$, that is, f_1 is convex on $\mathrm{Proj}_{\mathbb{R}^J}(\boldsymbol{\Theta})$ for each coordinate block $J \in \mathbb{J}$. In this case, g above is a multi-convex surrogate of f at $\boldsymbol{\theta}^\star$ and belongs to $\mathtt{Srg}_{2L,L}^{\mathbb{J}}(f, \boldsymbol{\theta}^\star)$. ▲

Example 6 (Convex Variational Surrogates) Let $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ be a two-block multi-convex function and let $\Theta_1 \subseteq \mathbb{R}^p$ and $\Theta_2 \subseteq \mathbb{R}^q$ be two convex sets. Define a function $f_* : \boldsymbol{\theta} \mapsto \inf_{H \in \Theta_2} f(\boldsymbol{\theta}, H)$. Then for each $\boldsymbol{\theta}^\star \in \Theta$, the following function

$$g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}, H^\star), \quad H^\star \in \underset{H \in \boldsymbol{\Theta}_2}{\arg\min} f(\boldsymbol{\theta}^\star, H)$$

is convex over Θ_1 and satisfies $g \geq f$ and $g(\boldsymbol{\theta}^\star) = f(\boldsymbol{\theta}^\star)$. Further, assume that

(i) $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}, H)$ is differentiable for all $H \in \Theta_2$ and $\boldsymbol{\theta} \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ is L'-Lipschitz for all $H \in \Theta_2$;

(ii) $H \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ is L-Lipschitz for all $\boldsymbol{\theta} \in \Theta_1$;

(iii) $H \mapsto f(\boldsymbol{\theta}, H)$ is ρ-strongly convex for all $\boldsymbol{\theta} \in \mathbb{R}^p$.

Then g belongs to $\mathtt{Srg}_{L'', \rho}(f_*, \boldsymbol{\theta}^{\star})$ for some $L'' > 0$. When f is jointly convex, then f_* is also convex and we can choose $L'' = L$. ▲

Example 7 (Multi-convex Variational Surrogates) *Let $f : \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_m} \times \mathbb{R}^{p_{m+1}} \to \mathbb{R}$ be a $(m+1)$-block multi-convex function and let $\Theta_1 \subseteq \mathbb{R}^p$ and $\Theta_i \subseteq \mathbb{R}^{p_i}$ for $i = 1, \ldots, m+1$ be convex sets. Denote $\boldsymbol{\Theta} := \Theta_1 \times \cdots \times \Theta_m$. Define a function $f_* : \boldsymbol{\theta} \mapsto \inf_{H \in \Theta_{p_{m+1}}} f(\boldsymbol{\theta}, H)$. Then for each $\boldsymbol{\theta}^{\star} \in \boldsymbol{\Theta}$, the following function*

$$g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}, H^{\star}), \quad H^{\star} \in \underset{H \in \boldsymbol{\Theta}_{p_{m+1}}}{\arg\min} \ f(\boldsymbol{\theta}^{\star}, H)$$

is m-block multi-convex over $\boldsymbol{\Theta}$ and satisfies $g \geq f$ and $g(\boldsymbol{\theta}^{\star}) = f(\boldsymbol{\theta}^{\star})$. Further, assume that

(i) $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}, H)$ *is differentiable for all $H \in \Theta_{p_{m+1}}$ and $\boldsymbol{\theta} \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ is L'-Lipschitz for all $H \in \Theta_{p_{m+1}}$* ;

(ii) $H \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ *is L-Lipschitz for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$;*

(iii) $H \mapsto f(\boldsymbol{\theta}, H)$ *is μ-strongly convex for all $\boldsymbol{\theta} \in \mathbb{R}^{p_{m+1}}$.*

Let $\mathbb{J} = \{J_1, \ldots, J_m\}$, where J_i denotes the ith coordinate block corresponding to $\mathbb{R}^{p_i}$. Then g belongs to $\mathtt{Srg}_{L'', \mu}^{\mathbb{J}}(f_*, \boldsymbol{\theta}^{\star})$ for some $L'' > 0$. When f is jointly convex, then f_* is also convex and we can choose $L'' = L$. ▲

