Stochastic Optimization with Arbitrary Recurrent Data Sampling

William Powell [*1] **Hanbaek Lyu** [*1]

Abstract

For obtaining optimal first-order convergence guarantees for stochastic optimization, it is necessary to use a recurrent data sampling algorithm that samples every data point with sufficient frequency. Most commonly used data sampling algorithms (e.g., i.i.d., MCMC, random reshuffling) are indeed recurrent under mild assumptions. In this work, we show that for a particular class of stochastic optimization algorithms, we do not need any further property (e.g., independence, exponential mixing, and reshuffling) beyond recurrence in data sampling to guarantee optimal rate of first-order convergence. Namely, using regularized versions of Minimization by Incremental Surrogate Optimization (MISO), we show that for non-convex and possibly non-smooth objective functions with constraints, the expected optimality gap converges at an optimal rate $O(n^{-1/2})$ under general recurrent sampling schemes. Furthermore, the implied constant depends explicitly on the 'speed of recurrence', measured by the expected amount of time to visit a farthest data point, either averaged ('target time') or supremized ('hitting time') over the initial locations. We discuss applications of our general framework to decentralized optimization and distributed non-negative matrix factorization.

1. Introduction

In this paper we consider the minimization of a non-convex weighted finite-sum objective $f : \mathbb{R}^p \to \mathbb{R}$:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ f(\boldsymbol{\theta}) := \sum_{v\in\mathcal{V}} f^v(\boldsymbol{\theta})\pi(v) \right\} \tag{1}$$

where $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is a convex, but not necessarily compact feasible set and $\boldsymbol{\theta}$ represents the parameters of a model to be optimized. Here $\mathcal{V}$ is a finite index set where one can view each index $v \in \mathcal{V}$ representing (a batch of) data that can be accessed at once. Then $f^v(\boldsymbol{\theta})$ is the loss incurred using parameter $\boldsymbol{\theta}$ with respect to data at v, which is weighted by $\pi(v) \geq 0$ when forming the overall objective f in (1). Without loss of generality, we assume the $\pi(v)$s sum to one. When $\pi(v) \equiv \frac{1}{|\mathcal{V}|}$ the problem (1) becomes the classical finite sum problem in the optimization literature. Instances of non-uniform π arise when training a model with imbalanced data as has been studied in (Steininger et al., 2021; Wang et al., 2022b; Sow et al., 2024).

We aim to solve this problem by developing an algorithm which produces iterative parameter updates $\boldsymbol{\theta}_n$ given only access to an *arbitrary sequence of data samples* $(v_n)_{n\geq 1}$. In order to reach a first-order stationary point of (1) for general objectives, it is necessary to use a sampling algorithm that is *recurrent*, meaning that every data point is sampled infinitely often with 'sufficient frequency'. Note that recurrence is satisfied by many common sampling schemes such as i.i.d. (independently and identically distributed) sampling, (irreducible) Markov Chain Monte-Carlo (MCMC), cyclic sampling (Bertsekas, 2011), and random-reshuffling (Ying et al., 2017). The main question we ask in this work is the following:

- *Is there any class of stochastic optimization algorithms for which recurrent sampling is enough to obtain optimal first-order convergence guarantee for* (1)*?*

In this paper, we show that for a class of suitable extensions of stochastic optimization algorithms known as *Minimization by Incremental Surrogate Optimization* (MISO) (Mairal, 2015), no additional property of a data sampling algorithm (e.g., independence, exponential mixing, reshuffling) other than recurrence is needed in order to guarantee convergence to first-order stationary points. Furthermore, we show that the rate of convergence depends crucially on either the averaged or supremized return time to the farthest data point, corresponding to the notion of 'target time' and 'hitting time' in Markov chain theory, respectively.

With the original MISO algorithm in (Mairal, 2013), even under the general recurrent data sampling, we are able to

[*]Equal contribution [1]Department of Mathematics, University of Wisconsin-Madison, WI, USA. Correspondence to: William Powell <wgpowell@wisc.edu>, Hanbaek Lyu <hlyu@math.wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

obtain asymptotically optimal iteration complexity if we can use strongly convex surrogate functions. However, there is a significant technical bottleneck in showing asymptotic convergence to stationary points, which was classically established in (Mairal, 2015) in case of the i.i.d. data sampling. We find that using additional regularization helps with improving the convergence rate and allows us to prove asymptotic convergence to stationary points under arbitrary recurrent data sampling. For these reasons, we propose a slight extention of MISO that we call the *Regularized Minimization by Incremental Surrogate Optimization* (RMISO), which takes the following form:

Step 1. Sample v_n according to a recurrent sampling algorithm

Step 2. $g_n^{v_n} \leftarrow$ Convex majorizing surrogate of f^{v_n} at $\boldsymbol{\theta}_{n-1}$; $g_n^v = g_{n-1}^v$ for $v \neq v_n$

Step 3. $\bar{g}_n \leftarrow \sum_{v\in\mathcal{V}} g_n^v \pi(v)$; Compute

$$\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big[\bar{g}_n(\boldsymbol{\theta}) + \Psi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)\Big].$$

The algorithm maintains a list of majorizing surrogate functions for each data point v. At each step, a new data sample v_n is drawn according to a recurrent data sampling algorithm. We then find a new majorizing convex surrogate $g_n^{v_n}$ that is tight at the current parameter $\boldsymbol{\theta}_{n-1}$. All other surrogates are unchanged. Then the new parameter $\boldsymbol{\theta}_n$ is found by minimizing the empirical mean of the current surrogates plus a regularization term $\Psi(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|)$ that penalizes large values of $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|$. To handle dependent data, many algorithms use some form of projection or regularization to achieve this property (Lyu, 2023; Bhandari et al., 2018; Roy & Balasubramanian, 2023). This allows one to control the bias introduced by dependent sampling schemes as well as use the broader class of convex surrogates instead of requiring them to be strongly convex. The original MISO (Mairal, 2015) is recovered by omitting this regularization term. The particular choice of this term is crucial for the success of the analysis under the general recurrent data sampling setting.

Applications of our work include distributed optimization over networks where $\mathcal{V}$ forms the vertex set of a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and each vertex v stores some data. Prior work (Johansson et al., 2010; 2007; Ram et al., 2009; Lopes & Sayed, 2007; Mao et al., 2020; Even, 2023; Sun et al., 2022) studies the performance of various optimization algorithms in this setting assuming the sequence $(v_n)_{n\geq 1}$ is a Markov chain on the graph $\mathcal{G}$. Here π is typically taken to be uniform and it is frequently assumed that the Markov chain is an MCMC sampling converging to π, see (Sun et al., 2022; Johansson et al., 2010; Wang et al., 2022a).

In this setting we find both theoretically and empirically that convergence of our algorithm can be accelerated by choosing sampling schemes that guarantee a higher frequency of visits to each $v \in \mathcal{V}$. Such schemes may be non-Markovian or not aperiodic and so not guaranteed to converge to a stationary distribution. Moreover, our analysis does not require π to agree with the stationary measure of the sampling process if it exists. As remarked in (Even, 2023), this additional flexibility may be advantageous as it allows one to opt for more efficient sampling schemes who's stationary measure may not agree with the data-weighting-distribution π. In our context, these are the schemes which minimize the measures of recurrence we define in the sequel.

1.1. Contribution

Our algorithms and analysis consider three cases which we briefly summarize in the following bullet points.

- We show convergence rates of $O(n^{-1/2})$ for MISO with strongly convex surrogates or constant quadratic proximal regularization, matching the rate shown for SAG in (Even, 2023). The implied constant depends on the potentially much smaller 'target time' rather than the hitting time.
- The same convergence rates hold for MISO with dynamic quadratic proximal regularization where, inspired by the dynamic step size used for SAG in (Even, 2023), the regularization parameter is adaptive to the state of the sampling process. Asymptotic convergence of stationarity measures in expectation is also proved.
- Convergence rates of $O(n^{-1/2}\log n)$ are shown for MISO with diminishing search radius restriction, where averaged surrogates are minimized within a diminishing radius. We show almost sure convergence to stationarity for this method.
- We experimentally validate our results for the tasks of non-negative matrix factorization and logistic regression. We find that our method is robust to data heterogeneity as it produces stable iterate trajectories while still maintaining fast convergence (see Sec. 4.2).

1.2. Related Work

MISO (Mairal, 2015) was originally developed to solve finite sum problems under i.i.d sampling and proceeds by repeatedly minimizing a surrogate of the empirical loss function. In (Mairal, 2015) it is shown that for MISO the expected objective optimality gap $\mathbb{E}[f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}^*)]$ decays at rate $O(1/n)$ when the objective function is convex and exponentially fast when it is strongly convex, just as batch gradient descent does (Bottou et al., 2018). For non-convex f it is shown that the iterates produced by MISO converge to

the set of stationary points of f over a convex constraint set, but no convergence rate analysis is given. Convergence rates for non-convex objectives were later provided for unconstrained problems in (Qian et al., 2019) where it was shown that the expected gradient norm $\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|]$ decays at rate $O(n^{-1/2})$. This rate was matched for the constrained setting in (Karimi et al., 2022). However, both papers only consider i.i.d sampling.

(R)MISO may be compared with *Stochastic Averaged Gradient* (SAG) (Schmidt et al., 2017) as both store the most recent information computed using the data v and output new parameter updates $\boldsymbol{\theta}_n$ depending on an average of this information over $\mathcal{V}$. Recently in (Even, 2023), it was shown that for non-convex objectives SAG produces iterates such that the expected gradient norm decays at rate $O(n^{-1/2})$ under Markovian sampling. In comparison, the expected gradient norm converges at rate $O(n^{-1/4})$ for other stochastic first-order methods such as Stochastic Gradient Descent (SGD) (Sun et al., 2018; Alacaoglu & Lyu, 2023; Even, 2023; Karimi et al., 2019). Other works devoted to the study of first order optimization methods under Markovian sampling include (Beznosikov et al., 2023; Bhandari et al., 2018; Wang et al., 2022a; Huo et al., 2023; Lyu, 2023).

There has also been a recent focus on proving faster convergence for SGD using without-replacement sampling methods such as random-reshuffling (Gürbüzbalaban et al., 2021; Ying et al., 2017). This has been further extended to variance reduced algorithms (Huang et al., 2021; Malinovsky et al., 2023; Beznosikov & Takáč, 2023) and distributed optimization (Mishchenko et al., 2022; Horváth et al., 2022). New sampling algorithms that aim to improve over random reshuffling have been suggested in (Rajput et al., 2022; Lu et al., 2022a; Mohtashami et al., 2022). In particular, in (Lu et al., 2022b) the authors show that the convergence of SGD can be accelerated provided a certain concentration inequality holds and propose leveraging this using a greedy sample selection strategy.

To obtain our results, we adopt a new analytical approach which is inspired in part by the analysis of SAG in (Even, 2023). This strategy differs significantly from mixing rate arguments used in the analysis of stochastic optimization methods with Markovian data (e.g (Sun et al., 2018; Bhandari et al., 2018; Nagaraj et al., 2020; Lyu et al., 2020; 2022; Lyu, 2023; Alacaoglu & Lyu, 2023)). We give a short sketch of our proofs in Section 3.4 and a brief overview of mixing rate techniques and the challenges of adapting them to the analysis of MISO in Appendix B. We believe that these techniques may be of interest in their own right and may further contribute to analyzing other stochastic optimization methods with recurrent data sampling.

1.3. Notation

In this paper, we let $\mathbb{R}^p$ denote the ambient space for the parameter space $\boldsymbol{\Theta}$ equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the induced Euclidean norm $\|\cdot\|$. For $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\varepsilon > 0$, we let $B_\varepsilon(\boldsymbol{\theta})$ represent the closed Euclidean ball of radius ε centered at $\boldsymbol{\theta}$. We let $\mathbb{1}(A)$ be the indicator function of an event A which takes value 1 on A and 0 on A^c. We denote $\pi_{\min} = \min_{v \in \mathcal{V}} \pi(v)$. We let $a \wedge b = \min\{a, b\}$ for real numbers a and b. For a set $\mathcal{X}$ we let $|\mathcal{X}|$ denote its cardinality.

2. Preliminary Definitions and Algorithm Statement

In this section we state the two main algorithms used to solve (1). To do this we start by defining first-order surrogate functions and then define a few random variables that will be important in both implementation and analysis. First-order surrogates are defined by

Definition 2.1 (First-order surrogates)**.** A convex function $g : \mathbb{R}^p \to \mathbb{R}$ is a first-order surrogate function of f at $\boldsymbol{\theta}$ if

(i) $g(\boldsymbol{\theta}') \geq f(\boldsymbol{\theta}')$ holds for all $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$

(ii) the approximation error $h := g - f$ is differentiable and ∇h is L-Lipschitz continuous for some $L > 0$; moreover $h(\boldsymbol{\theta}) = 0$ and $\nabla h(\boldsymbol{\theta}) = 0$.

We denote by $\mathcal{S}_L(f, \boldsymbol{\theta})$ the set of all first order surrogates of f at $\boldsymbol{\theta}$ such that ∇h is L-Lipschitz. We further define $\mathcal{S}_{L,\mu}(f, \boldsymbol{\theta})$ to be the set of all surrogates $g \in \mathcal{S}_L(f, \boldsymbol{\theta})$ such that g is μ-strongly convex.

Certain properties of the data sampling process are crucial in our analysis, especially in proving Lemmas D.1, E.2, and E.4. Below we define the *return time* and *last passage time*.

Definition 2.2 (Return time)**.** For $n \geq 0$ and $v \in \mathcal{V}$, the time to return to data v starting from time n is defined as

$$\tau_{n,v} = \inf\{j \geq 1 : v_{n+j} = v\}. \tag{2}$$

That is, $\tau_{n,v}$ is the amount of time which one has to wait after time n for the process to return to v. The return time may be viewed as a generalization of the return times of a Markov chain. Indeed, $\tau_{0,v} = \inf\{n \geq 1 : v_n = v\}$ agrees with the classical notion of return time from the Markov chain literature. This is closely related to the last passage time defined below.

Definition 2.3 (Last passage time)**.** For $n \geq 1$ and $v \in \mathcal{V}$ we define the *last passage time* of v before time n as

$$k^v(n) = \sup\{j \leq n : v_j = v\}. \tag{3}$$

If the process has not yet visited v, i.e. $\{j \leq n : v_j = v\} = \emptyset$, then we set $k^v(n) = 1$.

Algorithm 1 Incremental Majorization Minimization with Dynamic Proximal Regularization

Input: Initialize $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$; $N > 0$; $\rho \geq 0$
Option: $Regularization \in \{Dynamic, Constant\}$
Initialize surrogates $g_0^v \in \mathcal{S}_{L,\mu}(f^v, \boldsymbol{\theta}_0)$.
for $n = 1$ **to** N **do**
  sample a data point v_n
  choose $g_n^{v_n} \in \mathcal{S}_{L,\mu}(f^{v_n}, \boldsymbol{\theta}_{n-1})$; $g_n^v = g_{n-1}^v \ \forall v \neq v_n$
  $\bar{g}_n \leftarrow \sum_{v \in \mathcal{V}} g_n^v \pi(v)$
  if $Regularization = Dynamic$ **then**
    $\rho_n \leftarrow \rho + \max_{v \in \mathcal{V}}(n - k^v(n))$
  else if $Regularization = Constant$ **then**
    $\rho_n \leftarrow \rho$
  end if
  $\boldsymbol{\theta}_n \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\bar{g}_n(\boldsymbol{\theta}) + \frac{\rho_n}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \right]$
end for
output: $\boldsymbol{\theta}_N$

Algorithm 2 Incremental Majorization Minimization with Diminishing Radius

Input: Initialize $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$; $N > 0$; $(r_n)_{n \geq 1}$
Initialize surrogates $g_0^v \in \mathcal{S}_L(f^v, \boldsymbol{\theta}_0)$.
for $n = 1$ **to** N **do**
  sample a data point v_n
  choose $g_n^{v_n} \in \mathcal{S}_L(f^{v_n}, \boldsymbol{\theta}_{n-1})$; $g_n^v = g_{n-1}^v$ for all $v \neq v_n$
  $\bar{g}_n \leftarrow \sum_{v \in \mathcal{V}} g_n^v \pi(v)$
  $\boldsymbol{\theta}_n \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \cap B_{r_n}(\boldsymbol{\theta}_{n-1})} \bar{g}_n(\boldsymbol{\theta})$
end for
output: $\boldsymbol{\theta}_N$

The last passage time $k^v(n)$ appears naturally as it is the last time the surrogate for data point v has been updated during the execution of either Algorithm 1 or 2. Thus, g_n^v is a surrogate of f^v at $\boldsymbol{\theta}_{k^v(n)-1}$ and the corresponding surrogate error at this point $h_n^v(\boldsymbol{\theta}_{k^v(n)-1})$ and its gradient are equal to zero. We will use this fact crucially in the proof of the key lemma, Lemma D.1.

Our algorithms are stated formally in Algorithms 1 and 2. In Algorithm 1, the regularization term uses *Proximal Regularization* (PR) while Algorithm 2 utilizes a *Diminishing Radius* (DR) restriction.

3. Main Results

3.1. Optimality Conditions

We now introduce the optimality conditions used in this paper and related quantities. Here we denote f to be a general objective function $f : \boldsymbol{\Theta} \to \mathbb{R}$, but elsewhere f will refer to the objective function in (1) unless otherwise stated.

For a given function f and $\boldsymbol{\theta}^*, \boldsymbol{\theta} \in \boldsymbol{\Theta}$, we define its *directional derivative* at $\boldsymbol{\theta}^*$ in the direction $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ as

$$\nabla f(\boldsymbol{\theta}^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*) := \lim_{\alpha \to 0^+} \frac{f(\boldsymbol{\theta}^* + \alpha(\boldsymbol{\theta} - \boldsymbol{\theta}^*)) - f(\boldsymbol{\theta}^*)}{\alpha} \quad (4)$$

A necessary first order condition for $\boldsymbol{\theta}^*$ to be a local minimum of f is to require $\nabla f(\boldsymbol{\theta}^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (see (Mairal, 2015)). Thus we define the optimality of f at $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ as

$$O_f(\boldsymbol{\theta}^*) := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} -\nabla f(\boldsymbol{\theta}^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*). \quad (5)$$

Note that $O_f(\boldsymbol{\theta}^*)$ is non-negative (since we may take $\boldsymbol{\theta} = \boldsymbol{\theta}^*$) and only positive if there exists some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\nabla f(\boldsymbol{\theta}^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*) < 0$. Thus we say that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is a *stationary point* of f over $\boldsymbol{\Theta}$ if $O_f(\boldsymbol{\theta}^*) = 0$. If f is differentiable and $\boldsymbol{\Theta}$ is convex, this is equivalent to $-\nabla f(\boldsymbol{\theta}^*)$ being in the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$. If $\boldsymbol{\theta}^*$ is in the interior of $\boldsymbol{\Theta}$ then it implies that $\|\nabla f(\boldsymbol{\theta}^*)\| = 0$.

For iterative algorithms, this stationary point condition may hardly be satisfied in a finite number of iterations. A practically important question is how the worst case number of iterations required to achieve an ε-approximate solution scales with the desired precision ε. We say that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approximate stationary point* of f over $\boldsymbol{\Theta}$ if $O_f(\boldsymbol{\theta}^*) \leq \varepsilon$. This notion of ε-approximate solution is consistent with the corresponding notion for unconstrained problems. In fact, if f is differentiable, and if $\boldsymbol{\theta}^*$ is distance at least one away from the boundary $\partial\boldsymbol{\Theta}$, then it reduces to $\|\nabla f(\boldsymbol{\theta}^*)\| \leq \varepsilon$. For each $\varepsilon > 0$, we then define the *worst-case iteration complexity* of an algorithm for solving (1) as

$$N_\varepsilon(\boldsymbol{\theta}_0) := \inf\{n \geq 1 : O_f(\boldsymbol{\theta}_n) \leq \varepsilon\}, \quad (6)$$

where $(\boldsymbol{\theta}_n)_{n \geq 0}$ is a sequence of iterates produced by the algorithm with initial estimate $\boldsymbol{\theta}_0$.

3.2. Assumptions

In this subsection, we state our assumptions for establishing the main results. Throughout this paper, we denote by $\mathcal{F}_n$ the σ-algebra generated by the samples $v_1, \ldots, v_n$ and the parameters $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_n$ produced by Algorithm 1 or 2. With this definition, $(\mathcal{F}_n)_{n \geq 1}$ defines a filtration.

