Statistical Inference in Latent Convex Problems on Stream Data

Rohan Chauhan [◇] Emmanouil V. Vlatakis-Gkaragkounis[◇] Michael I. Jordan [†,◇]

Department of Electrical Engineering and Computer Sciences[◇]
Department of Statistics[†]
University of California, Berkeley

January 16, 2024

Abstract

Stochastic gradient methods are increasingly employed in statistical inference tasks, such as parameter and interval estimation. Yet, much of the current theoretical framework mainly revolves around scenarios with i.i.d. observations or strongly convex objectives, bypassing more complex models. To address this gap, our paper delves into the challenges posed by correlated stream data and the inherent intricacies of the non-convex landscapes in neural network applications. In this context, we present SHADE (Stochastic Hidden Averaging Data Estimator), a novel mini-batch gradient based estimator. We further substantiate its asymptotic normality through a tailored central limit theorem designed explicitly for its average scheme. From a technical perspective, our analysis integrates recent advancements in composite (hidden) convex optimization, stochastic processes, and dynamical systems.

1 Introduction

Nowadays we witness a surge in large-scale data-driven techniques, encompassing areas such as machine learning, statistical methodologies, and operational research [Council et al., 2013]. However, as these analytical tools gain prominence in areas with significant implications for individuals, like drug discovery and vaccine approval, there is an increasing realization that machine learning models, when used merely as "black boxes" for predictive purposes, could lead to hazardous misconceptions. For instance, consider FDA's surveillance models about adverse reactions for new CoViD-19 vaccines or drugs [Arora et al., 2021, McMurry et al., 2021]. In a simplified realizable scenario, such a model might evaluate two parameters ω_1 (reaction severity) and ω_2 (elapsed time since administration) and there exists an unknown parameter θ^* such that $y_{\theta^*} \sim L(\theta^*; (\omega_1, \omega_2))$ might describe the likelihood of hospitalizations for a given adverse reaction. Notably, while multiple θ could optimize prediction accuracy, like $\mathbb{E}_{(\omega_1,\omega_2)}[(y_\theta - y_{\theta^*})^2]$, discerning the magnitude or signs of θ^* becomes indispensable for gleaning correct causal insights about reaction severity and timing.

To tackle this inference task, a foundational methodology in statistics for estimating the true parameters $\theta^* \in \mathbb{R}^d$ of a d-dimensional model is through minimization of an objective function, typically expressed as the expected loss over a distribution spanning dataset sample space Ω, also known as the population risk:

$$\theta^* = \arg\min_{\theta \in \Theta} \left\{ \ell(\theta) = \mathbb{E}[L(\theta; \omega)] = \int_\omega L(\theta; \omega) \mathrm{d}\Pi(\omega) \right\}$$

Following the standard approach, $L(\theta; \omega)$ quantifies the empirical loss when estimating the parameter θ given the observed data ω sampled by a distribution Π.

However, direct computation of the expected loss or its gradient is often computationally infeasible, rendering traditional optimization methods unsuitable. A common remedy is to replace θ^* with its empirical counterpart θ_n^*:

$$\theta_n^* = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(\theta;\omega_s), \tag{ERM}$$

an approach frequently termed empirical risk minimization (ERM) or M-estimation [Hayashi, 2011].

To ensure the statistical soundness of this relaxation, two core criteria must be met (See Wasserman [2004]):

i) The *consistency* of the estimator θ_n^*, which ensures it converges in probability to the true parameter value as the sample size tends to infinity.

ii) The existence of a *central limit theorem* (CLT), which identifies an asymptotic limiting distribution, is pivotal for crafting confidence intervals for θ_n^*.

Building on these prerequisites, [Van der Vaart, 2000] demonstrates that under mild regularity conditions, ERM solutions exhibit asymptotic normality, meaning $\sqrt{n}(\theta_n^* - \theta^*)$ weakly converge to a normal distribution.

Attracted by the centrality of our question in ML, prior work has showcased that the framework of gradient-based estimators provides both promising solutions and inherent challenges. Even for the basic case of smooth strongly convex loss functions, the standard implementation for stochastic gradient descent (SGD) – traced back to the seminal work of Robbins and Monro [1951] – only partially meets the criteria set out earlier. Specifically, the last iteration of SGD, when viewed as a statistical estimator, demonstrates *a*) asymptotic normality[1] but *b*) with sub-optimal rate of consistency[2].

To ensure an optimally $\sqrt{n}$-consistent estimator, Polyak [1990b] and Ruppert [1988] independently independently proposed as statistical estimator the averaged SGD (ASGD) iterate:

$$\hat{\theta}_n^{\text{ASGD}} = n^{-1}\sum_{k=1}^{n}\theta_k, \text{ where } \theta_k = \theta_{k-1} - \eta_k \nabla L(\theta_{k-1};\omega_k) \qquad \text{(Polyak-Ruppert averaging scheme)}$$

for a diminishing learning rate $\eta_k \propto 1/k^\rho$, $\rho \in (0.5, 1)$. Indeed, Polyak and Juditsky [1992b] confirmed its $\sqrt{n}$-consistency and asymptotic normality, forming a bedrock for further advances in the domain.