Example 8 (Refining block structure) *Suppose $f : \boldsymbol{\Theta} \subseteq \mathbb{R}^p \to \mathbb{R}$ with $\boldsymbol{\Theta}$ convex, and we have a convex surrogate $g \in \mathtt{Srg}_{L, \rho}(f, \boldsymbol{\theta}^{\star})$. Let $\mathbb{J}$ denote an arbitrary set of coordinate bloks for $\mathbb{R}^p$ (e.g., $\mathbb{J} = \{\{1\}, \ldots, \{p\}\}$). Then one can also view the fuction g as a multi-convex surrogate in $\mathtt{Srg}_{L, \rho}^{\mathbb{J}}(f_*, \boldsymbol{\theta}^{\star})$. Instead of minimizing g over the convex set $\boldsymbol{\Theta}$ with respect to all p coordinates, one can block-minimize g with respect to the block structure $\mathbb{J}$. For instance, if $\mathbb{J} = \{\{1\}, \ldots, \{p\}\}$ and if $J_i(n)$ in Algorithm 3 is chosen uniformly at random from $\mathbb{J}$, then Algorithm 3 becomes a random coordinate descent for m iterations, which choses a random coordinate $i \in \{1, \ldots, p\}$ and minimize g over $\boldsymbol{\Theta}$ with respect to the ith coordinate. (See, e.g., Wright (2015)).*

A similar remark holds for multi-convex surrogates. Namely, suppose $g \in \mathtt{Srg}_{L, \rho}^{\mathbb{J}}(f, \boldsymbol{\theta}^{\star})$, where $\mathbb{J} = \{J_1, \ldots, J_m\}$ denotes a set of coordinate blocks for $\mathbb{R}^p$. Let $\mathbb{J}'$ denote an arbitrary set of coordinate bloks for $\mathbb{R}^p$ that refines $\mathbb{J}$. Namely, each coordinate block $J \in \mathbb{J}'$ is a subset of some J_i, $i = 1, \ldots, m$. Then one can also view the fuction g as a multi-convex surrogate in $\mathtt{Srg}_{L'', \rho}^{\mathbb{J}'}(f_*, \boldsymbol{\theta}^{\star})$ and accordingly, one can block-minimize g using the finer block structure given by $\mathbb{J}'$ instead of the original block structure $\mathbb{J}$. ▲

See Mairal (2013a) for other types of surrogate functions such as quadratic, Jensen, and saddle point surrogates.

Appendix C. Auxiliary lemmas

Lemma 34 (Convex Surrogate for Functions with Lipschitz Gradient) *Let $f : \mathbb{R}^p \to \mathbb{R}$ be differentiable and ∇f be L-Lipschitz continuous. Then for each $\theta, \theta' \in \mathbb{R}^p$,*

$$|f(\theta') - f(\theta) - \langle \nabla f(\theta), \theta' - \theta \rangle| \leq \frac{L}{2} \|\theta - \theta'\|^2.$$

Proof This is a classical Lemma. See (Nesterov, 1998, Lem 1.2.3). ∎

Lemma 35 (Characterization of weak convexity) *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a smooth function. Fix a convex set $\mathbf{\Theta} \subseteq \mathbb{R}^p$ and $\rho > 0$. The following conditions are equivalent.*

(i) *(Weak convexity) $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta}\|^2$ is convex on $\mathbf{\Theta}$;*

(ii) *(Hypermonotonicity) $\langle \nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle \geq -\rho\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}$;*

(iii) *(Quadratic lower bound) $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \langle \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}$.*

Proof See (Daniilidis and Malick, 2005, Thm. 7) for an equivalent statement for a more general case of local Lipschitz functions.

(i)$\Rightarrow$**(iii)**: For $s \in [0, 1]$, we have

$$s\left(f(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta}\|^2\right) + (1-s)\left(f(\boldsymbol{\theta}') + \frac{\rho}{2}\|\boldsymbol{\theta}'\|^2\right) \geq f(s\boldsymbol{\theta} + (1-s)\boldsymbol{\theta}') + \frac{\rho}{2}\|s\boldsymbol{\theta} + (1-s)\boldsymbol{\theta}'\|^2 \tag{C.1}$$

so we get

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \frac{f(\boldsymbol{\theta}' + s(\boldsymbol{\theta} - \boldsymbol{\theta}')) - f(\boldsymbol{\theta}')}{s} - \frac{\rho}{2}\left(\|\boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}'\|^2 - \frac{\|\boldsymbol{\theta}' + s(\boldsymbol{\theta} - \boldsymbol{\theta}')\|^2 - \|\boldsymbol{\theta}'\|^2}{s}\right)$$

Taking the limit $s \to 0$, we get

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \langle \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\rho}{2}\left(\|\boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}'\|^2 - 2\langle \boldsymbol{\theta}, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle\right).$$

This implies **(iii)**.