In what follows we will also define some important quantities in terms of the measure theoretic definition of the L_∞ norm for random variables:

$$\|X\|_\infty = \inf\{t > 0 : \mathbb{P}(|X| > t) = 0\}. \quad (7)$$

This is due to the technical consideration that the conditional expectation $\mathbb{E}[\tau_{n,v}|\mathcal{F}_n]$ is random and hence so is $\sup_{n \geq 1} \mathbb{E}[\tau_{n,v}|\mathcal{F}_n]$. Our analysis requires this supremum to be bounded by a non-random constant, while in the fully general case $\sup_{n \geq 1} \mathbb{E}[\tau_{n,v}|\mathcal{F}_n]$ may be an unbounded.

We first state our main assumption on the sampling scheme.

Assumption 3.1 (Recurrent data sampling)**.** The sequence $(v_n)_{n \geq 1}$ of data samples defines a stochastic process

which satisfies the following property: for each $v \in \mathcal{V}$, $\sup_{n\geq 1} \|\mathbb{E}[\tau_{n,v}|\mathcal{F}_n]\|_\infty < \infty$, i.e., the expected return time conditioned on $\mathcal{F}_n$ is uniformly bounded.

Assumption 3.1 states that the data $(v_n)_{n\geq 1}$ are sampled in such a way that the expected time between visits to a particular data point is finite and uniformly bounded. Generalizing the notion of positive recurrence in Markov chain theory, we say a sampling algorithm is *recurrent* if Assumption 3.1 is satisfied. We emphasize that recurrence is the *only* requirement we make of the sampling process in order to prove the convergence rate guarantees in Theorem 3.8 and the asymptotic convergence in Theorem 3.9 (We do not assume independence or Markovian dependence, etc.). We include below a list of some commonly used recurrent sampling algorithms.

1. (*i.i.d. sampling*) Sampling data i.i.d from a fixed distribution is the most common assumption in the literature (Mairal, 2013; 2015; Bottou et al., 2018; Schmidt et al., 2017; Johnson & Zhang, 2013). Suppose we sample v_n i.i.d from some distribution γ on $\mathcal{V}$. Then the $\tau_{n,v}$ are independent geometric random variables taking values from $\{1, 2, \dots, \}$ with success probability $\gamma(v)$, so $\mathbb{E}[\tau_{n,v}|\mathcal{F}_n] = 1/\gamma(v)$. In particular, if γ is uniform $\mathbb{E}[\tau_{n,v}|\mathcal{F}_n] = |\mathcal{V}|$ for all n and v.

2. (*MCMC*) Markov chain Monte Carlo methods (see e.g. Ch.3 of (Levin & Peres, 2017)) produce a Markov chain (v_n) on $\mathcal{V}$. If this chain is irreducible then $\max_{v,w} \mathbb{E}_w[\tau_{0,v}]$ is finite (Levin & Peres, 2017). For any n, v, and initial distribution ν, the Markov property implies $\mathbb{E}_\nu[\tau_{n,v}|\mathcal{F}_n] = \mathbb{E}_{v_n}[\tau_{0,v}]$. So any irreducible Markov chain satisfies 3.1.

3. (*Cyclic sampling*) In cyclic sampling one samples data in order according to some enumeration until the dataset is exhausted. This process is then repeated until convergence. The authors of (Lu et al., 2022b) show that iteration complexity for SGD can be improved from $O(\varepsilon^{-4})$ to $O(\varepsilon^{-3})$ using such methods. To see that 3.1 holds in this setting, we simply notice that $\tau_{n,v} \leq |\mathcal{V}|$ for all n and v.

4. (*Reshuffling*) Reshuffling is similar to cycling sampling except that the dataset is randomly permuted at the beginning of each epoch (Lu et al., 2022b). It was observed empirically that random reshuffling performs better that i.i.d sampling in (Bottou, 2012) and further studied in (Gürbüzbalaban et al., 2021; Lu et al., 2022b). The authors of (Lu et al., 2022b) show the same improvement in iteration complexity for SGD as for cyclic sampling. In this case, 3.1 is satisfied since $\tau_{n,v} \leq 2|\mathcal{V}|$.

In Markov chain theory, the quantity $\max_{v,w} \mathbb{E}_w[\tau_{0,v}]$ is commonly denoted t_{hit}. Adapting this notion, we define

$$t_{\text{hit}} := \max_{v\in\mathcal{V}} \sup_{n\geq 1} \|\mathbb{E}[\tau_{n,v}|\mathcal{F}_n]\|_\infty \tag{8}$$

for each v when 3.1 holds, for general sampling schemes.

Continuing the connection with Markov chains, we also let

$$t_\odot := \sup_{n\geq 1} \left\| \sum_{v\in\mathcal{V}} \mathbb{E}[\tau_{n,v}|\mathcal{F}_n]\pi(v) \right\|_\infty \tag{9}$$

where π is as in (1). Note that if v_n is an irreducible Markov chain then the Markov property implies

$$t_\odot = \max_{w\in\mathcal{V}} \sum_{v\in\mathcal{V}} \mathbb{E}_w[\tau_{0,v}]\pi(v). \tag{10}$$

This is closely related to the *target time* define by

$$t_\odot^w = \sum_{v\in V} \mathbb{E}_w[\tau_v]\pi(v) \tag{11}$$

with the difference being that here $\tau_v = \inf\{n \geq 0 : v_n = v\}$ is the first hitting time of v rather than the first return time to v. The *random target lemma* (Lemma 10.1 in (Levin & Peres, 2017)) states that if π is the stationary distribution of the Markov chain, then the target time is independent of the starting state w. In this case, the quantities (10) and (11) only differ by one. This can be seen by first noting that $\mathbb{E}_w[\tau_v] = \mathbb{E}_w[\tau_{0,v}]$ if $v \neq w$. This leaves only the difference $\mathbb{E}_w[\tau_{0,w}]\pi(w) - \mathbb{E}_w[\tau_w]\pi(w)$. The second term is equal to zero and the first equals one since if π is the unique stationary distribution for the chain, $\pi(w) = \frac{1}{\mathbb{E}_w[\tau_{0,w}]}$ (Levin & Peres, 2017).

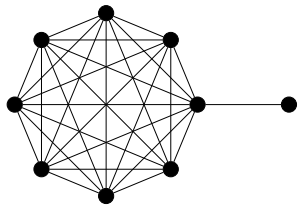


Figure 1. Lonely graph

For transitive irreducible Markov chains, t_{hit} and $t_\odot$ are comparable (Levin & Peres, 2017). However, in other situations $t_\odot$ may be much smaller that t_{hit}. For instance, consider the simple random walk on the graph in Figure 1, which we call the 'lonely' graph, and let π be its stationary distribution. In the lonely graph, $|\mathcal{V}| - 1$ vertices form a clique and the remaining vertex has degree one. A random walk on this graph has worst case hitting time $t_{\text{hit}} = O(|\mathcal{V}|^2)$: its value is tied to the lonely vertex with degree one which has low probability of being visited. On the other hand $t_\odot$ is only

$O(|\mathcal{V}|)$ since it only depends on an *average* hitting time instead of the worst case. See Example 10.4 in (Levin & Peres, 2017) for more details.

We again emphasize that in our general setting, we do not require (v_n) to be a Markov chain nor do we require π to be a stationary measure for the process. The above analysis serves to motivate the definition given in (9) and provide an example of where $t_\odot$ may be much smaller than t_{hit}.

We next list some assumptions of the functions f^v.

Assumption 3.2 (Lower-bounded objective and directional derivatives)**.** For all $v \in \mathcal{V}$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, the function f^v is bounded below, i.e. $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f^v(\boldsymbol{\theta}) > -\infty$. Moreover, the directional derivative $\nabla f^v(\boldsymbol{\theta}, \boldsymbol{\theta}' - \boldsymbol{\theta})$ exists.

Assumption 3.2 implies that the objective f is bounded below. For the remainder of this paper we will denote $\Delta_0 := \bar{g}_0(\boldsymbol{\theta}_0) - \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta})$. It is important to note that if the initial surrogates g_0^v are in $\mathcal{S}_L(f^v, \boldsymbol{\theta}_0)$ (as is the case in both Algorithms 1 and 2) then $g_0(\boldsymbol{\theta}_0) = f(\boldsymbol{\theta}_0)$ so $\Delta_0 = f(\boldsymbol{\theta}_0) - \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta})$. The regularity assumption in 3.2 was used in (Mairal, 2015) and is necessary in analyzing our algorithms using our definition of approximate stationarity. For Algorithm 2 we make the following stronger but common assumption which is crucial to our analysis:

Assumption 3.3. For each $v \in \mathcal{V}$, the function f^v is continuously differentiable and ∇f^v is L-Lipschitz continuous.

For simplicity, we assume that if Assumption 3.3 holds then the Lipschitz constant of f^v agrees with that of the corresponding approximation error h^v.

Finally, Assumption 3.4 states that the radii in Algorithm 2 decrease slowly, but not too slowly. This is analogous to square summability of step sizes in gradient descent.

Assumption 3.4 (Square-summable and non-summable radii)**.** The sequence $(r_n)_{n\geq 1}$ is non-increasing, $\sum_{n=1}^{\infty} r_n = \infty$, and $\sum_{n=1}^{\infty} r_n^2 < \infty$.

3.3. Statement of main results

In this section we state the two main results of this work. We consider the following three cases corresponding to the three variants of our main algorithm:

Case 3.5. Assumptions 3.1-3.2 hold. Use Algorithm 1 with $RegularizationSchedule = Constant$.

Case 3.6. Assumptions 3.1-3.2 hold. Use Algorithm 1 with $RegularizationSchedule = Dynamic$.

Case 3.7. Assumptions 3.1-3.4 hold. Use Algorithm 2.

Notice that in Case 3.5, if one chooses $\rho = 0$ and the surrogates are g_n^v are in $\mathcal{S}_{L,\mu}(f^v, \boldsymbol{\theta}_{n-1})$ for some $\mu > 0$, then Algorithm 1 reduces to the classical MISO algorithm in (Mairal, 2015).

Our first main result, Theorem 3.8, gives worst case upper-bounds on the expected rate of convergence to optimality. For each of the cases 3.5-3.7 we give rates of convergence for the objective function f.

Theorem 3.8 (Rate of Convergence to Stationarity)**.** *Algorithms 1 and 2 satisfy the following for any $N \geq 1$:*

(i) Assume Case 3.5. Further assume $\rho = 0$ and Algorithm 1 is run with μ-strongly convex surrogates. Then

$$\min_{1\leq n\leq N} \mathbb{E}[O_f(\boldsymbol{\theta}_n)] \leq \frac{Lt_\odot\sqrt{\frac{2\Delta_0}{\mu}}}{\sqrt{N}} \tag{12}$$

(ii) Assume Case 3.5. If $\rho \leq Lt_\odot \leq \rho + \mu$ then

$$\min_{1\leq n\leq N} \mathbb{E}\left[O_f(\boldsymbol{\theta}_n)\right] \leq 2\sqrt{\frac{2\Delta_0 Lt_\odot}{N}}. \tag{13}$$

(iii) Assume Case 3.6. If $\rho \leq Lt_\odot \leq \rho + \mu$ then

$$\min_{1\leq n\leq N} \mathbb{E}\left[O_f(\boldsymbol{\theta}_n)\right] \leq 2\sqrt{\frac{2\Delta_0(Lt_\odot + (2t_{hit}+1)\log_2(4|\mathcal{V}|))}{N}}. \tag{14}$$

(iv) Assume Case 3.7. Let $C_N = \sum_{n=1}^N r_n^2$. Then

$$\min_{1\leq n\leq N} \mathbb{E}\left[O_f(\boldsymbol{\theta}_n)\right] \leq \frac{\Delta_0 + \sqrt{\frac{2L}{\pi_{min}}C_N\Delta_0} + \left(3 + t_\odot\right)C_N L}{\sum_{n=1}^N (1 \wedge r_{n+1})}. \tag{15}$$

To our best knowledge, the rates of convergence given in Theorem 3.8 are entirely new for first-order algorithms with general recurrent data sampling. In contrast to the convergence result for SAG in (Even, 2023) that depends on the hitting time t_{hit}, Algorithm 1 with constant proximal regularization (case 3.5) depends on the possibly much smaller target time $t_\odot$. See Table 1 for a comparison of our results with other works concerning non-convex optimization with non-i.i.d data.

We remark that items (i) and (ii) show the potential benefit of using proximal regularization even if the surrogates are already strongly convex. For non-convex f, the strong convexity parameter μ of any surrogates cannot be larger than the Lipschitz constant L. However, we are free to choose ρ. So an optimal choice of ρ results in dependence on $\sqrt{Lt_\odot}$ in (ii) instead of the linear dependence in (i). However, it is not necessary to chose ρ in the range given in items (ii) and (iii). We include a more general version of Theorem 3.8, Theorem A.1, in Appendix A which shows that these convergence rates hold for arbitrary ρ and μ so long as $\rho + \mu > 0$. Overall, our theory suggests that one can

improve convergence by using sampling schemes that cover the dataset most efficiently, i.e. those that minimize $t_\odot$ and t_{hit}. See the remarks in Appendix A for more on this topic.

We also remark that Theorem 1 in (Even, 2023) gives a lower bound in terms of t_{hit}. While our results depend $t_\odot$, they do not contradict this lower bound. See Appendix A for more discussion.

Our second result, Theorem 3.9, concerns the asymptotic behavior of Algorithms 1 and 2.

Theorem 3.9 (Global Convergence)**.** *Algorithms 1 and 2 have the following asymptotic convergence properties:*

(i) For Case 3.6, we have $\lim_{n\to\infty}\mathbb{E}[O_f(\boldsymbol{\theta}_n)]=0$ *and* $\lim_{n\to\infty}\mathbb{E}[O_f(\boldsymbol{\theta}_n)^2]=0$.

(ii) For Case 3.7 almost surely every limit point of $(\boldsymbol{\theta}_n)_{n\geq 1}$ *is stationary for* f *over* $\boldsymbol{\Theta}$.

Theorem 3.9 shows that although RMISO with diminishing radius requires computing a projection at each step and has higher order dependence on $t_\odot$, it enjoys the strongest asymptotic guarantees. RMISO with dynamic proximal regularization is somewhere in the middle. It has lower order dependence on $t_\odot$ than the diminishing radius version but also depends on t_{hit}. However, we are able to show that both the first and second moments of the optimality gap converge to zero. In particular, notice that in the familiar case that $\boldsymbol{\Theta}=\mathbb{R}^p$ and f is differentiable (i) implies that $\lim_{n\to\infty}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|]=0$.

Though Algorithm 1 with constant proximal regularization is the simplest of our proposed methods and has the best dependence on the constants L and $t_\odot$, it appears that stronger regularization schemes as in cases 3.6 and 3.7 are needed for obtaining asymptotic convergence guarantees. We refer the reader to Appendix A as well as remark E.3 in Appendix E for a more detailed discussion on the technical difficulties of proving asymptotic convergence for Case 3.5 as well as proving it in the almost sure sense for Case 3.6.

3.4. Sketch of proofs

In this section we provide a short sketch of our analysis in order to convey the main ideas. Let $\bar{h}_n=\sum_{v\in V}h_n^v\pi(v)$ be the average surrogate approximation error at step n. The key step in our analysis (Lemma D.1) is to prove

$$\sum_{n=1}^N c_n\mathbb{E}[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|]=O\left(\left(\sum_{n=1}^N c_n^2\right)^{1/2}\right), \tag{16}$$

where c_n is any non-increasing sequence. For simplicity, assume that the surrogate functions g_n^v and the objective functions f^v are differentiable and we are in the unconstrained setting, i.e. $\boldsymbol{\Theta}=\mathbb{R}^p$. If $\boldsymbol{\theta}_n$ is a minimizer of $\bar{g}_n$ then $\nabla\bar{g}_n(\boldsymbol{\theta}_n)=0$. Therefore $\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|=\|\nabla f(\boldsymbol{\theta}_n)\|$ and (16) implies

$$\sum_{n=1}^N c_n\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|]=O\left(\left(\sum_{n=1}^N c_n^2\right)^{1/2}\right). \tag{17}$$

If we take $c_n=1$, we can then conclude that $\min_{1\leq n\leq N}\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|]=O(N^{-1/2})$. The addition of regularization introduces an added complication because we are no longer directly minimizing $\bar{g}_n$ on the entire feasible set and so do not have $\nabla\bar{g}_n(\boldsymbol{\theta}_n)=0$. However, as we argue in Section D, the added regularization is not too strong asymptotically.

The main idea is to focus in the individual error gradients because each has the property $\nabla h_n^v(\boldsymbol{\theta}_{k^v(n)-1})=0$. By Definition 2.1, we then have $\|\nabla h_n^v(\boldsymbol{\theta}_n)\|\leq L\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{k^v(n)-1}\|$, so to show (16) we only need to prove

$$\sum_{n=1}^N c_n\mathbb{E}[\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{k^v(n)-1}\|]=O\left(\left(\sum_{n=1}^N c_n^2\right)^{1/2}\right) \tag{18}$$

The triangle inequality and monotonicity of (c_n) imply

$$c_n\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{k^v(n)-1}\|\leq\sum_{i=k^v(n)}^n c_i\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_{i-1}\|, \tag{19}$$

so we can relate the error $\|\nabla h_n^v(\boldsymbol{\theta}_n)\|$ to the sequence $(\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|^2)_{n\geq 1}$. The crucial role played by the regularization or strong convexity is that we can prove the following iterate stability: $\sum_{n=1}^\infty\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|^2<\infty$ a.s. This idea was also used in (Lyu & Li, 2023; Lyu et al., 2022; Lyu, 2023).

Under Assumption 3.1 one can expect $\mathbb{E}[n-k^v(n)]\leq M$ for some M. One can then intuitively view the expectation of the right hand side of (19) similarly to $\sum_{i=n-M}^n c_i\mathbb{E}[\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_{i-1}\|]$. Summing this from $n=1$ to N we conclude that for a positive constant C,

$$\sum_{n=1}^N\mathbb{E}[\|\nabla h_n^v(\boldsymbol{\theta}_n)\|]\approx C\sum_{n=1}^N c_n\mathbb{E}[\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|]. \tag{20}$$

By Cauchy-Schwartz and the iterate stability obtained through regularization, the right hand side is $O\left(\left(\sum_{n=1}^N c_n^2\right)^{1/2}\right)$. Full details of our analysis are given in Appendices C, D, and E.

4. Applications and Experiments

4.1. Applications

In this section we give some applications of our general framework. These include applications to matrix factorization as well as a double averaging version of RMISO derived by using prox-linear surrogates.

Table 1. Comparison of iteration complexity for non-convex optimization with non-i.i.d data. The notation $\tilde{O}(\cdot)$ omits logarithmic factors. Here t_{mix} represents dependence on the mixing time of the Markov chain.

	Iteration complexity	Memory	Sampling	Sampling dependence
AdaGrad (Alacaoglu & Lyu, 2023)	$\tilde{O}(\varepsilon^{-4})$	$O(1)$	Markovian	$O(\sqrt{t_{\text{mix}}})$
SGD (Sun et al., 2018; Even, 2023)	$\tilde{O}(\varepsilon^{-4})$	$O(1)$	Markovian	$O(\sqrt{t_{\text{mix}}})$
SGD (Mishchenko et al., 2020; Lu et al., 2022b)	$O(\varepsilon^{-3})$	$O(1)$	Reshuffling	$O(\sqrt{\lvert\mathcal{V}\rvert})$
SAG (Even, 2023)	$O(\varepsilon^{-2})$	$O(\lvert\mathcal{V}\rvert)$	Markovian	$O(\sqrt{t_{\text{hit}}})$
RMISO Case 3.5	$O(\varepsilon^{-2})$	$O(\lvert\mathcal{V}\rvert)$	Recurrent	$O(\sqrt{t_{\odot}})$
RMISO Case 3.6	$O(\varepsilon^{-2})$	$O(\lvert\mathcal{V}\rvert)$	Recurrent	$O(\sqrt{t_{\odot}+t_{\text{hit}}})$
RMISO Case 3.7	$\tilde{O}(\varepsilon^{-2})$	$O(\lvert\mathcal{V}\rvert)$	Recurrent	$O(t_{\odot})$

4.1.1. Distributed Matrix Factorization

Before beginning this section we define some additional notation. For a collection of matrices $\{A_v\}_{v\in\mathcal{V}} \subset \mathbb{R}^{n\times m}$ we let $[A_v; v\in\mathcal{V}]$ be their concatenation along the horizontal axis. For a set $\boldsymbol{\Theta} \subset \mathbb{R}^{n\times m}$ we let $\boldsymbol{\Theta}^{\mathcal{V}} = \{[A_v; v\in\mathcal{V}] : A_v \in \boldsymbol{\Theta} \text{ for all } v\}$.