(iii)$\Rightarrow$**(ii)**: Adding the following two inequalities

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') \geq \langle \nabla f(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$$

$$f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) \geq \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle - \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2,$$

we obtain $0 \geq \langle \nabla f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle - \rho\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2$. This implies **(ii)**.

(ii)$\Rightarrow$**(i)**: Fix $s \in [0, 1]$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{\Theta}$. Denote $\boldsymbol{\theta}_s := s\boldsymbol{\theta} + (1-s)\boldsymbol{\theta}'$. Note that $\boldsymbol{\theta}_s - \boldsymbol{\theta}' = s(\boldsymbol{\theta} - \boldsymbol{\theta}')$ and $\boldsymbol{\theta}_s - \boldsymbol{\theta} = (1-s)(\boldsymbol{\theta}' - \boldsymbol{\theta})$. By the mean value theorem, there exists $s_* \in [s, 1]$ and $s'_* \in [0, s]$ such that

$$f(\boldsymbol{\theta}_s) - f(\boldsymbol{\theta}) = (1-s)\langle \nabla f(\boldsymbol{\theta}_{s_*}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle, \qquad f(\boldsymbol{\theta}_s) - f(\boldsymbol{\theta}') = s\langle \nabla f(\boldsymbol{\theta}_{s'_*}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle.$$

Multiplying by s and $1 - s$ respectively and adding the resulting equations, we get

$$f(\boldsymbol{\theta}_s) - sf(\boldsymbol{\theta}) - (1-s)f(\boldsymbol{\theta}') = -s(1-s)\langle \nabla f(\boldsymbol{\theta}_{s_*}) - \nabla f(\boldsymbol{\theta}_{s'_*}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle.$$

Note that since $s'_* \leq s \leq s_*$ and $\boldsymbol{\theta}_{s_*}, \boldsymbol{\theta}_{s'_*}$ are in the secant line between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, by **(ii)**,

$$\begin{aligned}
\langle \nabla f(\boldsymbol{\theta}_{s_*}) - \nabla f(\boldsymbol{\theta}_{s'_*}), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle &= \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|}{\|\boldsymbol{\theta}_{s^*} - \boldsymbol{\theta}_{s'_*}\|}\langle \nabla f(\boldsymbol{\theta}_{s_*}) - \nabla f(\boldsymbol{\theta}_{s'_*}), \boldsymbol{\theta}_{s_*} - \boldsymbol{\theta}_{s'_*} \rangle \\
&\geq -\rho\|\boldsymbol{\theta}_{s^*} - \boldsymbol{\theta}_{s'_*}\| \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \\
&\geq -\rho\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2.
\end{aligned}$$

It follows that

$$f(\boldsymbol{\theta}_s) - sf(\boldsymbol{\theta}) - (1-s)f(\boldsymbol{\theta}') \leq s(1-s)\rho\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2,$$

which is equivalent to (C.1).

Lemma 36 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function such that ∇f is L-Lipscthiz for some $L > 0$. Then f is L-weakly convex, that is, $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta}\|^2$ is convex.*

Proof Follows immediately by Lemmas 34 and 35. ∎

Lemma 37 (Second-Order Growth Property) *Let $f : \mathbb{R}^p \to [0, \infty)$ be μ-strongly convex and let Θ is a convex subset of $\mathbb{R}^p$. Let θ^* denote the minimizer of f over Θ. Then for all $\theta \in \Theta$,*

$$f(\theta) \geq f(\theta^*) + \frac{\mu}{2}\|\theta - \theta^*\|^2.$$

Proof See (Mairal, 2013b, Lem. A.4). ∎

Next, we provide some probabilistic lemmas.

Lemma 38 *Let $(X_n)_{n \geq 1}$ be a sequence of nonnegative random variables adapted to a filtration $(\mathcal{F}_n)_{n \geq 1}$ and $\mathbb{E}[X_n] < \infty$ for $n \geq 1$. Suppose there exists constants $\alpha \in [0, 1)$ and $M > 0$ such that for sufficiently large $n \geq 1$,*

$$\mathbb{E}\left[X_n \,|\, \mathcal{F}_{n-1}\right] \leq \alpha X_{n-1} + M. \tag{C.2}$$

Then $\lim_{n \to \infty} X_n$ exists and finite almost surely. In particular, $(X_n)_{n \geq 1}$ is bounded almost surely.