We consider the matrix factorization loss $f(W,H) = \frac{1}{2}\|X - WH\|_F^2 + \alpha\|H\|_1$ where $X \in \mathbb{R}^{p\times d}$ is a given data matrix to be factored into the product of dictionary $W \in \boldsymbol{\Theta}_W \subseteq \mathbb{R}^{p\times r}$ and code $H \in \boldsymbol{\Theta}_H \subseteq \mathbb{R}^{r\times d}$ with $\alpha \geq 0$ being the L_1-regularization parameter for H. Here $\boldsymbol{\Theta}_W$ and $\boldsymbol{\Theta}_H$ are convex constraint sets.

Suppose we have a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each vertex stores a matrix $X_v \in \mathbb{R}^{p\times d}$. For each $v \in \mathcal{V}$ define the loss function

$$f^v(W) = \inf_{H\in\boldsymbol{\Theta}_H} \frac{1}{2}\|X^v - WH\|_F^2 + \alpha\|H\|_1, \qquad (21)$$

which is the minimum reconstruction error for factorizing X_v using the dictionary W. In this context, the empirical loss to be minimized is $\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} f^v(W)$. Note that this problem is not convex. Indeed, letting $X = [X_v; v \in V]$ it is equivalent to finding $(W^*, H^*) \in \boldsymbol{\Theta}_W \times \boldsymbol{\Theta}_H^{\mathcal{V}}$ minimizing $\frac{1}{2}\|X - WH\|_F^2 + \alpha\|H\|_1$, which is a constrained non-convex optimization problem with a bi-convex loss function.

In order to apply RMISO let $W_{n-1} \in \boldsymbol{\Theta}_W$ be the previous dictionary and denote

$$H_n^v \in \operatorname*{arg\,min}_{H\in\boldsymbol{\Theta}_H^{\mathcal{V}}} \frac{1}{2}\|X_v - W_{n-1}H\|_F^2 + \alpha\|H\|_1 \qquad (22)$$

if $v_n = v$ and otherwise $H_n^v = H_{n-1}^v$. Then the function $g_n^v(W) := \frac{1}{2}\|X_v - W_{n-1}H_n^v\|_F^2 + \alpha\|H_n^v\|_1$ is a majorizing surrogate of f^v at W_{n-1} and belongs to $\mathcal{S}_{L'}(f^v, W_{n-1})$ for some $L' > 0$ (see Ex. G.5). Then Algorithms 1 and 2 can be used with these surrogates.

4.1.2. Prox-Linear Surrogates

Suppose each f^v is differentiable and has L-Lipschitz continuous gradients. Then the functions

$$g^v(\boldsymbol{\theta}) = f^v(\boldsymbol{\theta}') + \langle \nabla f^v(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \qquad (23)$$

are in $\mathcal{S}_{2L,L}(f^v, \boldsymbol{\theta}')$ (see Example G.2) . Further suppose that π is the uniform distribution. Using these surrogates, the update according to Algorithm 1 is

$$\begin{cases} \bar{\boldsymbol{\theta}}_{n-1} & \leftarrow \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \boldsymbol{\theta}_{k^v(n)-1} \\ \bar{\nabla}_{n-1} & \leftarrow \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \nabla f^v(\boldsymbol{\theta}_{k^v(n)-1}) \\ \tilde{\boldsymbol{\theta}}_{n-1} & \leftarrow \frac{\rho_n}{L+\rho_n}\boldsymbol{\theta}_{n-1} + \frac{L}{L+\rho_n}\bar{\boldsymbol{\theta}}_{n-1} \\ \boldsymbol{\theta}_n & \leftarrow \text{Proj}_{\boldsymbol{\Theta}}\left(\tilde{\boldsymbol{\theta}}_{n-1} - \frac{1}{L+\rho_n}\bar{\nabla}_{n-1}\right). \end{cases} \qquad (24)$$

Compared with MISO (obtained by setting $\rho_n = 0$ in (24)) we see that the additional proximal regularization has the effect of further averaging the iterates, putting additional weight of $\frac{\rho_n}{L+\rho_n}$ on the most recent parameter $\boldsymbol{\theta}_{n-1}$.

4.2. Experiments

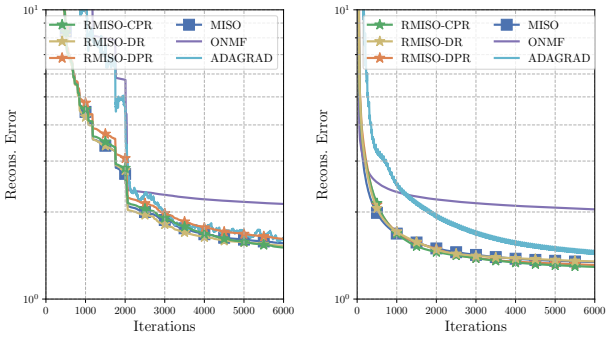
4.2.1. Distributed Nonnegative Matrix Factorization

In this section we compare the performance of the distributed matrix factorization version of RMISO from Sec. 4.1.1 against other well known optimization algorithms. We consider a randomly drawn collection of 5000 images from the MNIST (Deng, 2012) dataset where each sample X_v represents a subset of images. In all experiments, we set $\alpha = \frac{1}{28}$ and $r = 15$. The dictionary W is constrained to be non-negative and rows with euclidean norm at most one.

The set of vertices $\mathcal{V}$ is arranged in a cycle graph with $|\mathcal{V}| = 55$ with each vertex restricted to only contain samples with the same label. We consider two different sampling algorithms: the standard random walk where $t_{\odot}$ and t_{hit} are both $O(|\mathcal{V}|^2)$, and cyclic where both are $O(|\mathcal{V}|)$. Our theory suggests we should expect better performance for

cyclic sampling versus the random walk. We compare all three versions of RMISO: constant proximal regularization (RMISO-CPR), dynamic proximal regularization (RMISO-DPR), and diminishing radius (RMISO-DR), with MISO (Mairal, 2015), the online nonnegative matrix factorization (ONMF) algorithm of (Mairal et al., 2010), and AdaGrad (Duchi et al., 2011).

It is not guaranteed that the surrogates $g_n^v(W) := \frac{1}{2}\|X_v - W_{n-1}H_n^v\|_F^2 + \alpha\|H_n^v\|_1$ are strongly convex. However, while running the experiments, we find that the Hessian of the averaged surrogate is positive definite after only a few iterations and thus the results for MISO are also supported by Theorem 3.8 (ii). This phenomenon is also discussed in Assumption B of (Mairal et al., 2010).



(a) Random Walk (b) Cyclic

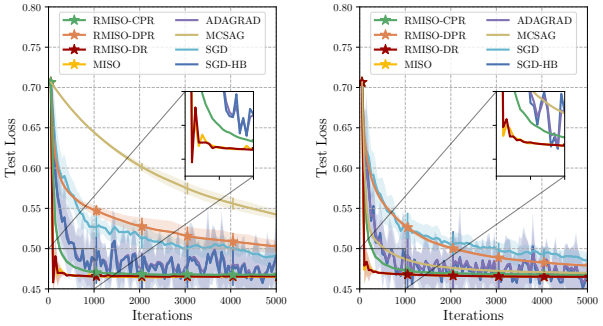
Figure 2. Plot of reconstruction error against interation number for NMF using two sampling algorithms. Results show the performance of algorithms RMISO, MISO (Algorithm 1 with $\rho_n = 0$), ONMF, and AdaGrad in factorizing a collection of MNIST (Deng, 2012) data matrices.

We ran the experiment ten times with ten different random seeds and plot the average reconstruction error versus iteration number in Figure 2. We see that RMISO outperforms ONMF and shows competitive performance against AdaGrad for both sampling schemes. As expected, there is a dramatic performance improvement under cyclic sampling versus the random walk.

4.2.2. Logistic regression with nonconvex regularization

We consider logistic regression with the non-convex regularization term $R(\theta) = 0.01 \cdot \sum_{i=1}^{p} \frac{\theta_i^2}{1+\theta_i^2}$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter to be optimized. We use the `a9a` dataset (Becker & Kohavi, 1996). Here we consider the random walk on two separate graph topologies: the complete graph and the 'lonely' graph as in Figure 1. Both graphs have $|\mathcal{V}| = 50$ and each vertex only stores data with the same label.

We compared eight different optimization algorithms: (1) the prox-linear version of Algorithm 1 (24) with non-zero proximal regularization (RMISO-CPR); (2) Algorithm 1 with dynamic proximal regularization; (3) Algorithm 2; (4) MISO (Algorithm 1 with $\rho_n = 0$); (5) AdaGrad (Duchi et al., 2011); (6) Markov Chain SAG (MCSAG) (Even, 2023); (7) SGD with decaying step size; (8) SGD-HB (SGD with momentum).



(a) Lonely graph (b) Complete graph

Figure 3. Plot of objective loss and standard deviation against the test dataset for `a9a` for two graph topologies and various optimization algorithms- RMISO, MISO (Algorithm 1 with $\rho_n = 0$), AdaGrad, MCSAG, SGD, Adam, and SGD-HB

We ran each experiment with ten different seeds. The results plotted in Figure 3 show the average loss against the test dataset for both graph topologies over these ten runs as well as a shaded region with boundaries given by the standard deviation. We see that RMISO-DPR and MCSAG display poorer performance on the lonely graph as t_{hit} increases from $O(|\mathcal{V}|)$ to $O(|\mathcal{V}|^2)$. The performance of RMISO-CPR, RMISO-DR, and MISO are unchanged since each only depends on $t_{\odot}$, with RMISO-DR and MISO performing the best and only narrowly outperforming RMISO-CPR.

Both SGD-HB and AdaGrad converge quickly in both settings but suffer from unstable trajectories compared to RMISO and MCSAG. A more stable algorithm may be advantageous in situations where the value of the objective function cannot easily be computed. See (Nesterov & Shikhman, 2015) for an example of such a situation.

5. Conclusion

In this paper we have established convergence and complexity results for our proposed extensions of MISO under the general assumption of recurrent data sampling. Our results show that convergence speed depends crucially on the average or supremized expected time to return to a given data point. In particular, the constant proximal regularization version of our algorithm depends only on the averaged target time, a potentially large improvement over the hitting time. Both our analysis and numerical experiments display the benefit of using possibly non-i.i.d or non-Markovian sampling schemes in order to accelerate convergence.

Acknowledgements

This work was supported in part by NSF Award DMS-2023239 and DMS-2206296. The authors thank Qiaomin Xie for helpful comments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alacaoglu, A. and Lyu, H. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 458–489. PMLR, 2023.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996.

Bertsekas, D. P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

Beznosikov, A. and Takáč, M. Random-reshuffled SARAH does not need full gradient computations. *Optimization Letters*, 2023.

Beznosikov, A., Samsonov, S., Sheshukova, M., Gasnikov, A., Naumov, A., and Moulines, E. First order methods with Markovian noise: from acceleration to variational inequalities. *arXiv preprint arXiv:2305:.15938*, 2023.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the 31st Conference On Learning Theory*, pp. 1691–1692. PMLR, 2018.

Bottou, L. *Stochastic Gradient Descent Tricks*, pp. 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

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

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

Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

Even, M. Stochastic gradient descent under Markovian sampling schemes. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 9412–9439. PMLR, 2023.

Gürbüzbalaban, M., Ozdaglar, A., and Parrilo, P. A. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

Horst, R. and Thoai, N. V. Dc programming: Overview. *Journal of Optimization Theory and Applications*, 103(1): 1–43, 1999.

Horváth, S., Sanjabi, M., Xiao, L., Richtárik, P., and Rabbat, M. Fedshuffle: Recipes for better use of local work in federated learning. *Transactions on Machine Learning Research*, 2022.

Huang, X., Yuan, K., Mao, X., and Yin, W. An improved analysis and rates for variance reduction under without-replacement sampling orders. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

Huo, D. L., Chen, Y., and Xie, Q. Bias and extrapolation in Markovian linear stochastic approximation with constant stepsizes. *ACM SIGMETRICS Performance Evaluation Review*, 51(1):81–82, 2023.

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

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

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. Non-asymptotic analysis of biased stochastic approximation scheme. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 1944–1974. PMLR, 2019.

Karimi, B., Wai, H.-T., Moulines, E., and Li, P. Minimization by incremental stochastic surrogate optimization for large scale nonconvex problems. In Dasgupta, S. and Haghtalab, N. (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167, pp. 606–637. PMLR, 2022.

Levin, D. and Peres, Y. *Markov Chains and Mixing Times*. MBK. American Mathematical Society, 2017.

Lopes, C. G. and Sayed, A. H. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4064–4077, 2007.

Lu, Y., Guo, W., and De Sa, C. M. Grab: Finding provably better data permutations than random reshuffling. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022a.

Lu, Y., Meng, S. Y., and Sa, C. D. A general analysis of example-selection for stochastic gradient descent. In *International Conference on Learning Representations*, 2022b.

Lyu, H. Stochastic regularized majorization-minimization with weakly convex and multi-convex surrogates. *arXiv preprint arXiv:2201.01652*, 2023.

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

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

Lyu, H., Strohmeier, C., and Needell, D. Online nonnegative cp-dictionary learning for Markovian data. *The Journal of Machine Learning Research*, 23(1):6630–6679, 2022.

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

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(2):19–60, 2010.

Malinovsky, G., Sailanbayev, A., and Richtárik, P. Random reshuffling with variance reduction: New analysis and better rates. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pp. 1347–1357. PMLR, 2023.

Mao, X., Yuan, K., Hu, Y., Gu, Y., Sayed, A. H., and Yin, W. Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528, 2020.

Mishchenko, K., Khaled, A., and Richtarik, P. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020.

Mishchenko, K., Khaled, A., and Richtarik, P. Proximal and federated random reshuffling. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 15718–15749. PMLR, 2022.

Mohtashami, A., Stich, S. U., and Jaggi, M. Characterizing & finding good data orderings for fast convergence of sequential gradient methods. *arXiv preprint arXiv:2202.01838*, 2022.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. Least squares regression with Markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676, 2020.

Nesterov, Y. *Introductory Lectures on Convex Optimization*. Applied Optimization. Springer, 2003.

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

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

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimimization*, 1(3):127–239, 2014.

Qian, X., Sailanbayev, A., Mishchenko, K., and Richtárik, P. MISO is making a comeback with better proofs and rates. *arXiv preprint arXiv:1906.01474*, 2019.

Rajput, S., Lee, K., and Papailiopoulos, D. Permutation-based SGD: Is random optimal? In *International Conference on Learning Representations*, 2022.

Ram, S. S., Nedić, A., and Veeravalli, V. V. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, 2009.

Roy, A. and Balasubramanian, K. Online covariance estimation for stochastic gradient descent under Markovian sampling. *arXiv preprint arXiv:2308.01481*, 2023.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

Sow, D., Lin, S., Wang, Z., and Liang, Y. Doubly robust instance-reweighted adversarial training. In *The Twelfth International Conference on Learning Representations*, 2024.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Sun, T., Sun, Y., and Yin, W. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sun, T., Li, D., and Wang, B. Adaptive random walk gradient descent for decentralized optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20790–20809. PMLR, 2022.

Wang, P., Lei, Y., Ying, Y., and Zhou, D.-X. Stability and generalization for Markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pp. 37735–37748, 2022a.

Wang, W., Xu, H., Liu, X., Li, Y., Thuraisingham, B., and Tang, J. Imbalanced adversarial training with reweighting. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1209–1214, Los Alamitos, CA, USA, dec 2022b. IEEE Computer Society.

Ying, B., Yuan, K., Vlaski, S., and Sayed, A. H. On the performance of random reshuffling in stochastic learning. In *2017 Information Theory and Applications Workshop*, pp. 1–5, 2017.

A. Further remarks on main results

We include in this section an extended version of Theorem 3.8 including convergence rates for arbitrary regularization parameters as well as two of its corollaries. The extended version of Theorem 3.8 is below.

Theorem A.1 (Extended Version of Theorem 3.8 in the main text)**.** *Algorithms 1 and 2 satisfy the following:*

(i) Assume Case 3.5. Then

$$\min_{1\le n\le N}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le 1}-\nabla f(\boldsymbol{\theta}_n,\boldsymbol{\theta}-\boldsymbol{\theta}_n)\right]\le\frac{\sqrt{2\Delta_0}\left(\frac{\rho}{\sqrt{\rho+\mu}}+\frac{Lt_\odot}{\sqrt{\rho+\mu}}\right)}{\sqrt{N}}. \tag{25}$$

In particular, if ρ is chosen so that $\rho\le Lt_\odot\le\rho+\mu$ then

$$\min_{1\le n\le N}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le 1}-\nabla f(\boldsymbol{\theta}_n,\boldsymbol{\theta}-\boldsymbol{\theta}_n)\right]\le 2\sqrt{\frac{2\Delta_0 Lt_\odot}{N}}. \tag{26}$$

(ii) Assume Case 3.6. Then

$$\min_{1\le n\le N}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le 1}-\nabla f(\boldsymbol{\theta}_n,\boldsymbol{\theta}-\boldsymbol{\theta}_n)\right]\le\frac{\sqrt{2\Delta_0}\left(\sqrt{\rho+(2t_{hit}+1)\log_2(4|\mathcal{V}|)}+\frac{Lt_\odot}{\sqrt{\rho+\mu}}\right)}{\sqrt{N}}. \tag{27}$$

If ρ satisfies the condition $\rho\le Lt_\odot\le\rho+\mu$ then

$$\min_{1\le n\le N}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\le 1\|}-\nabla f(\boldsymbol{\theta}_n,\boldsymbol{\theta}-\boldsymbol{\theta}_n)\right]\le\frac{\sqrt{2\Delta_0}\left(\sqrt{Lt_\odot+(2t_{hit}+1)\log_2(4|\mathcal{V}|)}+\sqrt{Lt_\odot}\right)}{\sqrt{N}}. \tag{28}$$

(iii) Assume Case 3.7. Then

$$\min_{1\le n\le N}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le 1}\langle-\nabla f(\boldsymbol{\theta}_n),\boldsymbol{\theta}-\boldsymbol{\theta}_n\rangle\right]\le\frac{\Delta_0+\sqrt{\frac{2L}{\pi_{min}}C_N\Delta_0}+\left(3+t_\odot\right)C_N L}{\sum_{n=1}^N 1\wedge r_{n+1}}, \tag{29}$$

where $C_N=\sum_{n=1}^N r_n^2$.

Next, Corollary A.2 specializes these results to the setting of unconstrained nonconvex optimization.

Corollary A.2. *Assume either $\boldsymbol{\Theta}=\mathbb{R}^p$ or that there exists $c\in(0,1]$ so that $dist(\boldsymbol{\theta}_n,\partial\boldsymbol{\Theta})\ge c$ for all $n\ge 1$.*

(i) Let $(\boldsymbol{\theta}_n)_{n\ge 0}$ be an output of Algorithm 1. Assume case 3.5 or 3.6. Then for any $N\ge 1$

$$\min_{1\le n\le N}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_n)\|\right]=O\left(N^{-1/2}\right). \tag{30}$$

(ii) Let $(\boldsymbol{\theta}_n)_{n\ge 0}$ be an output of Algorithm 2. Assume case 3.7. Then for any $N\ge 1$

$$\min_{1\le n\le N}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_n)\|\right]=O\left(\left(\sum_{n=1}^N 1\wedge r_n\right)^{-1}\right). \tag{31}$$

Notice that we may take $r_n=\frac{1}{\sqrt{n\log n}}$ in Algorithm 2. Then Corollary A.2 implies

$$\min_{1\le n\le N}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_n)\|\right]=O\left(\frac{\log N}{\sqrt{N}}\right) \tag{32}$$

holds for case 3.7.

Finally, Corollary A.3 states the iteration complexity of Algorithms 1 and 2.

Corollary A.3 (Iteration Complexity)**.** *Algorithms 1 and 2 have the following worst case iteration complexity:*

(i) Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1. Assume case 3.5 or 3.6. Then $N_\varepsilon(\boldsymbol{\theta}_0) = O(\varepsilon^{-2})$.

(ii) Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 2. Assume case 3.7. Then $N_\varepsilon(\boldsymbol{\theta}_0) = O(\varepsilon^{-2}(\log \varepsilon^{-1})^2)$.

Remark A.4 (Comparison with the lower bound of (Even, 2023) Theorem 1). Using our notation, the lower bound given in Theorem 1 of (Even, 2023) is

$$\|\nabla f(\boldsymbol{\theta}_N)\|^2 = \Omega\left(L\Delta_0\left(\frac{t_{\text{hit}}}{N}\right)^2\right). \tag{33}$$

Our convergence rates are given in terms of $\|\nabla f(\theta_n)\|$ rather than $\|\nabla f(\theta_n)\|^2$, so in our setting this is

$$\|\nabla f(\boldsymbol{\theta}_N)\| = \Omega\left(\sqrt{L\Delta_0}\left(\frac{t_{\text{hit}}}{N}\right)\right). \tag{34}$$

Notice that the rate of convergence in the lower bound is $O(N^{-1})$ while our upper bound gives a rate of convergence of $O(N^{-1/2})$. Thus, despite the dependence in our results on $t_\odot$, they do not contradict this lower bound.

Remark A.5 (Optimal sampling and estimating $t_\odot$ and t_{hit}). The dependence of the convergence rates on $t_\odot$ or t_{hit} in Theorem A.1 suggests one can accelerate convergence by choosing a sampling algorithm with the smallest values of these constants appropriate for the context. In general, an optimal sampling scheme is problem dependent. The best one can hope for, in terms of dependence on $|\mathcal{V}|$, is that both constants are $O(|\mathcal{V}|)$ which is achieved by i.i.d sampling. However, this may not be feasible in settings like decentralized optimization where communication can only occur between neighboring vertices in a graph.

Depending on the graph topology, it is likely that for the standard random walk t_{hit} and $t_\odot$ are much larger than $|\mathcal{V}|$, especially for sparse graphs. If a cycle containing all nodes in the graph exists, our theory suggests using cyclic sampling by traversing such a spanning cycle. In this case, both t_{hit} and $t_\odot$ have optimal order $O(|\mathcal{V}|)$. If no such cycle exists, a good way to minimized t_{hit} is to find the shortest path in the graph which contains all vertices and then sample the vertices deterministically in order by walking over this path. This idea holds in a more general setting beyond optimization on graphs: a good way to minimize t_{hit} is to sample data as efficiently as possible by covering the dataset with the fewest possible number of repeats.

For many specific instances, these quantities can be estimated analytically. For random walks on graphs, much about the hitting time and target time is known through classical Markov chain theory (Levin & Peres, 2017). For cyclic sampling and random reshuffling respectively, one has $t_{\text{hit}} = |\mathcal{V}|$ and $t_{\text{hit}} \leq 2|\mathcal{V}|$ since each data point is visited exactly once every epoch and no re-shuffling occurs in the cyclic case. Under cyclic sampling, for fixed n and for each $1 \leq k \leq |\mathcal{V}|$ there is a v with $\mathbb{E}[\tau_{n,v}|\mathcal{F}_n] = k$. So $t_\odot$ is the largest possible value of $\sum_{v\in\mathcal{V}} \sigma_v \pi(v)$ where σ ranges over all permutations of $1, ..., |\mathcal{V}|$. In particular, if π is uniform, $t_\odot = \frac{|\mathcal{V}|-1}{2}$. For reshuffling with uniform π, $t_\odot$ is at least this large by considering a time n at the beginning of an epoch. But it still holds $t_\odot \leq 2|\mathcal{V}|$ since $t_\odot \leq t_{\text{hit}}$. If these quantities cannot be easily estimated analytically, they can be approximated using Monte-Carlo.

Remark A.6 (Comparison with i.i.d sampling). If the sequence (v_n) is formed by sampling vertices uniformly at random from $\mathcal{V}$ then, as previously mentioned, the return times $\tau_{n,v}$ are i.i.d geometric random variables with parameter $\frac{1}{|\mathcal{V}|}$. Then $\mathbb{E}[\tau_{n,v}|\mathcal{F}_n] = |\mathcal{V}|$ for each n and v so $t_\odot = |\mathcal{V}|$. Substituting $|\mathcal{V}|$ for $t_\odot$ in the optimal bound in Theorem 3.8 (ii) we recover the result given for MISO in the i.i.d setting in (Karimi et al., 2022) up to a factor of two. This is in-spite of the fact that our analytical approach is necessarily different to handle general recurrent sampling and shows that our results are tight.

Remark A.7 (Iterate stability and regularization). Here we give some remarks on the use of diminishing radius and proximal regularization in Algorithms 1 and 2.

The diminishing radius restriction in Algorithm 2 is a 'hard' regularization technique. It bakes necessary iterate stability directly into the problem by enforcing the stronger condition $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$. In comparison with proximal regularization, diminishing radius bounds the one step iterate difference by a deterministic quantity. Moreover, as is argued in the proof of Theorem 3.9 (ii) and the preceding propositions, sufficiently often the iterate $\boldsymbol{\theta}_n$ obtained by minimizing the $\bar{g}_n$ over the trust region in fact minimizes $\bar{g}_n$ over the entire feasible set $\boldsymbol{\Theta}$. This allows us to prove that limit points of the iterates produces

by Algorithm 2 are stationary even for general recurrent sampling schemes. However, diminishing radius introduces the drawback of needing to compute a projection at each step of the optimization process.

Compared to diminishing radius, proximal regularization is a form of 'soft' regularization. It is less restrictive than diminishing radius, but only allows us to derive a weaker form of iterate stability: we only have $\sum_{n=1}^{\infty}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 < \infty$ instead of the stronger $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$ which makes some aspects of the analysis slightly more challenging.

As mentioned in Section 3.2, the use of dynamic proximal regularization in Case 3.6 adapts to the sampling process and increasingly penalizes large values of $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|$ as the amount of time any vertex is left un-visited increases. The drawback is that we are unable to prove almost sure asymptotic convergence in Theorem 3.9 for Case 3.6 as we are for Case 3.7. If we could show $\rho_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \to 0$, then asymptotic convergence would follow. However ρ_n is not bounded and can take arbitrarily large values, albeit with low probability. But as we show in Lemma E.1, $\mathbb{E}[\rho_n]$ is uniformly bounded, which allows us to deduce $\mathbb{E}[\rho_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|] \to 0$ and leads to the L^1-convergence result in Theorem 3.9.

For case 3.5 it is relatively straightforward to show $\rho_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \to 0$ since ρ_n is constant. In this case, the difficulty lies in showing $\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\| \to 0$. We *are* able to prove this for Case 3.6 however because the use of dynamic proximal regularization allows us to show that the sequence $\max_{v\in\mathcal{V}}(n - k^v(n))\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2$ is summable. More detail is given in the proofs and discussion in subsection E.1.

B. Convergence analysis using mixing times

In this section, we give an overview of the standard pipeline for analyzing stochastic optimization algorithms with Markovian data and discuss the difficulties of applying these techniques to the analysis of MISO.

The analysis of first order methods such as SGD generally rely on conditionally unbiased gradient estimates. In the context of solving problem (1), this is to require

$$\mathbb{E}[\nabla f^{v_n}(\boldsymbol{\theta}_{n-1})|\mathcal{F}_{n-1}] = \nabla f(\boldsymbol{\theta}_{n-1}) \tag{35}$$

where $\mathcal{F}_n$ is the filtration of information up to time n. However, in the dependent data setting (35) does not hold which complicates the analysis significantly. Previous works (e.g (Sun et al., 2018; Bhandari et al., 2018; Nagaraj et al., 2020; Lyu et al., 2020; 2022; Lyu, 2023; Alacaoglu & Lyu, 2023)) use a "conditioning on the distant past" argument. Specifically, for SGD, one assumes that (v_n) is a Markov chain on state space $\mathcal{V}$ which mixes exponentially fast to its stationary distribution π with parameter λ. One then considers the quantity

$$\mathbb{E}[\nabla f^{v_n}(\boldsymbol{\theta}_{n-a_n})|\mathcal{F}_{n-a_n}] \tag{36}$$

where a_n is a slowly growing sequence satisfying $\sum_{n\geq 1}\lambda^{a_n} < \infty$. By further assuming either uniformly bounded gradients as in (Sun et al., 2018) or that the conditional expectations $\mathbb{E}[\|\nabla f^{v_{n+1}}(\boldsymbol{\theta})\||\mathcal{F}_n]$ are uniformly bounded as in (Alacaoglu & Lyu, 2023), one can show that

$$\|\nabla f(\boldsymbol{\theta}_{n-a_n}) - \mathbb{E}[\nabla f^{v_n}(\boldsymbol{\theta}_{n-a_n})|\mathcal{F}_{n-a_n}]\| = O(\lambda^{a_n}). \tag{37}$$

Using Lipschitz continuity of gradients, it is then established that

$$\|\nabla f(\boldsymbol{\theta}_n) - \mathbb{E}[\nabla f^{v_n}(\boldsymbol{\theta}_n)|\mathcal{F}_{n-a_n}]\| = O(\lambda^{a_n}) + O(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-a_n}\|), \tag{38}$$

allowing one to control the bias in the stochastic gradient estimate. Conventional analysis may then be used to prove

$$\sum_{n=1}^{\infty}\gamma_n\mathbb{E}\left[\langle\nabla f(\boldsymbol{\theta}_n), \nabla f^{v_n}(\boldsymbol{\theta}_n)\rangle\right] < \infty, \tag{39}$$

where γ_n denotes the stepsize at iteration n. Combining this with (38), it can finally be deduced that

$$\sum_{n=1}^{\infty}\gamma_n\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|^2] < \infty \tag{40}$$

for an appropriate stepsize γ_n which gives the desired convergence rate.

This technique was further developed to analyze convergence rates of *stochastic majorization minimization* (SMM) (Mairal, 2013) algorithms under Markovian sampling in (Lyu, 2023). In the context of problem (1), SMM proceeds by minimizing a recursively defined majorizing surrogate of the empirical loss function. The algorithm is stated concisely as follows:

$$(\textbf{SMM}): \begin{cases} \text{Sample } v_n \text{ from the conditional distribution } \pi(\cdot|\mathcal{F}_{n-1}) \\ g_n \leftarrow \text{Strongly convex majorizing surrogate of } f^{v_n} \\ \boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1-w_n)\bar{g}_{n-1}(\boldsymbol{\theta}) + w_n g_n(\boldsymbol{\theta})\right) \end{cases} \tag{41}$$

where $(w_n)_{n\geq 1}$ is a non-increasing sequence of weights. A crucial step in the analysis of SMM is to show that

$$\sum_{n=1}^{\infty} w_n \mathbb{E}[|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)|] < \infty \tag{42}$$

where $\bar{f}_n$ is the empirical loss function satisfying the recursion $\bar{f}_n(\boldsymbol{\theta}) := (1-w_n)\bar{f}_{n-1}(\boldsymbol{\theta}) + w_n f^{v_n}(\boldsymbol{\theta})$. To do so, the problem is reduced to showing that

$$\mathbb{E}\left[\mathbb{E}[f^{v_n}(\boldsymbol{\theta}_{n-a_n}) - \bar{f}_n(\boldsymbol{\theta}_{n-a_n})|\mathcal{F}_{n-a_n}]^+\right] = O(w_{n-a_n}) + O(\lambda^{a_n}) \tag{43}$$

(see Proposition 8.1 in (Lyu, 2023)) which is similar to (37). By additionally assuming Lipschitz continuity of the individual loss functions f^v, it is then shown that

$$\mathbb{E}\left[\mathbb{E}[f^{v_n}(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)|\mathcal{F}_{n-a_n}]\right] = O(w_{n-a_n}) + O(\lambda^{a_n}) + O(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-a_n}\|) \tag{44}$$

which may be compared with (38). Finally, (42) is proved by showing

$$\sum_{n=1}^{\infty} w_n \mathbb{E}[|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)|] \leq \bar{g}_1(\boldsymbol{\theta}_1) + \sum_{n=1}^{\infty} w_n \mathbb{E}[f^{v_n}(\boldsymbol{\theta}_n) - \bar{f}_n(\boldsymbol{\theta}_n)] \tag{45}$$

and using the bound (44) to conclude that the latter sum is finite.

MISO and its extensions proposed in this paper are similar to SMM in their use of the majorization-minimization principle, but contain a few key differences. Put shortly, the original implementation of MISO is

$$\textbf{(MISO)}: \begin{cases} \text{Sample } v_n \text{ from the conditional distribution } \pi(\cdot|\mathcal{F}_{n-1}) \\ g_n^{v_n} \leftarrow \text{Convex surrogate of } f^{v_n} \text{ at } \boldsymbol{\theta}_{n-1};\ g_n^v = g_{n-1}^v \text{ for } v \neq v_n \\ \boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := \sum_{v\in\mathcal{V}} g_n^v \pi(v)\right) \end{cases} \tag{46}$$

In contrast with SMM, each surrogate defining $\bar{g}_n$ is given a constant weight. Consequently, the additional control provided by the decreasing weights (w_n) in SMM is not present, which makes the adaptation of the techniques used in the analysis of SMM non-trivial.

A key step in the original analysis of MISO given in (Mairal, 2015) is to show

$$\sum_{n=1}^{\infty} \mathbb{E}[\bar{h}_n(\boldsymbol{\theta}_n)] < \infty. \tag{47}$$

This is similar to (42) and shows that the averaged surrogate is an asymptotically accurate approximation of the true objective at all $\boldsymbol{\theta}_n$s. To prove this, one needs to relate the averaged error $\bar{h}_n(\boldsymbol{\theta}_n)$ to another quantity proven to be summable through other means. This is, in abstract, the role that mixing rate analysis plays for SGD and SMM. Using techniques from (Mairal, 2015) one can prove

$$\sum_{n=1}^{\infty} \mathbb{E}[h_n^{v_{n+1}}(\boldsymbol{\theta}_n)] < \infty. \tag{48}$$

If the (v_n) are drawn i.i.d from π, or more generally if the probability of transitioning between any two vertices is uniformly bounded below by a positive quantity, then $h_n^{v_{n+1}}(\boldsymbol{\theta}_n)$ is a conditionally unbiased estimate of $\bar{h}_n(\boldsymbol{\theta}_n)$ up to a constant. Then (47) follows by conditioning on the most recent information $\mathcal{F}_n$.

To adapt the analysis to a more general setting, one may attempt to use Markov chain mixing to show

$$|\bar{h}_n(\boldsymbol{\theta}_{n-a_n}) - \mathbb{E}[h_n^{v_{n+1}}(\boldsymbol{\theta}_{n-a_n})|\mathcal{F}_{n-a_n}]| = O(\lambda^{a_n}) \tag{49}$$

similar to (37) and (43). However, an additional complication arises because the function $h_n^{v_{n+1}}$ conditional on $\mathcal{F}_{n-a_n}$ depends on both the history of data samples v_k *and* the estimated parameters $\boldsymbol{\theta}_k$ for $n - a_n \leq k \leq n$. In contrast, f^{v_n} for SGD (in (37)) depends only on the last data sample v_n and $\bar{f}_n$ for SMM (in (43)) depends only on the history of data samples $v_{n-a_n}, \ldots, v_n$. So, one cannot use the Markov property to isolate the randomness in $h_n^{v_{n+1}}(\boldsymbol{\theta}_{n-a_n})$ due to the Markov chain transition over the interval $[n-a_n, n+1]$. To alleviate this problem, one may attempt to control the difference

$$|h_n^{v_{n+1}}(\boldsymbol{\theta}_{n-a_n}) - h_{n-a_n}^{v_{n+1}}(\boldsymbol{\theta}_{n-a_n})| \tag{50}$$

but doing so is not straightforward without access to something resembling the weights (w_n) in SMM.

C. Preliminary Lemmas

In this section we state and prove some preliminary lemmas which will be used to prove in the proofs of both Theorem 3.8 and 3.9.

We first introduce some additional notation. Throughout this section as well as the remainder of the paper we let

$$t_{\text{cov}} := \sup_{n\geq 1} \left\| \mathbb{E}\left[\max_{v\in\mathcal{V}} \tau_{n,v} \middle| \mathcal{F}_n \right] \right\|_\infty . \tag{51}$$

We recall that the L_∞ norm here is taken for the conditional expectation viewed as a random variable. This quantity is a generalization of the worst case expected cover time from Markov chain theory (Levin & Peres, 2017) and will be important in the analysis of Algorithm 1 with dynamic proximal regularization.

Our first result, Proposition C.1, states that under Assumption 3.1 the return times have finite moments of all orders and gives an upper-bound on t_{cov} in terms of t_{hit}. The first item will is used in Section E to prove asymptotic results while the second is used in the proof of iteration complexity for Case 3.6.

Proposition C.1 (Recurrence implies finite exponential moments of return time)**.** *Let Assumption 3.1 hold and let t_{hit} and t_{cov} be as in (8) and (51) respectively.*

(i) There exists $s_0 > 0$ so that for all $0 < s < s_0$ there is a constant $C_s > 0$ with

$$\max_{v\in\mathcal{V}} \sup_{n\geq 1} \mathbb{E}\left[e^{s\tau_{n,v}} \middle| \mathcal{F}_n \right] \leq C_s < \infty. \tag{52}$$

Consequently, for each $p \geq 1$, $\max_{v\in\mathcal{V}} \sup_{n\geq 1} \mathbb{E}[\tau_{n,v}^p|\mathcal{F}_n] \leq C < \infty$ for some $C > 0$.

(ii) We have the following bound on t_{cov}:

$$t_{cov} \leq (2t_{hit} + 1)\log_2(4|\mathcal{V}|). \tag{53}$$

Proof. Let m be the smallest integer satisfying $m \geq 2t_{\text{hit}}$. For any $n \geq 1$ and $v \in \mathcal{V}$ notice that if $\tau_{n,v} \geq km$ then we must have $\tau_{n+jm,v} \geq m$ for each $0 \leq j \leq k-1$. Then

$$\mathbb{P}(\tau_{n,v} \geq km|\mathcal{F}_n) \leq \mathbb{P}\left(\bigcap_{j=0}^{k-1} \{\tau_{n+jm,v} \geq m\} \middle| \mathcal{F}_n \right) = \mathbb{E}\left[\prod_{j=0}^{k-1} \mathbb{1}(\tau_{n+jm,v} \geq m) \middle| \mathcal{F}_n \right]. \tag{54}$$

We have

$$\mathbb{E}\left[\prod_{j=0}^{k-1} \mathbb{1}(\tau_{n+jm,v} \geq m) \middle| \mathcal{F}_n \right] = \mathbb{E}\left[\mathbb{E}\big[\mathbb{1}(\tau_{n+(k-1)m,v} \geq m)|\mathcal{F}_{n+(k-1)m}\big] \prod_{j=0}^{k-2} \mathbb{1}(\tau_{n+jm,v} \geq m) \middle| \mathcal{F}_n \right] \tag{55}$$

where we have used that $\{\tau_{n+jm,v} \geq m\}$ is measurable with respect to $\mathcal{F}_{n+(k-1)m}$ for each $j \leq k-2$. By Markov's inequality and Assumption 3.1

$$\mathbb{E}\big[\mathbb{1}(\tau_{n+(k-1)m,v} \geq m)|\mathcal{F}_{n+(k-1)m}\big] = \mathbb{P}\left(\tau_{n+(k-1)m,v} \geq m|\mathcal{F}_{n+(k-1)m}\right) \leq \frac{t_{\text{hit}}}{m}. \tag{56}$$

So

$$\mathbb{E}\left[\prod_{j=0}^{k-1}\mathbb{1}(\tau_{n+jm,v} \geq m)\Big|\mathcal{F}_n\right] \leq \frac{t_{\text{hit}}}{m}\mathbb{E}\left[\prod_{j=0}^{k-2}\mathbb{1}(\tau_{n+jm,v} \geq m)\Big|\mathcal{F}_n\right]. \tag{57}$$

Proceeding by induction it follows that

$$\mathbb{P}(\tau_{n,v} \geq km|\mathcal{F}_n) \leq \left(\frac{t_{\text{hit}}}{m}\right)^k \leq 2^{-k} \tag{58}$$

with the second inequality using our choice of m. Now,

$$\mathbb{E}\left[e^{s\tau_{n,v}}|\mathcal{F}_n\right] = \sum_{\ell=1}^{\infty} e^{s\ell}\mathbb{P}(\tau_{n,v} = \ell|\mathcal{F}_n) \leq \sum_{k=0}^{\infty} e^{s(k+1)m}\mathbb{P}(\tau_{n,v} \geq km|\mathcal{F}_n) \leq \sum_{k=0}^{\infty} e^{s(k+1)m}2^{-k}. \tag{59}$$

The latter sum is finite if $s < \frac{\log 2}{2m}$ and does not depend on n or v which shows (i).