Proof Fix a constant $c > 0$ and let $\widetilde{X}_n := X_n + c$ for $n \geq 1$. Substituting for X_n in (C.2), we get

$$\mathbb{E}[\tilde{X}_n \,|\, \mathcal{F}_{n-1}] \leq \alpha \tilde{X}_{n-1} + (1 - \alpha)c + M$$

for large enough $n \geq 1$. Thus if we choose $c = -M/(1-\alpha)$, then $(\widetilde{X}_n)_{n \geq N}$ is a supermartingale with respect to the filtration $(\mathcal{F}_n)_{n \geq N}$ for some $N \geq 1$. Furthermore, $\widetilde{X}_n$ is uniformly lower bounded by $c = -M/(1 - \alpha)$ since $X_n \geq 0$ for $n \geq 1$. By the martingale convergence theorem (see, e.g., (Durrett, 2019, Thm. 4.2.11)), $\widetilde{X}_n$ converges almost surely to some random variable, say, $\widetilde{X}_\infty$, such that $\mathbb{E}[\widetilde{X}_\infty] \leq \mathbb{E}[\widetilde{X}_0] < \infty$. In particular, $\widetilde{X}_\infty < \infty$ almost surely. It follows that X_n converges to $\widetilde{X}_\infty - c$ almost surely. Lastly, note that

$$\mathbb{P}((X_n)_{n \geq 1} \text{ is not bounded }) \leq \mathbb{P}(X_n \text{ does not converge to } \widetilde{X}_\infty - c) = 0.$$

This shows the assertion. ∎

Lemma 39 *Let $(\boldsymbol{\theta}_n)_{n \geq 1}$ denote the output of Algorithm 1 with Algorithm 2 used for (3.1) with $\lambda_n \equiv 0$. Assume (A1), (A3), and (A5') hold. Suppose that the data sequence $(\mathbf{x}_n)_{n \geq 1}$ is contained in some compact subset $\mathfrak{X}_0 \subseteq \mathfrak{X}$. Then $\|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_n\| = O(w_n)$ and $\Delta_n = O(w_n^2)$ for $n \geq 1$. In particular, if $\sum_{n=1}^{\infty} w_n^2 < \infty$, then (A5) holds.*

Proof The argument here follows the one given in Mensch et al. (2017). Let $\boldsymbol{\theta}_n^\star$ be the exact minmizer of the ρ-strongly convex function $\bar{g}_n$ over $\boldsymbol{\Theta}$ for $n \geq 1$. We introduce the following random variables

$$A_n = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|, \quad B_n = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^\star\|, \quad C_n = \|\boldsymbol{\theta}_{n-1}^\star - \boldsymbol{\theta}_n^\star\|, \quad D_n = \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_n^\star).$$

By ρ-strong convexity of $\bar{g}_n$, almost surely, $\frac{\rho}{2}B_n^2 \leq D_n$. By the second-order growth property (Lemma 37) and using L'-Lipschitz continuity of g_n (see Lemma 16), almost surely,

$$\begin{aligned}
\frac{\rho}{2}C_n^2 &\leq \bar{g}_n(\boldsymbol{\theta}_{n-1}^\star) - \bar{g}_n(\boldsymbol{\theta}_n^\star) \\
&= (1 - w_n)\left(\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}^\star) - \bar{g}_{n-1}(\boldsymbol{\theta}_n^\star)\right) + w_n\left(g_n(\boldsymbol{\theta}_{n-1}^\star) - g_n(\boldsymbol{\theta}_n^\star)\right) \\
&\leq w_n\left(g_n(\boldsymbol{\theta}_{n-1}^\star) - g_n(\boldsymbol{\theta}_n^\star)\right) \leq w_n L' C_n.
\end{aligned}$$

This implise $C_n \leq c_1 w_n$ for all $n \geq 1$ almost surely for some constant $c_1 > 0$. By triangle inequality, we can write $A_n \leq B_n + B_{n-1} + C_n$, so by Cauchy-Schwarz inequality,

$$\begin{aligned}
A_n^2 &\leq (B_n + B_{n-1} + C_n)^2 \leq 3(B_n^2 + B_{n-1}^2 + C_n^2) \tag{C.3} \\
&\leq 3((2/\rho)(D_n + D_{n-1}) + (2L'/\rho)^2 w_n^2).
\end{aligned}$$

Next, we assume (A5') and show that $D_n = O(w^2)$. Denote $\widetilde{D}_n := D_n/w_n^2$. We will show that

$$\mathbb{E}\left[\widetilde{D}_n \mid \mathcal{F}_{n-1}\right] \leq \alpha \widetilde{D}_{n-1} + M \tag{C.4}$$

for all sufficiently large $n \geq 1$, where $\alpha \in [0, 1)$ and $M > 0$ are constants. Then by Lemma 38, it follows that $\widetilde{D}_n$ is uniformly bounded almost surely. This yields $D_n = O(w_n^2)$. The assertion then follows by combining this with (C.3).