With n still fixed let $\tau_{\text{cov}} = \max_{v\in\mathcal{V}} \tau_{n,v}$. We have

$$\mathbb{E}[\tau_{\text{cov}}|\mathcal{F}_n] = \sum_{\ell=1}^{\infty}\mathbb{P}(\tau_{\text{cov}} \geq \ell|\mathcal{F}_n) \leq \sum_{k=0}^{\infty} m\mathbb{P}(\tau_{\text{cov}} \geq km|\mathcal{F}_n) \leq (2t_{\text{hit}}+1)\sum_{k=0}^{\infty}\mathbb{P}(\tau_{\text{cov}} \geq km|\mathcal{F}_n) \tag{60}$$

since $\mathbb{P}(\tau_{\text{cov}} \geq \ell|\mathcal{F}_n)$ is a decreasing function of ℓ and $m \leq 2t_{\text{hit}}+1$. By a union bound

$$\mathbb{P}(\tau_{\text{cov}} \geq km|\mathcal{F}_n) \leq 1 \wedge \sum_{v\in\mathcal{V}}\mathbb{P}(\tau_{n,v} \geq km|\mathcal{F}_n) \leq 1 \wedge |\mathcal{V}|2^{-k}. \tag{61}$$

Summing a geometric series we get

$$\sum_{k=0}^{\infty}\mathbb{P}(\tau_{\text{cov}} \geq km|\mathcal{F}_n) \leq \log_2|\mathcal{V}| + |\mathcal{V}|\sum_{k>\log_2|\mathcal{V}|} 2^{-k} \leq \log_2|\mathcal{V}| + 2. \tag{62}$$

Combining (60) and (62) shows (ii). □

The next proposition states some general properties of first order surrogate functions.

Proposition C.2 (Properties of Surrogates)**.** *Fix $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ and $f : \boldsymbol{\Theta} \to \mathbb{R}$. Let $g \in \mathcal{S}_L(f, \bar{\boldsymbol{\theta}})$ and let $\boldsymbol{\theta}'$ be a minimizer of g over $\boldsymbol{\Theta}$. Then for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*

$$|h(\boldsymbol{\theta})| \leq \frac{L}{2}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2 \tag{63}$$

Proof. This follows from using the classical upperbound for L-smooth functions

$$h(\boldsymbol{\theta}) \leq h(\bar{\boldsymbol{\theta}}) + \langle\nabla h(\bar{\boldsymbol{\theta}}), \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2 \tag{64}$$

(see Lemma F.1) and noting that $h(\bar{\boldsymbol{\theta}})$ and $\nabla h(\bar{\boldsymbol{\theta}})$ are both equal to zero according to Definition 2.1. □

Next, we show that the surrogate objective value $\bar{g}_n(\boldsymbol{\theta}_n)$ evaluated at $\boldsymbol{\theta}_n$ is non-increasing.

Lemma C.3 (Surrogate Monotonicity)**.** *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of either Algorithm 1 or 2 . Then $\bar{g}_n(\boldsymbol{\theta}_{n-1}) \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1})$ for all $n \geq 1$. Moreover, the sequence $(\bar{g}_n(\boldsymbol{\theta}_n))_{n\geq 0}$ is non-increasing. As a consequence, by Assumption 3.2 and Definition 2.1, $(\bar{g}_n(\boldsymbol{\theta}_n))_{n\geq 0}$ is bounded below with probability one and therefore $\lim_{n\to\infty} \bar{g}_n(\boldsymbol{\theta}_n)$ exists almost surely.*

Proof. Since $g_n^{v_n} \in \mathcal{S}_L(f^{v_n}, \boldsymbol{\theta}_{n-1})$, Definition 2.1 implies that $g_n^{v_n}(\boldsymbol{\theta}_{n-1}) = f^{v_n}(\boldsymbol{\theta}_{n-1})$. Then

$$\bar{g}_n(\boldsymbol{\theta}_{n-1}) = \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) + \Big[g_n^{v_n}(\boldsymbol{\theta}_{n-1}) - g_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\Big]\pi(v_n) \tag{65}$$

$$= \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) + \Big[f^{v_n}(\boldsymbol{\theta}_{n-1}) - g_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\Big]\pi(v_n) \tag{66}$$

$$\leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) \tag{67}$$

where the last inequality used $g_{n-1}^{v_n}$ is a majorizing surrogate of f^{v_n}.

Suppose now that $(\boldsymbol{\theta}_n)_{n\geq 0}$ is an output of Algorithm 1. Then by definition of $\boldsymbol{\theta}_n$,

$$\begin{aligned}
\bar{g}_n(\boldsymbol{\theta}_n) &\leq \bar{g}_n(\boldsymbol{\theta}_n) + \frac{\rho_n}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \\
&\leq \bar{g}_n(\boldsymbol{\theta}_{n-1}) + \frac{\rho_n}{2}\|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_{n-1}\|^2 \\
&= \bar{g}_n(\boldsymbol{\theta}_{n-1}) \\
&\leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}).
\end{aligned}$$

If instead $(\boldsymbol{\theta}_n)_{n\geq 1}$ is an output of Algorithm 2, then we can directly conclude $\bar{g}_n(\boldsymbol{\theta}_n) \leq \bar{g}_n(\boldsymbol{\theta}_{n-1})$ by definition of $\boldsymbol{\theta}_n$. The remainder of the proof is identical to the above. □

The next lemma establishes the summability of the sequence $h_n^{v_{n+1}}(\boldsymbol{\theta}_n)$. This was used in (Mairal, 2015) to prove asymptotic convergence of MISO under i.i.d. sampling. We use it primarily in the analysis of Algorithm 2.

Lemma C.4. *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of either Algorithm 1 or 2. Then almost surely*

$$\sum_{n=1}^{\infty} h_n^{v_{n+1}}(\boldsymbol{\theta}_n) \leq \frac{1}{\pi_{min}}\Delta_0. \tag{68}$$

Proof. By Definition 2.1, for each n the quantity $h_n^{v_{n+1}}(\boldsymbol{\theta}_n)$ is non-negative. Therefore, it suffices to show that the sequence of partial sums, $\sum_{n=1}^{N} h_n^{v_{n+1}}(\boldsymbol{\theta}_n)$ is uniformly bounded.

Recall that

$$\bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) \leq \bar{g}_{n+1}(\boldsymbol{\theta}_n) = \bar{g}_n(\boldsymbol{\theta}_n) + (g_{n+1}^{v_{n+1}}(\boldsymbol{\theta}_n) - g_n^{v_{n+1}}(\boldsymbol{\theta}_n))\pi(v_{n+1}) \tag{69}$$

$$= \bar{g}_n(\boldsymbol{\theta}_n) + (f^{v_{n+1}}(\boldsymbol{\theta}_n) - g_n^{v_{n+1}}(\boldsymbol{\theta}_n))\pi(v_{n+1}). \tag{70}$$

We then have

$$\begin{aligned}
\sum_{n=1}^{N} h_n^{v_{n+1}}(\boldsymbol{\theta}_n) &= \sum_{n=1}^{N} g_n^{v_{n+1}}(\boldsymbol{\theta}_n) - f^{v_{n+1}}(\boldsymbol{\theta}_n) \\
&\leq \sum_{n=1}^{N} \frac{1}{\pi(v_{n+1})}\left(\bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1})\right) \\
&\leq \frac{1}{\pi_{\min}} \sum_{n=1}^{N} \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_{n+1}(\boldsymbol{\theta}_{n+1}) \\
&\leq \frac{1}{\pi_{\min}}\Delta_0
\end{aligned}$$

which is what we needed to show. □

Proposition C.5. *Suppose $(\boldsymbol{\theta}_n)_{n\geq 0}$ is an output of Algorithm 2. The for all $n \geq 1$, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$.*

Proof. This follows directly from the definition of $\boldsymbol{\theta}_n$ in Algorithm 2. □

Lemma C.6 establishes the iterate stability. These results are crucially used to control the surrogate error gradient $\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|$ in Lemma D.1 as well as in the asymptotic analysis of RMISO in Section E.

Lemma C.6 (Finite variation of iterate differences)**.** *The following hold almost surely:*

(i) For Case 3.7,

$$\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \sum_{n=1}^{\infty} r_n^2 < \infty. \tag{71}$$

(ii) In either of the Cases 3.5 or 3.6,

$$\sum_{n=1}^{\infty} \frac{\rho_n + \mu}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \Delta_0 < \infty. \tag{72}$$

Proof. The proof of (i) can be deduced from Proposition C.5 and Assumption 3.4.

Now assume either of the Cases 3.5 or 3.6. Define $G_n(\boldsymbol{\theta}) = \bar{g}_n(\boldsymbol{\theta}) + \frac{\rho_n}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|$. Then G_n is $\rho_n + \mu$ strongly convex. Since θ_n is a minimizer of G_n over $\boldsymbol{\Theta}$ we get

$$G_n(\boldsymbol{\theta}_n) + \frac{\rho_n + \mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq G_n(\boldsymbol{\theta}_{n-1}) = \bar{g}_n(\boldsymbol{\theta}_{n-1}) \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) \tag{73}$$

where the last inequality is due to Lemma C.3. So

$$\frac{\rho_n + \mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - G_n(\boldsymbol{\theta}_n) \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n). \tag{74}$$

Hence,

$$\sum_{n=1}^{N} \frac{\rho_n + \mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \sum_{n=1}^{N} \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) = \bar{g}_0(\boldsymbol{\theta}_0) - \bar{g}_N(\boldsymbol{\theta}_N) \leq \Delta_0. \tag{75}$$

Letting $N \to \infty$ shows that

$$\sum_{n=1}^{\infty} \frac{\rho_n + \mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \Delta_0 \tag{76}$$

as desired. This shows (ii). □

The remaining results in this section concern Algorithm 2 and are used both in the convergence rate analysis in Section D as well as the asymptotic analsyis of Section E. Recall that in this case we are assuming that ∇f^v is L-Lipschitz continuous for each $v \in V$. Proposition C.7 states that this assumption implies $\nabla \bar{g}_n$ is differentiable and $2L$ Lipschitz for each n.

Proposition C.7. *Let $\{\boldsymbol{\theta}_v\}_{v\in\mathcal{V}}$ be a collection of $|\mathcal{V}|$ points in $\boldsymbol{\Theta}$. Suppose that Assumption 3.3 holds and that $g_n^v \in \mathcal{S}_L(f^v, \boldsymbol{\theta}_v)$ for each v. Then*

(i) The gradient of the objective function $\nabla f = \sum_{v\in\mathcal{V}} \nabla f^v \pi(v)$ is L-Lipschitz over $\boldsymbol{\Theta}$.

(ii) For each v, ∇g_n^v is $2L$-Lipchitz over $\boldsymbol{\Theta}$. In addition, $\nabla \bar{g}_n$ is $2L$-Lipchitz.

Proof. Since π is a probability distribution, (i) follows easily from the triangle inequality.

For (ii) note that $\nabla(g_n^v - f^v) = \nabla h_n^v$ is L-Lipshitz by Definition 2.1. Then since $\nabla g_n^v = \nabla h_n^v + \nabla f^v$ it follows from the triangle inequality that ∇g_n^v is $2L$-Lipschitz. Then recalling that $\nabla \bar{g}_n = \sum_{v\in\mathcal{V}} \nabla g_n^v \pi(v)$ another application of the triangle inequality shows that $\nabla \bar{g}_n$ is $2L$-Lipschitz continuous. □

Proposition C.8. *Assume Case 3.7 and let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 2. Then*

$$\sum_{n=1}^{N} |\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle| \leq \Delta_0 + L\sum_{n=1}^{N} r_n^2 \tag{77}$$

almost surley.

Proof. Since $\nabla \bar{g}_n$ is $2L$-Lipschitz continuous by Proposition C.7, by Lemma F.1

$$|\bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle| \leq L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2. \tag{78}$$

Under Algorithm 2 we have $\bar{g}_n(\boldsymbol{\theta}_n) \leq \bar{g}_n(\boldsymbol{\theta}_{n-1})$ by the definition of $\boldsymbol{\theta}_n$ and $\bar{g}_n(\boldsymbol{\theta}_{n-1}) \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1})$ by Lemma C.3. These observations together with (78) imply

$$\begin{aligned}
|\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle| &\leq |\bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n)| + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &(79)\\
&= \bar{g}_n(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &(80)\\
&\leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2. &(81)
\end{aligned}$$

We have

$$\sum_{n=1}^{N} \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) \leq \Delta_0 \tag{82}$$

almost surely. Therefore

$$\begin{aligned}
\sum_{n=1}^{N} |\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle| &\leq \sum_{n=1}^{N} \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n) + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &(83)\\
&\leq \Delta_0 + L\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &(84)\\
&\leq \Delta_0 + L\sum_{n=1}^{N} r_n^2, &(85)
\end{aligned}$$

where the last line uses $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$. □

The next lemma is a key to establishing iteration complexity of Algorithm 2. A similar lemma was used to analyze block majorization-minimization with diminishing radius in (Lyu & Li, 2023).

Lemma C.9 (Approximate first order optimality)**.** *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 2 and let $b_n = \min\{1, r_n\}$. Then*

$$b_n \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\|\leq 1} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\rangle \leq \langle -\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle + r_n\|\nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\| + 2Lr_n^2. \tag{86}$$

Proof. Fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$. By definition of $\boldsymbol{\theta}_n$ we have $\bar{g}_n(\boldsymbol{\theta}_n) \leq \bar{g}_n(\boldsymbol{\theta})$. Subtracting $\bar{g}_n(\boldsymbol{\theta}_{n-1})$ from both sides and using Proposition C.7 and Lemma F.1,

$$\begin{aligned}
\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\rangle - L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 &\leq \bar{g}_n(\boldsymbol{\theta}_n) - \bar{g}_n(\boldsymbol{\theta}_{n-1}) &(87)\\
&\leq \bar{g}_n(\boldsymbol{\theta}) - \bar{g}_n(\boldsymbol{\theta}_{n-1}) &(88)\\
&\leq \langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\rangle + L\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2. &(89)
\end{aligned}$$

Notice that

$$\begin{aligned}
\nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}) &= \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) + \left[\nabla g_n^{v_n}(\boldsymbol{\theta}_{n-1}) - \nabla g_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\right]\pi(v_n) &(90)\\
&= \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) + \left[\nabla f^{v_n}(\boldsymbol{\theta}_{n-1}) - \nabla g_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1}\right]\pi(v_n) &(91)\\
&= \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\pi(v_n). &(92)
\end{aligned}$$

where the second line used $g_n^{v_n} \in \mathcal{S}_L(f^{v_n}, \boldsymbol{\theta}_{n-1})$ and item (ii) of Definition 2.1, and the third line used the definition of $h_{n-1}^{v_n}$. Therefore, adding and subtracting $\langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle$ from the right hand side of (89) we get

$$\begin{aligned}
\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle &\leq \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle - \pi(v_n) \langle \nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + L\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 && (93)\\
&\quad + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 && (94)\\
&\leq \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \|\nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\| \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| + L\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 && (95)\\
&\quad + L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 && (96)\\
&\leq \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + r_n \|\nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\| + 2Lr_n^2 && (97)
\end{aligned}$$

where the last line used $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n \leq r_n$. Since the above holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n$ we obtain

$$\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1} \rangle \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + r_n \|\nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\| + 2Lr_n^2. \tag{98}$$

Finally notice that since $b_n \leq 1$, the convexity of $\boldsymbol{\Theta}$ implies that $\boldsymbol{\theta}_{n-1} + b_n(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}) \in \boldsymbol{\Theta}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, if there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq 1$ there is $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta}' - \boldsymbol{\theta}_{n-1}\| \leq b_n$ such that the direction of $\boldsymbol{\theta}' - \boldsymbol{\theta}_{n-1}$ agrees with that of $\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}$. Therefore

$$\begin{aligned}
b_n \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq 1} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle &= \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq 1} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), b_n(\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}) \rangle && (99)\\
&\leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq b_n} \langle -\nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle. && (100)
\end{aligned}$$

combining this with (98) we complete the proof. □

D. Convergence Rate Analysis

In this subsection we prove the convergence rate guarantees of Theorem 3.8.

D.1. The key lemma

First we state and prove Lemma D.1 which lies at the heart of our analysis. It allows us to relate the surrogate error gradient $\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|$ to the sequence of parameter differences $(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2)$ which is known to be summable by Lemma C.6. It is important to note that we only use the recurrence of data sampling Assumption 3.1 and the structure of the algorithm in the proof.

Lemma D.1 (Key lemma)**.** *Let $(c_n)_{n \geq 1}$ be a non-increasing sequence of positive numbers. For any of the cases 3.5-3.7 and any $v \in V$,*

$$\mathbb{E}\left[\sum_{n=1}^{N} c_n \|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq L t_{\odot} \left(\sum_{n=1}^{N} c_n^2\right)^{1/2} \mathbb{E}\left[\left(\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2}\right]. \tag{101}$$

Proof. Fix some $v \in V$. We recall that $k^v(n)$ is the last time before n that the sampling process visited v and therefore the last time the surrogate g_n^v was updated. We then have $g_n^v \in \mathcal{S}_L(f^v, \boldsymbol{\theta}_{k^v(n)-1})$ so by the definition of first-order surrogates (Definition 2.1) $\nabla h_n^v(\boldsymbol{\theta}_{k^v(n)-1}) = 0$. Combining this with the Lipschitz continuity of ∇h_n^v we get

$$c_n \|\nabla h_n^v(\boldsymbol{\theta}_n)\| = c_n \|\nabla h_n^v(\boldsymbol{\theta}_n) - \nabla h_n^v(\boldsymbol{\theta}_{k^v(n)-1})\| \leq L c_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{k^v(n)-1}\| \leq L \sum_{i=k^v(n)}^{n} c_n \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|. \tag{102}$$

Therefore by the triangle inequality,

$$c_n \|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\| \leq c_n \sum_{v \in \mathcal{V}} \|\nabla h_n^v(\boldsymbol{\theta}_n)\| \pi(v) \leq L \sum_{v \in \mathcal{V}} \left(\sum_{i=k^v(n)}^{n} c_n \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|\right) \pi(v). \tag{103}$$

For an integer n, let $p^v(n) = \inf\{j > n : v_j = v\}$ be the first time strictly after n that the sampling algorithm visits v. Denote $a \wedge b := \min(a, b)$. We have

$$\sum_{n=1}^{N}\sum_{v\in\mathcal{V}}\left(\sum_{i=k^v(n)}^{n} c_n\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|\right)\pi(v) = \sum_{v\in\mathcal{V}}\left(\sum_{n=1}^{N}\sum_{i=k^v(n)}^{n} c_n\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|\right)\pi(v) \tag{104}$$

$$= \sum_{v\in\mathcal{V}}\left(\sum_{i=1}^{N}\sum_{n=i}^{N\wedge p^v(i)-1} c_n\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|\right)\pi(v) \tag{105}$$

$$\leq \sum_{v\in\mathcal{V}}\left(\sum_{i=1}^{N} c_i\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|(p^v(i) - i)\right)\pi(v) \tag{106}$$

$$= \sum_{v\in\mathcal{V}}\left(\sum_{i=1}^{N} c_i\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|\tau_{i,v}\right)\pi(v), \tag{107}$$

where the third line used that (c_n) is non-increasing. So we get

$$\mathbb{E}\left[\sum_{n=1}^{N} c_n\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq L\mathbb{E}\left[\sum_{v\in\mathcal{V}}\left(\sum_{n=1}^{N} c_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\tau_{n,v}\right)\pi(v)\right] \tag{108}$$

$$= L\mathbb{E}\left[\sum_{n=1}^{N} c_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\left(\sum_{v\in V}\tau_{n,v}\pi(v)\right)\right] \tag{109}$$

$$= L\mathbb{E}\left[\sum_{n=1}^{N} c_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\left(\sum_{v\in\mathcal{V}}\mathbb{E}[\tau_{n,v}|\mathcal{F}_n]\pi(v)\right)\right] \tag{110}$$

$$\leq Lt_{\odot}\mathbb{E}\left[\sum_{n=1}^{N} c_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\right] \tag{111}$$

$$\leq Lt_{\odot}\left(\sum_{n=1}^{N} c_n^2\right)^{1/2}\mathbb{E}\left[\left(\sum_{n=1}^{N}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2}\right] \tag{112}$$

with the second to last line using Assumption 3.1 and the last using Cauchy-Schwartz. □

D.2. The constant proximal regularization case 3.5

In this section we prove Theorem A.1 for Case 3.5.