It suffices to verify (C.4). First, by using (A5') and the previous inequalities involving A_n and B_n, we have

$$\begin{aligned}
\mathbb{E}\left[D_n \mid \mathcal{F}_{n-1}\right] &\leq (1 - \mu)\left(\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n^\star)\right) \\
&= (1 - \mu)w_n\left(g_n(\boldsymbol{\theta}_{n-1}) - g_n(\boldsymbol{\theta}_n) + g_n(\boldsymbol{\theta}_n) - g_n(\boldsymbol{\theta}_n^\star)\right) \\
&\quad + (1 - \mu)(1 - w_n)\left(\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}^\star)\right) \\
&\quad + (1 - \mu)(1 - w_n)\left(\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}^\star) - \bar{g}_{n-1}(\boldsymbol{\theta}_n^\star)\right) \\
&\leq (1 - \mu)L' w_n(A_n + B_n) + (1 - \mu)D_{n-1} \\
&\leq (1 - \mu)c_1 w_n\left(\sqrt{D_n + D_{n-1} + c_2 w_n^2} + \sqrt{D_n}\right) + (1 - \mu)D_{n-1},
\end{aligned}$$

where $c_1, c_2 > 0$ are constants. The second inequality above uses the fact that $\boldsymbol{\theta}_{n-1}^\star$ is an exact minimizer of $\bar{g}_{n-1}$ over $\boldsymbol{\Theta}$. Setting $\widetilde{D}_n := D_n/w_n^2$, this gives

$$\mathbb{E}\left[\widetilde{D}_n \mid \mathcal{F}_{n-1}\right] \leq (1 - \mu)c_1\left(\sqrt{\widetilde{D}_n + \widetilde{D}_{n-1}(w_{n-1}/w_n)^2 + c_2} + \sqrt{\widetilde{D}_n}\right) + (1 - \mu)\widetilde{D}_{n-1}(w_{n-1}/w_n)^2.$$

Fix $\varepsilon > 0$. Using the inequality $\sqrt{t} \leq \varepsilon t + \varepsilon^{-1}$ for all $t \geq 0$, this gives

$$\mathbb{E}\left[\widetilde{D}_n \mid \mathcal{F}_{n-1}\right] \leq c_3 \varepsilon \widetilde{D}_n + (1 - \mu)(1 + c_1 \varepsilon)(w_{n-1}/w_n)^2 \widetilde{D}_{n-1} + c_4 \varepsilon^{-1},$$

where $c_3, c_4 > 0$ are constants. Recall the hypothesis that $w_{n-1}/w_n \to 1$ as $n \to \infty$. Hence taking the expectation conditional on $\mathcal{F}_{n-1}$ and choosing $\varepsilon > 0$ sufficiently small, we verify (C.4), as desired. ∎

Lemma 40 *Fix a bounded and measurable function $\psi : \mathfrak{X} \times \boldsymbol{\Theta} \to \mathbb{R}$. Under Assumptions (A2) and (A3),*

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sqrt{n}\left|\mathbb{E}_{\mathbf{x} \sim \pi}\left[\psi(\mathbf{x}, \boldsymbol{\theta})\right] - \frac{1}{n}\sum_{k=1}^{n}\psi(\mathbf{x}_k, \boldsymbol{\theta})\right|\right] = O(1).$$

Furthermore, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\left|\mathbb{E}_{\mathbf{x} \sim \pi}\left[\psi(\mathbf{x}, \boldsymbol{\theta})\right] - \frac{1}{n}\sum_{k=1}^{n}\psi(\mathbf{x}_k, \boldsymbol{\theta})\right| \to 0$ almost surely as $t \to \infty$.

Proof Omitted. See (Lyu et al., 2020, Lem. 10) ∎

Lemma 41 *Fix a continuous and measurable function $\psi : \mathfrak{X} \times \boldsymbol{\Theta} \to \mathbb{R}$. Fix a sequence $(w_n)_{n \geq 1}$ in $(0, 1]$ and define functions $\bar{\psi}_n : \boldsymbol{\Theta} \to \mathbb{R}$ by*

$$\bar{\psi}_n(\boldsymbol{\theta}) = \sum_{k=1}^{n} \hat{w}_k^n \psi(\mathbf{x}_k, \boldsymbol{\theta}),$$

where $w_k^n := w_k \prod_{i=k+1}^{n}(1 - w_i)$. Suppose (A2)-(A3) hold. Assume that there exist an integer $T \geq 1$ such that $w_T \geq 1/2$ and $w_{n+1}^{-1} - w_n^{-1} \leq 1$ for all $n \geq T$. Suppose $w_n \geq cn^{-\gamma}$ for some constant $c > 0$ and $\gamma \in (0, 1]$ for all $n \geq 1$.

Then there exists a constant $C = C(T) > 0$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\left|\mathbb{E}_{\mathbf{x} \sim \pi}\left[\psi(\mathbf{x}, \boldsymbol{\theta})\right] - \bar{\psi}_n(\boldsymbol{\theta})\right|\right] \leq C w_n \sqrt{n}.$$

Furthermore, if $w_n\sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$, then $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\left|\mathbb{E}_{\mathbf{x} \sim \pi}\left[\psi(\mathbf{x}, \boldsymbol{\theta})\right] - \bar{\psi}_n(\boldsymbol{\theta})\right| \to 0$ almost surely as $t \to \infty$.

The following lemma establishes the fluctuation bound in Lemma 20 for univariate observables.

Lemma 42 *Fix compact subsets $\mathcal{X} \subseteq \mathbb{R}^q$, $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and a bounded Borel measurable function $\psi : \mathcal{X} \times \boldsymbol{\Theta} \to \mathbb{R}$. Let $(\mathbf{x}_n)_{n \geq 1}$ denote a sequence of points in $\mathcal{X}$ such that $\mathbf{x}_n = \varphi(X_n)$ for $n \geq 1$, where $(X_n)_{n \geq 1}$ is a Markov chain on a state space Ω and $\varphi : \Omega \to \mathcal{X}$ is a measurable function. Fix a sequence of weights $w_n \in (0, 1]$, $n \geq 1$ and define functions $\bar{\psi}(\cdot) := \mathbb{E}_{\mathbf{x} \sim \pi}\left[\psi(\mathbf{x}, \cdot)\right]$ and $\bar{\psi}_n : \boldsymbol{\Theta} \to \mathbb{R}$ recursively as $\bar{\psi}_0 \equiv \mathbf{0}$ and*

$$\bar{\psi}_n(\cdot) = (1 - w_n)\bar{\psi}_{n-1}(\cdot) + w_n\psi(\mathbf{x}_n, \cdot).$$

Assume the following:

(a1) *The Markov chain $(X_n)_{n \geq 1}$ mixes exponentially fast to its unique stationary distribution and the stochastic process $(\mathbf{x}_n)_{n \geq 1}$ on $\mathcal{X}$ has a unique stationary distribution π.*

(a2) *w_n is non-increasing in n and $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all sufficiently large $n \geq 1$.*

Then there exists a constant $C > 0$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\left\|\bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta})\right\|\right] \leq C w_n \sqrt{n}. \tag{C.5}$$

Furthermore, if $w_n\sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$, then $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\left\|\bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta})\right\| \to 0$ as $t \to \infty$ almost surely.

Proof We first argue for the special case when $w_n = 1/n$ for $n \geq 1$. In this case, $w_n^{-1} - w_{n-1}^{-1} \equiv 1$ so **(a2)** is satisfied, and also $w_k^n \equiv 1/n$ for $1 \leq k \leq n$ so $\bar{\psi}_n(\boldsymbol{\theta})$ becomes the sample mean

$$\bar{\psi}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{k=1}^{n} \psi(\mathbf{x}_k, \boldsymbol{\theta}).$$

If $\mathbf{x}_k$'s are i.i.d. from π, then (C.5) holds by a standard empirical process theory. More generally under the hypothesis **(a1)**, (C.5) also holds by using a uniform functional central limit theorem for Markov chains. This statement has been shown in (Lyu et al., 2020, Lem. 10).