Proof of Theorem A.1 for case 3.5. We first use the linearity of the limit and the differentiability of the average surrogate error $\bar{h}_n$ to get

$$|\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n) - \nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)| = \left|\langle\nabla\bar{h}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle\right| \leq \|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq \|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\| \tag{113}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1$. It the follows from the triangle inequality, taking supremums, and then expectations that

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] + \mathbb{E}\left[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|\right]. \tag{114}$$

Our goal will be to control the sum of right hand side.

We first address the first term on the right hand side of (114). Recall that in this case we are using constant proximal regularization, so $\rho_n \equiv \rho$ for some $\rho \geq 0$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n) + \rho\langle\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}, \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle \geq 0 \tag{115}$$

since $\boldsymbol{\theta}_n$ is a minimizer of $\bar{g}_n(\boldsymbol{\theta}) + \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2$ over $\boldsymbol{\Theta}$. Then

$$-\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n) \leq \rho\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}, \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle \leq \rho\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq \rho\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \tag{116}$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq 1$. Therefore,

$$\mathbb{E}\left[\sum_{n=1}^{N} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \rho\mathbb{E}\left[\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\right] \tag{117}$$

$$\leq \rho\sqrt{N}\mathbb{E}\left[\left(\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2}\right] \tag{118}$$

where we used the Cauchy-Schwartz inequality in the last line. By Lemma C.6

$$\left(\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2} \leq \sqrt{\frac{2\Delta_0}{\rho + \mu}} \tag{119}$$

almost surely. Thus,

$$\mathbb{E}\left[\sum_{n=1}^{N} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \rho\sqrt{\frac{2N\Delta_0}{\rho + \mu}}. \tag{120}$$

We now turn to the second term on the right hand side of (114). By Lemma D.1 with $c_n = 1$ and Lemma C.6

$$\mathbb{E}\left[\sum_{n=1}^{N} \|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq \sqrt{N}Lt_{\odot}\mathbb{E}\left[\left(\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2}\right] \leq \sqrt{\frac{2N\Delta_0}{\rho + \mu}}Lt_{\odot}. \tag{121}$$

Now, summing both sides of (114) and using (120) and (121),

$$\sum_{n=1}^{N} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] = \mathbb{E}\left[\sum_{n=1}^{N} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \tag{122}$$

$$\leq \sqrt{2N\Delta_0}\left(\frac{\rho}{\sqrt{\rho + \mu}} + \frac{Lt_{\odot}}{\sqrt{\rho + \mu}}\right). \tag{123}$$

This shows

$$\min_{1\leq n\leq N} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \frac{\sqrt{2\Delta_0}\left(\frac{\rho}{\sqrt{\rho+\mu}} + \frac{Lt_{\odot}}{\sqrt{\rho+\mu}}\right)}{\sqrt{N}}. \tag{124}$$

□

D.3. The dynamic proximal regularization case 3.6

In this section we prove Theorem A.1 for Case 3.6. Recall the definition of t_{cov} from (51). Before proving the theorem we introduce a Lemma adapted from (Even, 2023) Lemma A.5. This is used to bound the expected sum of the first N dynamic regularization parameters ρ_n in terms of t_{cov}.

Lemma D.2. *Let* $a_n = \max_{v\in V}(n - k^v(n))$. *Then*

$$\sum_{n=1}^{N} \mathbb{E}[a_n] \leq Nt_{cov}. \tag{125}$$

Proof. Since $a_n \leq n-1$ we have

$$\sum_{n=1}^{N} \mathbb{E}[a_n] = \sum_{n=1}^{N} \sum_{i=1}^{n-1} \mathbb{P}(a_n \geq i). \tag{126}$$

Let $b_n = \max_{v \in \mathcal{V}} \tau_{n,v}$. We note that if $a_n \geq i$ then there is $v \in V$ with $v_j \neq v$ for all $n-i < j \leq n$ and so $\tau_{n-i,v} \geq i$. So we have the inclusion $\{a_n \geq i\} \subseteq \{b_{n-i} \geq i\}$. Therefore

$$\sum_{n=1}^{N} \sum_{i=1}^{n-1} \mathbb{P}(a_n \geq i) = \sum_{i=1}^{N} \sum_{n=i+1}^{N} \mathbb{P}(a_n \geq i) \tag{127}$$

$$\leq \sum_{i=1}^{N} \sum_{n=i+1}^{N} \mathbb{P}(b_{n-i} \geq i) \tag{128}$$

$$= \sum_{s=1}^{N-1} \sum_{t=1}^{N-s} \mathbb{P}(b_s \geq t) \tag{129}$$

$$\leq \sum_{s=1}^{N-1} \sum_{t=1}^{\infty} \mathbb{P}(b_s \geq t) \tag{130}$$

$$\leq N \sup_{n \geq 1} \mathbb{E}[b_n]. \tag{131}$$

We have

$$\mathbb{E}[b_n] = \mathbb{E}\left[\mathbb{E}\left[\max_{v \in \mathcal{V}} \tau_{n,v} \middle| \mathcal{F}_n\right]\right] \leq t_{\text{cov}} \tag{132}$$

so we are done. □

Proof of Theorem A.1 for Case 3.6. The proof in this case follows the same strategy as Case 3.5, but is slightly more complicated due to the randomness of the dynamic proximal regularization parameter.

Define $\delta_n := \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n)$. By optimality of $\boldsymbol{\theta}_n$ and Lemma C.3,

$$\bar{g}_n(\boldsymbol{\theta}_n) + \frac{\rho_n}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) \tag{133}$$

so $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq \sqrt{2\rho_n^{-1}\delta_n}$. Using similar reasoning as in the proof for case 3.5

$$-\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n) \leq \rho_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq \sqrt{2\rho_n \delta_n}. \tag{134}$$

We have using Cauchy-Schwartz twice,

$$\sum_{n=1}^{N} \mathbb{E}[\sqrt{\rho_n \delta_n}] \leq \sum_{n=1}^{N} (\mathbb{E}[\rho_n])^{1/2} (\mathbb{E}[\delta_n])^{1/2} \tag{135}$$

$$\leq \left(\sum_{n=1}^{N} \mathbb{E}[\rho_n]\right)^{1/2} \left(\sum_{n=1}^{N} \mathbb{E}[\delta_n]\right)^{1/2} \tag{136}$$

$$\leq \sqrt{N(\rho + t_{\text{cov}})\Delta_0}. \tag{137}$$

The last inequality here uses $\rho_n = \rho + \max_{v \in V}(n - k^v(n))$ and Lemma D.2 as well as $\sum_{n=1}^{N} \delta_n \leq \Delta_0$ a.s. It follows that

$$\sum_{n=1}^{N} \mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \leq 1} -\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \sqrt{2N(\rho + t_{\text{cov}})\Delta_0}. \tag{138}$$

To handle the gradient error $\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|$ we first use Lemma C.6 and $\rho_n \geq \rho$ to conclude

$$\sum_{n=1}^{N} \frac{\rho+\mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \sum_{n=1}^{N} \frac{\rho_n+\mu}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \Delta_0 \tag{139}$$

almost surely. It then follows from Lemma D.1 that

$$\sum_{n=1}^{N} \mathbb{E}\left[\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq \sqrt{N} L t_{\odot} \mathbb{E}\left[\left(\sum_{n=1}^{N}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|\right)^{1/2}\right] \leq \sqrt{\frac{2N\Delta_0}{\rho}} L t_{\odot}. \tag{140}$$

Finally, combining (138) and (140) we get

$$\sum_{n=1}^{N} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right] \leq \sqrt{2N\Delta_0}\left(\sqrt{\rho + t_{\text{cov}}} + \frac{L t_{\odot}}{\sqrt{\rho+\mu}}\right) \tag{141}$$

and so we deduce

$$\min_{1\leq n\leq N} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right] \leq \frac{\sqrt{2\Delta_0}\left(\sqrt{\rho + t_{\text{cov}}} + \frac{L t_{\odot}}{\sqrt{\rho+\mu}}\right)}{\sqrt{N}}. \tag{142}$$

We complete the proof by substituting the bound for t_{cov} in Proposition C.1. □

D.4. The diminishing radius case 3.7

Here we prove Theorem A.1 for Case 3.7.

Proof of Theorem A.1 for Case 3.7. Similar to the proof in Case 3.5 we have

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}-\boldsymbol{\theta}_n\rangle\right] \leq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta}-\boldsymbol{\theta}_n\rangle\right] + \mathbb{E}\left[\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \tag{143}$$

Let $b_n = r_n \wedge 1$. Then by Lemma C.9

$$\sum_{n=1}^{N} b_{n+1}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta}-\boldsymbol{\theta}_n\rangle\right] \tag{144}$$

$$\leq \sum_{n=1}^{N} \mathbb{E}\left[\langle -\nabla \bar{g}_{n+1}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1}-\boldsymbol{\theta}_n\rangle\right] + \sum_{n=1}^{N} r_{n+1}\mathbb{E}\left[\|\nabla h_n^{v_{n+1}}(\boldsymbol{\theta}_n)\|\right] + \sum_{n=1}^{N} 2Lr_{n+1}^2. \tag{145}$$

Because $h_n^{v_{n+1}}$ is non-negative and has L-Lipschitz continuous gradients $\|\nabla h_n^{v_{n+1}}(\boldsymbol{\theta}_n)\| \leq \sqrt{2Lh_n^{v_{n+1}}(\boldsymbol{\theta}_n)}$ (see Lemma F.2). Then

$$\sum_{n=1}^{N} r_{n+1}\mathbb{E}\left[\|\nabla h_n^{v_{n+1}}(\boldsymbol{\theta}_n)\|\right] \leq \sum_{n=1}^{N} r_{n+1}\mathbb{E}\left[\sqrt{2Lh_n^{v_{n+1}}(\boldsymbol{\theta}_n)}\right] \tag{146}$$

$$\leq \left(\sum_{n=1}^{N} r_{n+1}^2\right)^{1/2}\left(\sum_{n=1}^{N}\mathbb{E}\left[2Lh_n^{v_{n+1}}(\boldsymbol{\theta}_n)\right]\right)^{1/2} \tag{147}$$

$$\leq \sqrt{\frac{2L\Delta_0}{\pi_{\min}}\sum_{n=1}^{N} r_{n+1}^2}. \tag{148}$$

Here the second line used Cauchy-Schwartz and then Jensen's inequality to move the square inside the expectation and the last line used Lemma C.4. From Proposition C.8,

$$\sum_{n=1}^{N} \mathbb{E}\left[\langle -\nabla \bar{g}_{n+1}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle\right] \leq \Delta_0 + L \sum_{n=1}^{N} r_{n+1}^2 \tag{149}$$

so from (145)

$$\sum_{n=1}^{N} b_{n+1} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle\right] \leq \Delta_0 + \sqrt{\frac{2L\Delta_0}{\pi_{\min}} \sum_{n=1}^{N} r_{n+1}^2} + 3L \sum_{n=1}^{N} r_{n+1}^2. \tag{150}$$

From Lemma D.1 with $c_n = r_{n+1}$

$$\sum_{n=1}^{N} r_{n+1} \mathbb{E}\left[\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq \sum_{v\in V} \mathbb{E}\left[\sum_{n=1}^{N} r_{n+1} \|\nabla h_n^v(\boldsymbol{\theta}_n)\|\right] \pi(v) \leq L t_{\odot} \left(\sum_{n=1}^{N} r_{n+1}^2\right)^{1/2} \mathbb{E}\left[\left(\sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right)^{1/2}\right]. \tag{151}$$

Since $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$ and (r_n) is non-increasing, this bound reduces to

$$\sum_{n=1}^{N} r_{n+1} \mathbb{E}\left[\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \leq L t_{\odot} \sum_{n=1}^{N} r_n^2. \tag{152}$$

Combining (150) and (152) with (143) we have

$$\sum_{n=1}^{N} b_n \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle\right] \leq \Delta_0 + \sqrt{\frac{2L\Delta_0}{\pi_{\min}} \sum_{n=1}^{N} r_n^2} + \Big(3 + t_{\odot}\Big) L \sum_{n=1}^{N} r_n^2 \tag{153}$$

so

$$\min_{1\leq n\leq N} \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle\right] \leq \frac{\Delta_0 + \sqrt{\frac{2L\Delta_0}{\pi_{\min}} \sum_{n=1}^{N} r_n^2} + \Big(3 + t_{\odot}\Big) L \sum_{n=1}^{N} r_n^2}{\sum_{n=1}^{N} b_n}. \tag{154}$$

□

D.5. Proofs of Corollaries

In this section, we prove Corollaries A.2 and A.3.

Proof of corollary A.2. Fix N and let k be the integer recognizing the minimum in (25). If $\boldsymbol{\Theta} = \mathbb{R}^p$ then we may choose $\boldsymbol{\theta}^*$ so that $\boldsymbol{\theta}^* - \boldsymbol{\theta}_k = -\frac{\nabla f(\boldsymbol{\theta}_k)}{\|\nabla f(\boldsymbol{\theta}_k)\|}$. Thus,

$$\min_{1\leq n\leq N} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|] \leq \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|] = \mathbb{E}\left[\langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] \tag{155}$$

$$\leq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle\right] = O\left(N^{-1/2}\right). \tag{156}$$

If instead $\boldsymbol{\Theta} \neq \mathbb{R}^p$ but the second condition $\operatorname{dist}(\boldsymbol{\theta}_k, \partial\boldsymbol{\Theta}) \geq c$ holds, we can take $\boldsymbol{\theta} - \boldsymbol{\theta}_k = -c\frac{\nabla f(\boldsymbol{\theta}_k)}{\|\nabla f(\boldsymbol{\theta}_k)\|}$. In doing so we obtain

$$\min_{1\leq n\leq N} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_n)\|] \leq \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_k)\|] = \frac{1}{c}\mathbb{E}\left[\langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] \tag{157}$$

$$\leq \frac{1}{c}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\leq 1} \langle -\nabla f(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle\right] = O\left(N^{-1/2}\right). \tag{158}$$

This shows (30). The proof of (31) is similar. □

Proof of corollary A.3. The convergence rates of Theorem 3.8 are of order $O(N^{-1/2})$ for Algorithm 1. Then we can prove (i) by choosing N large enough so that $N^{-1/2} \leq \varepsilon$.

If we take $r_n = \frac{1}{\sqrt{n}\log n}$ in Algorithm 2 then its corresponding convergence rate in 3.8 is of order $O(n^{-1/2}\log n)$. Then (ii) follows by using the fact that $n \geq \varepsilon^{-2}(3\log\varepsilon^{-1})^2$ implies $n^{-1/2}\log n \leq \varepsilon$ for sufficiently large ε. Indeed since $\frac{\log x}{\sqrt{x}}$ is decreasing for sufficiently large x we have

$$n^{-1/2}\log n \leq \frac{\varepsilon}{3\log\varepsilon^{-1}}\left(2\log\varepsilon^{-1} + 2\log(3\log\varepsilon^{-1})\right) \leq \varepsilon \tag{159}$$

for ε sufficiently small. □

E. Asymptotic Analysis

We use this section to prove Theorem 3.9. Recall that by Proposition C.1, there are constants C_1 and C_2 with $\sup_{n\geq 1}\mathbb{E}[\tau_{n,v}^2|\mathcal{F}_n] \leq C_1$ and $\sup_{n\geq 1}\mathbb{E}[\tau_{n,v}^4|\mathcal{F}_n] \leq C_2$ for each $v \in \mathcal{V}$. Accordingly, we let $\mu_2 = \max_{v\in V}\sup_{n\geq 1}\left\|\mathbb{E}[\tau_{n,v}^2|\mathcal{F}_n]\right\|_\infty$ and $\mu_4 = \max_{v\in\mathcal{V}}\sup_{n\geq 1}\|\mathbb{E}[\tau_{n,v}^4|\mathcal{F}_n]\|_\infty$.

The first Lemma of this section states that the first and second moments of the dynamic regularization parameter ρ_n in Algorithm 1 are uniformly bounded. While ρ_n only appears in Algorithm 1, the random variable $\max_{v\in V}(n - k^v(n))$ is also present in the analysis of Algorithm 2. Therefore, this Lemma is used in the analysis of both algorithms in this section.

Lemma E.1. *Assume 3.1. Then there is a constant $C > 0$ such that*

$$\sup_{n\geq 1}\mathbb{E}[\rho_n] + \sup_{n\geq 1}\mathbb{E}[\rho_n^2] \leq C. \tag{160}$$

Proof. Fix $v \in \mathcal{V}$. For a positive integer j we have

$$\{n - k^v(n) \geq j\} = \{k^v(n) \leq n - j\} \subseteq \{\tau_{n-j,v} \geq j\}. \tag{161}$$

Therefore, we get

$$\mathbb{E}[(n - k^v(n))] = \sum_{j=1}^{\infty}\mathbb{P}(n - k^v(n) \geq j) \tag{162}$$

$$\leq \sum_{j=1}^{\infty}\mathbb{P}(\tau_{n-j,v} \geq j) \tag{163}$$

$$\leq \sum_{j=1}^{\infty}\frac{\mathbb{E}[\tau_{n-j,v}^2]}{j^2} \tag{164}$$

$$\leq \mu_2\sum_{j=1}^{\infty}j^{-2} \tag{165}$$

since $\mathbb{E}[\tau_{n-j,v}^2] \leq \mu_2$. Finally,

$$\mathbb{E}[\rho_n] = \rho + \mathbb{E}\left[\max_{v\in\mathcal{V}}(n - k^v(n))\right] \leq \rho + \sum_{v\in V}\mathbb{E}[n - k^v(n)] \leq \rho + |\mathcal{V}|\mu_2\sum_{j=1}^{\infty}j^{-2}. \tag{166}$$

To bound the second moment we follow the same approach:

$$\mathbb{E}[(n-k^v(n))^2] = \sum_{j=1}^{\infty} \mathbb{P}\left((n-k^v(n))^2 \geq j\right) \tag{167}$$

$$\leq \sum_{j=1}^{\infty} \mathbb{P}(\tau_{n-j,v}^2 \geq j) \tag{168}$$

$$\leq \sum_{j=1}^{\infty} \frac{\mathbb{E}[\tau_{n-j,v}^4]}{j^2} \tag{169}$$

$$\leq \mu_4 \sum_{j=1}^{\infty} j^{-2}. \tag{170}$$

The proof is completed by mimicking the last line of the proof bounding the first moment.

□

E.1. The dynamic proximal regularization case 3.6

Here we prove Theorem 3.9 (i). Our first lemma, Lemma E.2, is similar to D.1 and key to showing that $\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\| \to 0$. The difference is that we must deal with $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{k^v(n)-1}\|^2$ instead of $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{k^v(n)-1}\|$. In order to relate this to the sequence of one step iterate differences $(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2)$ we need to use the Cauchy-Schwartz inequality which introduces a dependence on μ_2 as well as t_{hit}.