Now we consider a more general weighting scheme under **(a2)**. Fix $t \in \mathbb{N}$. Define $\Psi_i(\boldsymbol{\theta}) = (t - i + 1)^{-1} \sum_{j=1}^{t} \psi(\mathbf{x}_j, \boldsymbol{\theta})$ for each $1 \leq i \leq t$. Denote $\bar{\psi}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x} \sim \pi}[\psi(\mathbf{x}, \boldsymbol{\theta})]$. By the previous case, there exists a constant $c_1 > 0$ such that

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\Psi_i(\boldsymbol{\theta}) - \bar{\psi}(\boldsymbol{\theta})|\right] \leq \frac{c_1}{\sqrt{n - i + 1}}$$

for all $1 \leq i \leq n$. A simple calculation shows the following identity

$$\bar{\psi}_n - \bar{\psi} = \sum_{i=1}^{n} (w_i^n - w_{i-1}^n)(n - i + 1)(\Psi_i - \bar{\psi}),$$

with the convention of $w_0^n = 0$. Also, suppose $T \geq 1$ is such that $w_k^{-1} - w_{k-1}^{-1} \leq 1$ for $k \geq T$. Note that for $i \geq 2$, $w_{i-1}^n \leq w_i^n$ if and only if $w_{i-1}(1 - w_i) \leq w_i$ if and only if $w_i^{-1} - w_{i-1}^{-1} \leq 1$. Hence for each $n > T$ and $k \geq T$, we have $w_k^n \leq w_{k+1}^n \leq \cdots \leq w_n^n = w_n$. Then observe that

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\bar{\psi}_n(\boldsymbol{\theta}) - \bar{\psi}(\boldsymbol{\theta})|\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} |w_i^n - w_{i-1}^n|(t - i + 1) \sup_{W \in \boldsymbol{\Theta}} |\Psi_i(\boldsymbol{\theta}) - \bar{\psi}(\boldsymbol{\theta})|\right]$$

$$= \sum_{i=1}^{n} |w_i^n - w_{i-1}^n|(n - i + 1) \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\Psi_i(\boldsymbol{\theta}) - \bar{\psi}(\boldsymbol{\theta})|\right]$$

$$\leq \sum_{i=1}^{n} |w_i^n - w_{i-1}^n| c_1 \sqrt{n - i + 1}$$

$$\leq c_1 \sqrt{n} \left(\sum_{i=1}^{T} |w_i^n - w_{i-1}^n| + \sum_{i=T}^{n} (\hat{w}_i^n - \hat{w}_{i-1}^n)\right)$$

$$\leq c_1 \sqrt{n} \left(w_n + \sum_{i=1}^{T} w_i^n\right).$$

By using Lemma 43, we have $\sum_{i=1}^{T} w_i^n = O(1/n)$. Hence the above shows (C.5). Lastly, note that **(a2)** implies $w_n \geq c_2/n$ for all $n \geq 1$ for some constant $c_2 > 0$ (see the proof of Lemma 43) for details. Therefore we can conclude (C.5).

To show the second part of the assertion, suppose $w_n \sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some fixed $\varepsilon > 0$. Denote $X_n := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta})|$, which is a nonnegative random variable. We wish to show $X_n \to 0$ almost surely as $n \to \infty$. Suppose for contradiction that there exists a diverging sequence $n(k)$ such that $X_{n(k)}$ does not converge to 0 with some positive probability. According to the first part, there exists a constant $c > 0$ such that $\mathbb{E}[X_n] \leq c(\log n)^{-1-\varepsilon}$ for all $n \geq 1$. By Markov inequality, we have

$$\mathbb{P}\left(|X_n| \geq \frac{1}{(\log n)^{\varepsilon/2}}\right) \leq \frac{c}{(\log n)^{1+(\varepsilon/2)}}.$$

We may choose a subsequence $m(r) = n(k(r))$ of $n(k)$ such that there exists a constant $c' > 0$ for which $m(r) \geq c' \exp(r)$ for all $r \geq 1$. Then it follows that

$$\mathbb{P}\left(|X_{m(r)}| \geq \frac{1}{(\log m(r))^{\varepsilon/2}}\right) \leq \frac{c}{r^{1+(\varepsilon/2)}}.$$

Since $m(r)$ diverges, by the Borel-Cantelli lemma, $X_{m(r)} \to 0$ almost surely as $r \to \infty$. This contradicts the assumption that $X_{n(r)}$ does not converge to zero almost surely. This completes the proof. ∎

The following lemma was used in the proof of Lemma 42.