Lemma E.2. *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 1. Assume case 3.6. Then*

$$\sum_{n=1}^{\infty} \mathbb{E}[\bar{h}_n(\boldsymbol{\theta}_n)] < \infty. \tag{171}$$

Proof. Since $\bar{h}_n(\boldsymbol{\theta}_n) = \sum_{v\in V} h_n^v(\boldsymbol{\theta}_n)\pi(v)$ and $\mathcal{V}$ is finite, it is sufficient to show $\sum_{n=1}^{\infty} \mathbb{E}[h_n^v(\boldsymbol{\theta}_n)] < \infty$ for each $v \in V$. Before starting recall that by Lemma C.6, and $\rho_n \geq \rho > 0$

$$\sum_{n=1}^{\infty} \frac{\rho_n}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \Delta_0 \quad \text{and} \quad \sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \frac{2}{\rho}\Delta_0 \tag{172}$$

almost surely. This implies

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \rho_n \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right] + \mathbb{E}\left[\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2\right] < \infty. \tag{173}$$

Fix $v \in V$. For each n we have $g_n^v \in \mathcal{S}_L(f^v, \boldsymbol{\theta}_{k^v(n)-1})$. Then using Proposition C.2, the triangle inequality and Cauchy Schwartz

$$|h_n^v(\boldsymbol{\theta}_n)| \leq \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{k^v(n)-1}\|^2 \leq \frac{L}{2}(n-k^v(n)+1)\sum_{i=k^v(n)}^{n} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2. \tag{174}$$

Let $B_n = (n-k^v(n)+1)\sum_{i=k^v(n)}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2$. We claim that $\mathbb{E}\left[\sum_{n=1}^{\infty} B_n\right] < \infty$. As in Lemma D.1 let $p^v(n) = \inf\{j > n : v_j = v\}$ be the next time strictly after time n the sampling algorithm visits v. Exchanging the order of summation we have

$$\mathbb{E}\left[\sum_{n=1}^{\infty}(n-k^v(n)+1)\sum_{i=k^v(n)}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] = \mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{p^v(i)-1}(n-k^v(n)+1)\right] \tag{175}$$

$$= \mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{\infty}(n-k^v(n)+1)\mathbb{1}(p^v(i) > n)\right]. \tag{176}$$

The equality $\{p^v(i) > n\} = \{\tau_{i,v} \geq n - i + 1\}$ holds as both are equal to the event $\{v_j \neq v : i < j \leq n + 1\}$. Moreover, $k^v(n) = k^v(i)$ on $\{p^v(i) > n\}$ since there is no visit to v between times i and n. Therefore,

$$\mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{\infty}(n - k^v(n) + 1)\mathbb{1}(p^v(i) > n)\right] \tag{177}$$

$$= \mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{1}(\tau_{i,v} \geq n - i + 1)\right] \tag{178}$$

$$= \mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i)\right] \tag{179}$$

where the last line used $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i-1}$ and $k^v(i)$ are all measurable with respect to $\mathcal{F}_i$. We have

$$\sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i) \tag{180}$$

$$= \sum_{n=i}^{\infty}(n - i + 1)\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i) + (i - k^v(i))\sum_{n=i}^{\infty}\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i) \tag{181}$$

$$= \mathbb{E}[\tau_{i,v}^2|\mathcal{F}_i] + (i - k^v(i))\mathbb{E}[\tau_{i,v}|\mathcal{F}_i] \tag{182}$$

$$\leq \mu_2 + (i - k^v(i))t_{\text{hit}}. \tag{183}$$

Finally,

$$\mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i)\right] \tag{184}$$

$$\leq \mu_2\mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] + t_{\text{hit}}\mathbb{E}\left[\sum_{i=1}^{\infty}(i - k^v(i))\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] \tag{185}$$

$$\leq \mu_2 + \mathbb{E}\left[\sum_{i=1}^{\infty}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] + t_{\text{hit}}\mathbb{E}\left[\sum_{i=1}^{\infty}\rho_i\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] < \infty. \tag{186}$$

This shows $\mathbb{E}\left[\sum_{n=1}^{\infty} B_n\right] < \infty$. The proof is completed by using Fubini's Theorem and (174) to conclude

$$\sum_{n=1}^{\infty}\mathbb{E}[h_n^v(\boldsymbol{\theta}_n)] = \mathbb{E}\left[\sum_{n=1}^{\infty}h_n^v(\boldsymbol{\theta}_n)\right] \leq \frac{L}{2}\mathbb{E}\left[\sum_{n=1}^{\infty}B_n\right] < \infty. \tag{187}$$

This completes the proof. □

We now prove Theorem 3.9 (i).

Proof of Theorem 3.9 (i). Starting as in the proofs of Theorem 3.8

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] + \mathbb{E}\left[\|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|\right] \tag{188}$$

$$\tag{189}$$

Using the same argument as in the proof of Theorem 3.8 for Case 3.6,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla \bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta} - \boldsymbol{\theta}_n)\right] \leq \mathbb{E}[\rho_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|] \leq \mathbb{E}[\sqrt{2\rho_n\delta_n}] \tag{190}$$

where $\delta_n = \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}) - \bar{g}_n(\boldsymbol{\theta}_n)$. By Cauchy-Schwartz and Lemma E.1

$$\mathbb{E}[\sqrt{2\rho_n\delta_n}] \leq \sqrt{2\mathbb{E}[\rho_n]\mathbb{E}[\delta_n]} \leq C\sqrt{\mathbb{E}[\delta_n]} \tag{191}$$

for some $C > 0$ independent of n. By Jensen's inequality and Lemma F.2

$$\mathbb{E}[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|] \leq \sqrt{\mathbb{E}[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|^2]} \leq \sqrt{2L\mathbb{E}[\bar{h}_n(\boldsymbol{\theta}_n)]} \tag{192}$$

We have

$$\sum_{n=1}^{\infty}\mathbb{E}[\delta_n] = \sum_{n=1}^{\infty}\mathbb{E}[\bar{g}_{n-1}(\boldsymbol{\theta}_{n-1})] - \mathbb{E}[\bar{g}_n(\boldsymbol{\theta}_n)] \leq \Delta_0 < \infty \tag{193}$$

so $\mathbb{E}[\delta_n] \to 0$ as $n \to \infty$. Also, $\sqrt{\mathbb{E}[\bar{h}_n(\boldsymbol{\theta}_n)]} \to 0$ by Lemma E.2. Therefore

$$\limsup_{n\to\infty}\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right] \leq \limsup_{n\to\infty}\left(\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right] + \mathbb{E}\left[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|\right]\right) \tag{194}$$

$$\leq \lim_{n\to\infty}\left(C\sqrt{\mathbb{E}[\delta_n]} + \sqrt{2L\mathbb{E}[\bar{h}_n(\boldsymbol{\theta}_n)]}\right) = 0. \tag{195}$$

We follow a similar approach to show that

$$\mathbb{E}\left[\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right)^2\right] \to 0. \tag{196}$$

Notice that the sub-optimality measure $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)$ is always non-negative, since we can take $\boldsymbol{\theta} = \boldsymbol{\theta}_n$. Then from the inequality

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n) \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n) + \|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\| \tag{197}$$

and Cauchy-Schwartz we get

$$\mathbb{E}\left[\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla f(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right)^2\right] \leq 2\mathbb{E}\left[\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right)^2\right] + 2\mathbb{E}\left[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|^2\right]. \tag{198}$$

Mimicking the proof above and using Lemma E.1

$$\mathbb{E}\left[\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} -\nabla\bar{g}_n(\boldsymbol{\theta}_n, \boldsymbol{\theta}-\boldsymbol{\theta}_n)\right)^2\right] \leq \mathbb{E}[\rho_n^2\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2] \leq 2\mathbb{E}[\rho_n\delta_n] \tag{199}$$

$$\leq 2\sqrt{\mathbb{E}[\rho_n^2]\mathbb{E}[\delta_n^2]}. \tag{200}$$

$$\leq C\sqrt{\mathbb{E}[\delta_n^2]}. \tag{201}$$

Since $\sum_{n=1}^{\infty}\delta_n \leq \Delta_0$ we have $\delta_n \to 0$ almost surely. Therefore, an application of the dominated convergence theorem shows $\mathbb{E}[\delta_n^2] \to 0$. Using Lemmas E.2 and F.2 again to show $\mathbb{E}[\|\nabla\bar{h}_n(\boldsymbol{\theta}_n)\|^2] \to 0$ completes the proof.

□

Remark E.3. The proof of Lemma E.2 demonstrates one of the main difficulties in proving asymptotic convergence for constant proximal regularization. In particular, our techniques require us to show that

$$\mathbb{E}\left[\sum_{i=1}^{\infty}(i - k^v(i))\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] < \infty. \tag{202}$$

The term $i - k^v(i)$ appears as the residual when we swap $n - k^v(i) + 1$ for $n - i + 1$ in order to show

$$\sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{P}(\tau_{i,v} \geq n - i + 1|\mathcal{F}_i) \leq \mu_2 + (i - k^v(i))t_{\text{hit}}$$

To avoid this, an idea is to notice that $\tau_{i,v} \geq n - i + 1$ only if $\tau_{k^v(i),v} \geq n - k^v(i) + 1$ and instead compute

$$\sum_{n=i}^{\infty}(n - k^v(i) + 1)\mathbb{P}(\tau_{k^v(i),v} \geq n - k^v(i) + 1|\mathcal{F}_{k^v(i)}).$$

The problem here is that $k^v(i)$ is *not* a stopping time so, among other things, the σ-algebra $\mathcal{F}_{k^v(i)}$ may not be well defined. Intuitively, $i - k^v(i)$ represents a gap in knowledge since we must wait until time i to know the last time v was visited.

The use of dynamic proximal regularization bakes (202) into the algorithm. Lemmas E.1 and C.6 suggests that it may be true with constant proximal regularization: if $(i - k^v(n))$ and $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2$ were independent then

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{\infty}(i - k^v(i))\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2\right] &= \sum_{i=1}^{\infty}\mathbb{E}[i - k^v(i)]\mathbb{E}[\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2] \\ &\leq C\sum_{i=1}^{\infty}\mathbb{E}[\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2] < \infty.\end{aligned}$$

However, $k^v(i)$, $\boldsymbol{\theta}_i$, and $\boldsymbol{\theta}_{i-1}$ are all determined by the behavior of the sampling process so we do not have this independence.

We will see in the next subsection that diminishing radius overcomes this issue by bounding the difference $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2$ by a deterministic quantity.

E.2. The diminishing radius case 3.7

Here we prove Theorem 3.9 (ii). Lemma E.4 is an analogue of Lemma E.2 for the diminishing radius case. The remaining argument is similar to that used in (Lyu & Li, 2023) and (Lyu, 2023) to analyze block majorization-minimization and SMM with diminishing radius respectively.

Lemma E.4. *Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be an output of Algorithm 2. Assume Case 3.7. Then almost surely*

$$\sum_{n=1}^{\infty}\bar{h}_n(\boldsymbol{\theta}_n) < \infty. \tag{203}$$

Proof. The strategy here is nearly the same as in Lemma E.2 except that we use $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$.

Again, it is sufficient to show $\sum_{n=1}^{\infty} h_n^v(\boldsymbol{\theta}_n) < \infty$ almost surely for each $v \in \mathcal{V}$. Fixing v we have $g_n^v \in \mathcal{S}_L(f^v, \boldsymbol{\theta}_{k^v(n)-1})$. Then Proposition C.2, the triangle inequality, and Cauchy-Schwartz give us

$$|h_n^v(\boldsymbol{\theta}_n)| \leq \frac{L}{2}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{k^v(n)-1}\|^2 \leq \frac{L}{2}(n - k^v(n) + 1)\sum_{i=k^v(n)}^{n}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2 \tag{204}$$

$$\leq \frac{L}{2}(n - k^v(n) + 1)\sum_{i=1}^{n} r_i^2. \tag{205}$$

Let $B_n = (n - k^v(n) + 1)\sum_{i=k^v(n)}^{n} r_i^2$. We mimic the proof of Lemma E.2 with r_i^2 in place of $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|^2$ to conclude

$$\mathbb{E}\left[\sum_{n=1}^{\infty} B_n\right] \leq \mu_2\sum_{i=1}^{\infty} r_i^2 + t_{\text{hit}}\mathbb{E}\left[\sum_{i=1}^{\infty}\rho_i r_i^2\right]. \tag{206}$$

The first term on the right hand side is finite by Assumption 3.4. Moreover, by Lemma E.1 and Fubini's Theorem

$$\mathbb{E}\left[\sum_{i=1}^{\infty}\rho_i r_i^2\right] = \sum_{i=1}^{\infty}\mathbb{E}[\rho_i]r_i^2 \leq C\sum_{i=1}^{\infty} r_i^2 < \infty. \tag{207}$$

Hence,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} h_n^v(\boldsymbol{\theta}_n)\right] \leq \frac{L}{2}\mathbb{E}\left[\sum_{n=1}^{\infty} B_n\right] < \infty. \tag{208}$$

It then follows that $\sum_{n=1}^{\infty} h_n^v(\boldsymbol{\theta}_n)$ is finite almost surely.

□

Proposition E.5. *Assume Case 3.7. Suppose there exists a sequence $(n_k)_{k\geq 1}$ such that almost surely either*

$$\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}\| = \infty \quad \text{or} \quad \liminf_{k\to\infty} \left| \left\langle \nabla \bar{g}_{n_k+1}(\boldsymbol{\theta}_{n_k}), \frac{\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}}{\|\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}\|} \right\rangle \right| = 0. \tag{209}$$

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

Proof. By Proposition C.8,

$$\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}\| \left| \left\langle \nabla \bar{g}_{n_k+1}(\boldsymbol{\theta}_{n_k}), \frac{\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}}{\|\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}\|} \right\rangle \right| < \infty \quad \text{a.s.} \tag{210}$$

Therefore, the former condition implies the latter almost surely. So, it suffices to show the the latter condition implies the assertion. Assume the latter condition in (209) and let $(m_k)_{k\geq 1}$ be a subsequence of $(n_k)_{k\geq 1}$, satisfying

$$\lim_{k\to\infty} \left| \left\langle \nabla \bar{g}_{m_k+1}(\boldsymbol{\theta}_{m_k}), \frac{\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}}{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|} \right\rangle \right| = 0. \tag{211}$$

Since $\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\| \leq r_{m_k}$, it follows that

$$\lim_{k\to\infty} \frac{\|\boldsymbol{\theta}_{m_{k+1}} - \boldsymbol{\theta}_{m_k}\|}{b_{m_k+1}} \left| \left\langle \nabla \bar{g}_{m_k+1}(\boldsymbol{\theta}_{m_k}), \frac{\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}}{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|} \right\rangle \right| = 0. \tag{212}$$

where $b_n = \min\{1, r_n\}$. If $\boldsymbol{\theta}_\infty$ is not a stationary point of f over $\boldsymbol{\Theta}$, then we may find $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_\infty\| \leq 1$ and $\varepsilon > 0$ so that

$$\langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_\infty \rangle \leq -\varepsilon < 0. \tag{213}$$

On the other hand by the triangle inequality and Cauchy-Schwartz

$$\begin{aligned}
&|\langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k} \rangle - \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_\infty \rangle| && (214)\\
&= |\langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k}) - \nabla f(\boldsymbol{\theta}_{m_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k} \rangle + \langle \nabla f(\boldsymbol{\theta}_{m_k}) - \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k} \rangle + \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{m_k} \rangle| && (215)\\
&\leq \|\nabla \bar{h}_{m_k}(\boldsymbol{\theta}_{m_k})\| \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k}\| + \|\nabla f(\boldsymbol{\theta}_{m_k}) - \nabla f(\boldsymbol{\theta}_\infty)\| \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_\infty\| + \|\nabla f(\boldsymbol{\theta}_\infty)\| \|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{m_k}\| && (216)
\end{aligned}$$

Since $(\boldsymbol{\theta}_{m_k})_{k\geq 1}$ converges, $\sup_k \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k}\| \leq M$ for some $M < \infty$. Furthermore,

$$\sum_{n=1}^{\infty} \|\nabla \bar{h}_n(\boldsymbol{\theta}_n)\|^2 \leq 2L \sum_{n=1}^{\infty} \bar{h}_n(\boldsymbol{\theta}_n) < \infty \tag{217}$$

by Lemmas E.4 and F.2 so $\|\nabla \bar{h}_{m_k}(\boldsymbol{\theta}_{m_k})\| \to 0$ almost surely as $k \to \infty$. This, together with continuity of ∇f and $\boldsymbol{\theta}_{m_k} \to \boldsymbol{\theta}_\infty$, shows that right hand side above tends to zero as $k \to \infty$. Then we can choose K sufficiently large so that

$$\langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k} \rangle \leq -\frac{\varepsilon}{2} \tag{218}$$

for $k \geq K$. Recall that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| \leq r_n$ and $r_n = o(1)$. Applying Lemma C.9 we get

$$\frac{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|}{b_n} \left\langle \nabla \bar{g}_n(\boldsymbol{\theta}_{n-1}), \frac{\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}}{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|} \right\rangle \leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\|\leq 1} \langle \nabla \bar{g}_{n-1}(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\rangle + \|\nabla h_{n-1}^{v_n}(\boldsymbol{\theta}_{n-1})\| + L r_n. \quad (219)$$

It then follows that for sufficiently large k

$$\frac{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|}{b_{m_k+1}} \left\langle \nabla \bar{g}_{m_k+1}(\boldsymbol{\theta}_{m_k}), \frac{\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}}{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|} \right\rangle \quad (220)$$

$$\leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_{m_k}\|\leq 1} \langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{m_k}\rangle + \|\nabla h_{m_k}^{v_{m_k+1}}(\boldsymbol{\theta}_{m_k})\| + L r_{m_k} \quad (221)$$

$$\leq \langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k}), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{m_k}\rangle + \|\nabla h_{m_k}^{v_{m_k+1}}(\boldsymbol{\theta}_{m_k})\| + r_{m_k} \quad (222)$$

$$\leq -\frac{\varepsilon}{2} + \|\nabla h_{m_k}^{v_{m_k+1}}(\boldsymbol{\theta}_{m_k}\| + r_{m_k} \quad (223)$$

Recall Lemma C.4 which shows

$$\sum_{n=1}^{\infty} h_n^{v_{n+1}}(\boldsymbol{\theta}_n) < \infty \quad (224)$$

almost surely. Therefore, since $h_n^{v_n}$ is non-negative, $\sqrt{h_n^{v_{n+1}}(\boldsymbol{\theta}_n)} \to 0$ almost surely as $n \to \infty$. Moreover, by Lemma F.2, $\|\nabla h_n^{v_{n+1}}(\boldsymbol{\theta}_n)\| \leq \sqrt{2L h_n^{v_{n+1}}(\boldsymbol{\theta}_n)}$. So letting $k \to \infty$ shows

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|}{b_{m_k+1}} \left\langle \nabla \bar{g}_{m_k}(\boldsymbol{\theta}_{m_k+1}), \frac{\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}}{\|\boldsymbol{\theta}_{m_k+1} - \boldsymbol{\theta}_{m_k}\|} \right\rangle \leq -\frac{\varepsilon}{2} \quad (225)$$

contradicting (212). □

Recall that under Algorithm 2, the one step parameter difference $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|$ is at most r_n. For each $n \geq 1$ we say that $\boldsymbol{\theta}_n$ is a *long point* if $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| < r_n$ and a *short point* if $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = r_n$. The next proposition shows that if $\boldsymbol{\theta}_n$ is a long point, then $\boldsymbol{\theta}_n$ is obtained by directly minimizing $\bar{g}_n$ over the full parameter space $\boldsymbol{\Theta}$. It is here the we crucially use the convexity of $\bar{g}_n$ from Definition 2.1.

Proposition E.6. *For $n \geq 1$, suppose that $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\cap B_{r_n}(\boldsymbol{\theta}_{n-1})} \bar{g}_n(\boldsymbol{\theta})$ and that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| < r_n$. Then $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \bar{g}_n(\boldsymbol{\theta})$.*

Proof. By Definition 2.1, $\bar{g}_n$ is convex. Thus, it suffices to verify the first order stationarity condition

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle \geq 0 \quad (226)$$

to conclude $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \bar{g}_n(\boldsymbol{\theta})$. To this end, assume the conclusion is false. Then there is $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}$ with $\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_n\rangle < 0$. Moreover, as $\boldsymbol{\theta}_n$ is obtained by minimizing $\bar{g}_n$ over $\boldsymbol{\Theta} \cap B_{r_n}(\boldsymbol{\theta}_{n-1})$ we must have $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_{n-1}\| > r_n$. As we are assuming $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| < r_n$, there is $\alpha \in (0,1)$ so that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = \alpha r_n$. Notice that

$$\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_n\| \geq \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_{n-1}\| - \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| > (1-\alpha) r_n. \quad (227)$$

Hence if we set $a = \frac{(1-\alpha)r_n}{\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_n\|}$ then $a \in (0,1)$. So, the convexity of $\boldsymbol{\Theta}$ implies that $\tilde{\boldsymbol{\theta}} := a(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_n) + \boldsymbol{\theta}_n \in \boldsymbol{\Theta}$. Furthermore,

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{n-1}\| \leq a\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_n\| + \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = (1-\alpha) r_n + \alpha r_n = r_n \quad (228)$$

and

$$\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\rangle = a\langle \nabla \bar{g}_n(\boldsymbol{\theta}_n), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_n\rangle < 0. \quad (229)$$

This contradicts $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\cap B_{r_n}(\boldsymbol{\theta}_{n-1})} \bar{g}_n(\boldsymbol{\theta})$ and completes the proof. □

Proposition E.7. *Assume the Case 3.7. If $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ is a sequence consisting of long points such that $\boldsymbol{\theta}_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$ exist almost surely, then $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$.*

Proof. By the assumption that $\boldsymbol{\theta}_{n_k}$ is a long point and Proposition E.6 we have $\boldsymbol{\theta}_{n_k} \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \bar{g}_{n_k}(\boldsymbol{\theta})$. Therefore, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\langle \nabla \bar{g}_{n_k}(\boldsymbol{\theta}_{n_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n_k} \rangle \geq 0. \tag{230}$$

We then notice that

$$\langle \nabla f(\boldsymbol{\theta}_{n_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{m_k} \rangle = \langle \nabla \bar{g}_{n_k}(\boldsymbol{\theta}_{n_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n_k} \rangle - \langle \nabla \bar{h}_{n_k}(\boldsymbol{\theta}_{n_k}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n_k} \rangle. \tag{231}$$

By Lemmas F.2 and E.4, $\|\nabla \bar{h}_{n_k}(\boldsymbol{\theta}_{n_k})\|^2 \leq 2L\bar{h}_{n_k}(\boldsymbol{\theta}_{n_k}) \to 0$ almost surely as $k \to \infty$. Therefore, by taking limits we get

$$\langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0. \tag{232}$$

Since this holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_\infty\|\leq 1} \langle -\nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \leq 0 \tag{233}$$

which means that $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$. □

Proposition E.8. *Suppose there exists a sub-sequence $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ such that $\lim_{k\to\infty} \boldsymbol{\theta}_{n_k} = \boldsymbol{\theta}_\infty$ exists almost surely and that $\boldsymbol{\theta}_\infty$ is not a stationary point of f over $\boldsymbol{\Theta}$. Then there is $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty)$ has the following properties:*

(a) $B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any stationary points of f over $\boldsymbol{\Theta}$.