Lemma 43 *Fix a sequence $(w_n)_{n \geq 1}$ of numbers in $(0, 1]$. Denote $w_k^n := w_k \prod_{i=k+1}^n (1 - w_i)$ for $1 \leq k \leq n$. Suppose $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all sufficiently large $n \geq 1$. Fix $T \geq 1$. Then for all $n \geq T$,*

$$\sum_{i=1}^T w_i^n = O(1/n).$$

Proof Suppose $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all $n \geq N$ for some $N \geq 1$. It follows that $w_n^{-1} - w_N^{-1} \leq n - N$, so $w_n \geq \frac{1}{n - N + w_N^{-1}}$. Hence for some constant $c > 0$, $w_n \geq \frac{1}{n+c}$ for all $n \geq N$. Denote $a \vee b = \max(a, b)$. Then note that

$$w_k^n = w_k \exp\left(\sum_{i=k+1}^n \log(1 - w_i)\right) \leq \exp\left(-\sum_{i=k+1}^n w_i\right) \leq \exp\left(-\int_{N \vee (k+1)}^n \frac{1}{x+c}\,dx\right) = \frac{[N \vee (k+1)] + c}{n+c},$$

where the second inequality uses $w_k \leq 1$ and the following inequality uses $\log(1 - a) \leq -a$ for $a < 1$. Hence for each fixed $1 \leq T \leq n$, we have

$$\sum_{k=1}^T w_k^n \leq T\left((N \vee (T+1)) + c\right) \frac{1}{n+c}.$$

This shows the assertion. ∎

Remark 44 In the original statement of (Mairal, 2013b, Lem B.7), the assumption that $w_{n+1}^{-1} - w_n^{-1} \leq 1$ for sufficiently large n was not used, and it seems that the argument in Mairal (2013b) needs this assumption. To give more detail, the argument begins with writing the empirical loss $f_n(\cdot) = \sum_{k=1}^n w_k^n \ell(\mathcal{X}_k, \cdot)$, where $w_k^n := w_k(1 - w_{k-1}) \cdots (1 - w_n)$, and proceeds with assuming the monotonicity $w_1^n \leq \cdots \leq w_n^n$, which is equivalent to $w_1 \geq 1 - w_1$ and $w_k \geq w_{k-1}(1 - w_k)$ for $2 \leq k \leq n$. In turn, this is equivalent to $w_k^{-1} - w_{k-1}^{-1} \leq 1$ for $2 \leq k \leq n$ and $w_1 \geq 1/2$. Note that this condition implies $w_k^{-1} \leq k + 1$, or $w_k \geq \frac{1}{k+1}$. This means that, asymptotically, w_k cannot decay faster than the balanced weight $1/k$. We have used this observation in the proof of Lemma 43.

Next, we will argue that (A4') implies (A4). This observation was made in (Lyu et al., 2022, Remark 20), but we repeat the argument here for completeness. It is clear that if the sequence $w_n \in (0, 1]$ satisfies (A4'), then $\sum_{n=1}^\infty w_n = \infty$ and $\sum_{n=1}^\infty w_n^2 \sqrt{n} < \infty$. So it remains to verify $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for sufficiently large t. Suppose $w_n = \Theta(n^{-\beta}(\log t)^{-\delta})$ for some $\beta \in [0, 1]$ and $\delta \geq 0$. Let $c_1, c_2 > 0$ be constants such that $w_n t^\beta (\log t)^\delta \in [c_1, c_2]$ for all $n \geq 1$. Then by the mean value

theorem,

$$
\begin{aligned}
w_{n+1}^{-1} - w_n^{-1} &\leq c_2 \left((n+1)^\beta (\log(n+1))^\delta - n^\beta (\log n)^\delta \right) \\
&\leq c_2 \sup_{n \leq s \leq n+1} \left(\beta s^{\beta-1} (\log s)^\delta + \delta s^{\beta-1} (\log s)^{\delta-1} \right) \\
&\leq c_2 \sup_{n \leq s \leq n+1} s^{\beta-1} (\log s)^{\delta-1} \left((\log s) + \delta \right).
\end{aligned}
$$

Since $n \geq 1$, the last expression is of $o(1)$ if $\beta < 1$. Otherwise, $w_n = n^{-1}$ for $t \geq 1$ by (A4'). Then $w_{n+1}^{-1} - w_n^{-1} \equiv 1$ for all $t \geq 1$.