(b) There are infinitely many n for which $\boldsymbol{\theta}_n$ is outside of $B_\varepsilon(\boldsymbol{\theta}_\infty)$.

Proof. We first show that there exists $\varepsilon > 0$ so that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any long points. Suppose for contradiction that for each $\varepsilon > 0$, there is a long point in $B_\varepsilon(\boldsymbol{\theta}_\infty)$. Then one may construct a sequence of long points converging to $\boldsymbol{\theta}_\infty$. But then by Proposition E.7, $\boldsymbol{\theta}_\infty$ is a stationary point for f over $\boldsymbol{\Theta}$, a contradiction.

Next we show that there exists ε so that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies (a). In fact, suppose not. Then we can find a sequence of stationary points $(\boldsymbol{\theta}_{\infty,k})_{k\geq 1}$ converging to $\boldsymbol{\theta}_\infty$. But then by continuity of ∇f,

$$\langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle = \lim_{k\to\infty} \langle \nabla f(\boldsymbol{\theta}_{k,\infty}), \boldsymbol{\theta} - \boldsymbol{\theta}_{k,\infty} \rangle \geq 0 \tag{234}$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then $\boldsymbol{\theta}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$, contradicting our assumptions.

Now let $\varepsilon > 0$ be such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any long points and satisfies (a). We will show that $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies (b) and thus $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies both (a) and (b) as desired. Aiming for a contradiction, suppose there are only finitely many n for which $\boldsymbol{\theta}_n$ is outside $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$. Then there exists N so that $\boldsymbol{\theta}_n \in B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ for all $n \geq N$. Then $\boldsymbol{\theta}_n$ is a short point for each $n \geq N$ so $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = r_n$ for all $n \geq N$. This, in turn, implies that $\sum_{n=1}^\infty \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = \infty$. By Proposition E.5, there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ such that $\boldsymbol{\theta}'_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$ exists and is stationary for f. But since $\boldsymbol{\theta}'_\infty \in B_\varepsilon(\boldsymbol{\theta}_\infty)$, this contradicts (a). The proof is complete. □

We now prove Theorem 3.9 (ii).

Proof of Theorem 3.9 (ii). Suppose for contradiction that there exists a non-stationary limit point $\boldsymbol{\theta}_\infty$ of $(\boldsymbol{\theta}_n)_{n\geq 0}$. By Proposition E.8, there is $\varepsilon > 0$ so that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies the conditions (a) and (b). Choose N large enough so that $r_n \leq \frac{\varepsilon}{4}$ for $n \geq N$. 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 intersection. Fix a crossing $I = [\ell, \ell')$. It follows by the triangle inequality,

$$\sum_{n=\ell}^{\ell'-1} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \geq \|\boldsymbol{\theta}_{\ell'} - \boldsymbol{\theta}_\ell\| \geq \varepsilon/3. \tag{235}$$

Note that since $\boldsymbol{\theta}_\infty$ is a limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$, we have $\boldsymbol{\theta}_n \in B_{\varepsilon/3}(\boldsymbol{\theta}_\infty)$ infinitely often. In addition, by condition (b) of Proposition E.8, $\boldsymbol{\theta}_n$ also exits $B_\varepsilon(\boldsymbol{\theta}_\infty)$ infinitely often. Therefore, there must be infinitely many crossings. Let n_k be the k-th smallest integer that appears in some crossing, noting importantly that $\boldsymbol{\theta}_{n_k} \in B_{2\varepsilon/3}$ for $k \geq 1$. Then $n_k \to \infty$ as $k \to \infty$ and by (235),

$$\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{n_k+1} - \boldsymbol{\theta}_{n_k}\| \geq (\#\text{ of crossings})\frac{\varepsilon}{3} = \infty. \tag{236}$$

Then by Proposition E.5, there is a further subsequence $(\boldsymbol{\theta}_{m_k})_{k\geq 1}$ of $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ so that $\boldsymbol{\theta}'_\infty = \lim_{k\to\infty} \boldsymbol{\theta}_{m_k}$ exists and is stationary. However, since $\boldsymbol{\theta}_{n_k} \in B_{2\varepsilon/3}(\boldsymbol{\theta}_\infty)$ the stationary point $\boldsymbol{\theta}'_\infty$ is in $B_\varepsilon(\boldsymbol{\theta}_\infty)$. This contradicts property (a) of Proposition E.8 which shows the assertion. □

E.3. Details for numerical experiments

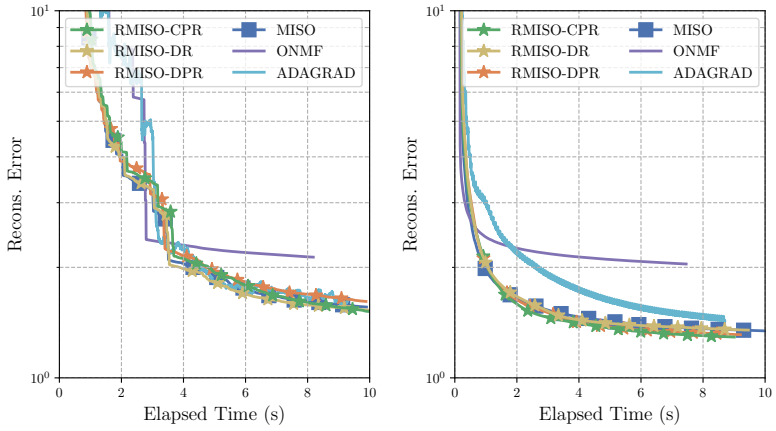
E.3.1. Distributed Nonnegative Matrix Factorization

The MNIST samples X_v at each node were formed by concatenating a collection of images $\{X_i\}_{i=1}^k \subset \mathbb{R}_+^{28\times 28}$ along the horizontal axis so that $X_v \in \mathbb{R}_+^{28\times 28k}$. We selected 5000 images from the full dataset at random and divided them into groups based on class label. New nodes were formed by adding batches of 100 images from each group until fewer than 100 images remained. Then a final node was added for the remaining images.

We include here a list of hyperparamters used for the NMF experiments.

For AdaGrad we used constant step size parameter $\eta = 0.5$. For both RMISO-DPR and RMISO-CPR we set $\rho = 2500$ for the random walk and $\rho = 50$ for cyclic sampling. For the diminishing radius version RMISO-DR we set $r_n = \frac{1}{\sqrt{n}\log(n+1)}$.

Figure 4 displays the results of these experiments vs compute time.



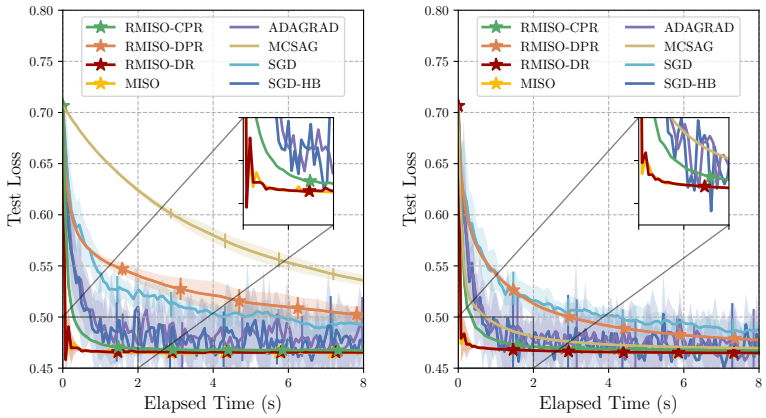
(a) Random Walk (b) Cyclic

Figure 4. Plot of reconstruction error against compute time for NMF using two sampling algorithms. Results show the performance of algorithms RMISO, MISO (Algorithm 1 with $\rho_n = 0$), ONMF, and AdaGrad in factorizing a collection of MNIST (Deng, 2012) data matrices.

E.3.2. Logistic Regression with nonconvex regularization

The hyperparameters for the logistic regression experiments were chosen as follows. For MCSAG and RMISO/MISO we took $L = 2/5$. The random walk on the complete graph has $t_{\text{hit}} = O(|\mathcal{V}|)$ while $t_{\text{hit}} = O(|\mathcal{V}|^2)$ for the lonely graph but $t_\odot = O(|\mathcal{V}|)$ for both. Accordingly for MCSAG we set the hitting time parameter in the step size $t_{\text{hit}} = 50$ for the complete graph and $t_{\text{hit}} = 2500$ for the lonely graph. For RMISO we set $\rho = 50$ for both the constant proximal regularization version and the dynamic proximal regularization version. We ran SGD with a decaying step size of the form $\alpha_n = \frac{\alpha}{n^\gamma}$ where $\alpha = 0.1$ and $\gamma = 0.5$. For SGD-HB and AdaGrad we used step sizes $\alpha = 0.05$ and SGD-HB momentum parameter $\beta = 0.9$.

Figure 5 shows the results of our experiments plotted vs compute time.



(a) Lonely graph (b) Complete graph

Figure 5. Plot of objective loss and standard deviation vs compute time for `a9a` for two graph topologies and various optimization algorithms- RMISO, MISO (Algorithm 1 with $\rho_n = 0$), AdaGrad, MCSAG, SGD, Adam, and SGD-HB

F. Auxiliary Lemmas

Lemma F.1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a continuously differentiable function with L-Lipschitz continuous gradient. Then for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$,*

$$|f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle| \leq \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{237}$$

Proof. This is a classical lemma. See (Nesterov, 2003) Lemma 1.2.3. □

Lemma F.2. *Let $f : \mathbb{R}^p \to [0, \infty)$ be a continuously differentiable function with L-Lipschitz continuous gradient. Then for all $\boldsymbol{\theta} \in \mathbb{R}^p$, it holds $\|\nabla f(\boldsymbol{\theta})\|^2 \leq 2Lf(\boldsymbol{\theta})$.*

Proof. Fix $\boldsymbol{\theta} \in \mathbb{R}^p$. By Lemma F.1 we have

$$\inf_{\boldsymbol{\theta}' \in \mathbb{R}^p} f(\boldsymbol{\theta}') \leq \inf_{\boldsymbol{\theta}' \in \mathbb{R}^p} \left\{ f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \right\}. \tag{238}$$

It is easy to compute that

$$\inf_{\boldsymbol{\theta}' \in \mathbb{R}^p} \left\{ f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \right\} = f(\boldsymbol{\theta}) - \frac{1}{2L}\|\nabla f(\boldsymbol{\theta})\|_2^2. \tag{239}$$

Therefore

$$\|\nabla f(\boldsymbol{\theta})\|_2^2 \leq 2L(f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \mathbb{R}^p} f(\boldsymbol{\theta}')) \leq 2Lf(\boldsymbol{\theta}) \tag{240}$$

since $\inf_{\boldsymbol{\theta}' \in \mathbb{R}^p} f(\boldsymbol{\theta}') \geq 0$. □

G. Examples of Surrogate Functions

Example G.1 (Proximal surrogates for L-smooth functions)**.** Suppose f is continuously differentiable with L-Lipschitz continuous gradients. Then f is L-weakly convex, meaning $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta}\|^2$ is convex (see (Lyu, 2023) Lemma C.2). For each $\gamma \geq L$, the following function belongs to $\mathcal{S}_{L+\gamma}(f, \boldsymbol{\theta}^*)$:

$$g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \tag{241}$$

Indeed, $g \geq f$, $g(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^*)$, $\nabla h(\boldsymbol{\theta}^*) = 0$, and ∇h is $(L+\rho)$ Lipschitz. Minimizing the above function over $\boldsymbol{\Theta}$ is equivalent to applying a proximal mapping of f where the resulting estimate is denoted $\text{prox}_{f/\rho}(\boldsymbol{\theta}^*)$ (see (Parikh & Boyd, 2014; Davis & Drusvyatskiy, 2019)).

Example G.2 (Prox-linear surrogates)**.** If f is L-smooth, then the following quadratic function g belongs to $\mathcal{S}_{2L}(f, \boldsymbol{\theta}^*)$:

$$g : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}^*) + \langle \nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2. \tag{242}$$

Indeed, $g(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^*)$, $\nabla g(\boldsymbol{\theta}^*) = \nabla f(\boldsymbol{\theta}^*)$. Moreover,

$$\|\nabla h(\boldsymbol{\theta}) - \nabla h(\boldsymbol{\theta}')\| = \|\nabla f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}) + L(\boldsymbol{\theta} - \boldsymbol{\theta}')\| \leq 2L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \tag{243}$$

since f is L-smooth.

Example G.3 (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 $\mathcal{S}_{2L}(f, \boldsymbol{\theta}^*)$:

$$g : \boldsymbol{\theta} \mapsto f_1(\boldsymbol{\theta}^*) + \langle \nabla f_1(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 + f_2(\boldsymbol{\theta}). \tag{244}$$

Minimizing g over $\boldsymbol{\Theta}$ amounts to performing a proximal gradient step (Beck & Teboulle, 2009; Nesterov, 2013).

Example G.4 (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 $\mathcal{S}_{2L}(f, \boldsymbol{\theta}^*)$:

$$g : \boldsymbol{\theta} \mapsto f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{\theta}^*) + \langle \nabla f_2(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle. \tag{245}$$

Such surrogates are important in DC programming (Horst & Thoai, 1999).

Example G.5 (Variational Surrogates)**.** Let $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ be a two-block multi-convex function and let $\boldsymbol{\Theta}_1 \subseteq \mathbb{R}^p$ and $\boldsymbol{\Theta}_2 \subseteq \mathbb{R}^q$ be two convex sets. Define a function $f_* : \inf_{H \in \boldsymbol{\Theta}_2} f(\boldsymbol{\theta}, H)$. Then for each $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, the following function

$$g : \boldsymbol{\theta} \to f(\boldsymbol{\theta}, H^*), \quad H^* \in \operatorname*{arg\,min}_{H \in \boldsymbol{\Theta}_2} f(\boldsymbol{\theta}^*, H) \tag{246}$$

is convex over $\boldsymbol{\Theta}_1$ and satisfies $g \geq f$ and $g(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^*)$. Further, assume that

(i) $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}, H)$ is differentiable for all $H \in \boldsymbol{\Theta}_2$ and $\boldsymbol{\theta} \to \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ is L'-Lipschitz for all $H \in \boldsymbol{\Theta}_2$;

(ii) $H \mapsto \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, H)$ is L-Lipschitz for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$;

Then g belongs to $\mathcal{S}_L(f_*, \boldsymbol{\theta}^*)$ for some $L'' > 0$. When f is jointly convex, then f_* is also convex and we can choose $L'' = L$.

H. Matrix factorization algorithms

Here we formally state the non-negative matrix factorization algorithms derived in Section 4.1.1. They may be compared to the celebrated online nonnegative matrix factorization algorithm in (Mairal et al., 2010) which is a special case of SMM.

With surrogates $g_n(W)$ as defined in Section 4.1.1, one can show that minimizing the averaged surrogate $\bar{g}_n(W) = \frac{1}{|\mathcal{V}|}\sum_{v \in V} g_n^v(W)$ is equivalent to minimizing

$$\text{tr}(W A_n W^T) - 2\text{tr}(W B_n), \tag{247}$$

with A_n and B_n, defined recursively as

$$A_n := A_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n}(H_n^{v_n})^T - H_{n-1}^{v_n}(H_{n-1}^{v_n})^T\right] \tag{248}$$

$$B_n := B_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n} X_v^T - H_{n-1}^{v_n} X_v^T\right]. \tag{249}$$

With this, we state the full algorithms below.

Algorithm 3 Distributed Matrix Factorization with Proximal Regularization

Input: $(X^v)_{v\in\mathcal{V}}$ (Data matrices in $\mathbb{R}^{p\times d}$); $W_0 \in \boldsymbol{\Theta}_W$ (initial dictionary); N (number of iterations); $\rho > 0$ (regularization parameter)
Option: $Regularization \in \{Dynamic, Constant\}$
Compute initial codes $H_0^v \in \arg\min_{H\in\boldsymbol{\Theta}_H^{\mathcal{V}}} \frac{1}{2}\|X^v - W_0 H\|_F^2 + \alpha\|H\|_1$ for each $v \in \mathcal{V}$
for $n = 1$ **to** N **do**
  sample an index v_n
  update $H_n^v \in \arg\min_{H\in\boldsymbol{\Theta}_H^{\mathcal{V}}} \frac{1}{2}\|X_v - W_{n-1}H\|_F^2 + \alpha\|H\|_1$; $H_n^v = H_{n-1}^v$ for $v \neq v_n$
  $A_n \leftarrow A_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n}(H_n^{v_n})^T - H_{n-1}^{v_n}(H_{n-1}^{v_n})^T\right]$
  $B_n \leftarrow B_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n}(X^{v_n})^T - H_{n-1}^{v_n}(X^{v_n})^T\right]$
  if $Regularization = Dynamic$ **then**
    $\rho_n \leftarrow \rho + \max_{v\in\mathcal{V}}(n - k^v(n))$
  else
    $\rho_n \leftarrow \rho$
  end if
  update dictionary W_n:

$$W_n \in \underset{W\in\boldsymbol{\Theta}_W}{\arg\min}\left[\text{tr}(WA_nW^T) - 2\text{tr}(WB_n) + \frac{\rho_n}{2}\|W - W_{n-1}\|_F^2\right] \tag{250}$$

end for
output: $\boldsymbol{\theta}_N$

Algorithm 4 Distributed Matrix Factorization with Diminishing Radius

Input: $(X^v)_{v\in\mathcal{V}}$ (Data matrices in $\mathbb{R}^{p\times d}$); $W_0 \in \boldsymbol{\Theta}_W$ (initial dictionary); N (number of iterations); $(r_n)_{n\geq 1}$ (diminishing radius search constraints)
Compute initial codes $H_0^v \in \arg\min_{H\in\boldsymbol{\Theta}_H^{\mathcal{V}}} \frac{1}{2}\|X^v - W_0 H\|_F^2 + \alpha\|H\|_1$ for each $v \in \mathcal{V}$
for $n = 1$ **to** N **do**
  sample an index v_n
  update $H_n^v \in \arg\min_{H\in\boldsymbol{\Theta}_H^{\mathcal{V}}} \frac{1}{2}\|X_v - W_{n-1}H\|_F^2 + \alpha\|H\|_1$; $H_n^v = H_{n-1}^v$ for $v \neq v_n$
  $A_n \leftarrow A_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n}(H_n^{v_n})^T - H_{n-1}^{v_n}(H_{n-1}^{v_n})^T\right]$
  $B_n \leftarrow B_{n-1} + \frac{1}{|\mathcal{V}|}\left[H_n^{v_n}(X^{v_n})^T - H_{n-1}^{v_n}(X^{v_n})^T\right]$
  update dictionary W_n:

$$W_n \in \underset{W\in\boldsymbol{\Theta}_W\cap B_{r_n}(W_{n-1})}{\arg\min}\left[\text{tr}(WA_nW^T) - 2\text{tr}(WB_n)\right] \tag{251}$$

end for
output: $\boldsymbol{\theta}_N$