However, real-world applications introduce their own complexities: Modern systems frequently process online streaming data [Agarwal and Duchi, 2012], handle large datasets tainted by numerical errors [Schroeder and Gibson, 2009] or data intentionally obfuscated for privacy considerations [Song et al., 2013]. Moreover, practicalities, like dropout in neural training or storage limitations, lead to periodic data exclusion [Srebro and Tewari, 2010]. These factors have necessitated a further adaptation of unbiased gradient estimates instead of complete ones yielding novel classes of stochastic gradient algorithms (See Bottou et al. [2018] recent review). While the simplicity and performance of such algorithms have cemented their quintessential status, a significant portion of the theoretical literature remains tethered to the idealistic assumptions of a strictly convex underlying objective and i.i.d. data.

[1] An estimator $\hat{\theta}_n^{\mathcal{A}lg}$ for a parameter θ^* is said asymptotically normal if, where:

$$\sqrt{n}(\hat{\theta}_n^{\mathcal{A}lg} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

indicating that $\hat{\theta}_n^{\mathcal{A}lg}$ converges to a normal distribution as the sample size n increases.

[2] An estimator $\hat{\theta}_n^{\mathcal{A}lg}$ for a parameter θ^* is said to be rate(n)-consistent when, where:

$$\text{rate}(n)\cdot(\hat{\theta}_n^{\mathcal{A}lg} - \theta^*) \xrightarrow{p} 0$$

indicating that the likelihood of the difference between $\hat{\theta}_n^{\mathcal{A}lg}$ and θ^* exceeding any fixed $\varepsilon > 0$ diminishes as n grows.

Against this backdrop, the crux of this study seeks to probe these theoretical limits, asking:

Is it possible to design an estimator that is both consistent and asymptotically normal in structured ML-driven non-convex landscapes, especially when dealing with correlated streamed datasets?

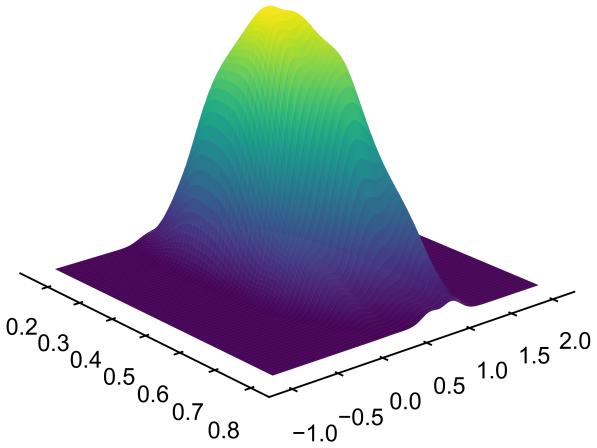
In the above quest, we pursue a twofold approach:

i) *Composite convex optimization.* In a series of machine learning tasks — from proximal gradient methods in reinforcement learning [**?**], Kalman smoothing [Aravkin et al., 2013], to generative models [**?**] — the objective often arises from merging a strongly convex function (typically a penalty function like squared-mean or negative log-likelihood) with an often intractable non-convex differentiable map. In these contexts, the map's outcomes are defined by a low-dimensional space of *control variables*, akin to the bias and weight parameters found in neural networks. Then through this map, the *control variables* are expanded into a higher-dimensional manifold providing the latent variables within the penalty function. From an optimization standpoint, this composition "hides" the strongly convex geometry of the penalty function, leading to highly non-convex optimization landscapes. This class of loss penalties has been of particular interest as many forms of non-convex programming, e.g. SDP, QP, etc. can be formulated in this manner [**?**].

ii) *Correlated Streamed Data:* In a separate axis of research, online data, which is frequently updated in real-time, has become indispensable in today's digital landscape. This trend is particularly evident in time series applications, from financial market trends [Pincus and Kalman, 2004] and weather forecasting [Wang et al., 2015] to patient health monitoring [Fassois and Kopsaftopoulos, 2013]. Moreover, models in these domains sometimes exhibit noise with correlated patterns. Such phenomena are prominent in deep reinforcement learning [Chen et al., 2022] techniques where data are decision-dependent from the previous outcomes through a Markov Decision Process (MDP).

1.1 Our Results & Techniques.

Figure 1: Model parameter distribution given by bootstrap samples. The distribution is asymptotically normal and centered at the true values



In this paper, we aim to offer an *affirmative answer* to the outlined challenges. For this purpose, we assert that the input steaming data only needs to comply with the $\phi-$mixing property [Ibragimov, 1959]—a broader form of time series dependence encompassing classical Markovian dependencies. Fur-

ther, we engage with the "hidden convexity" paradigm, a concept originally sparked by the dynamics of games induced by neural networks [Goodfellow et al., 2014, Perolat et al., 2022].

In this setting, we introduce SHADE, an innovative gradient-based estimator that capitalizes on the representation function linking control and latent variables. This estimator amalgamates techniques like the Polyak-Ruppert acceleration, a natural gradient approach [Mladenovic et al., 2021], and a preconditioned discretization [Sakos et al., 2023].

Aligning with the criteria outlined in our introduction, Theorem 1 confirms the consistency of our estimator in both the latent and control spaces. This outcome emerges from synergizing (1) modern standard descent inequalities in composite convex optimization, (2) the foundational work of [Yu, 1994] offering robust statistical assurances for a stochastic first oracle, and (3) the Robbins-Siegmund convergence theorem tailored to stochastic approximation concerning non-negative near-super-martingales.

Building on these foundations, Theorem 2 demonstrates the SHADE estimator's asymptotic normality through a central limit theorem. Leveraging this asymptotic behavior, we formulate a bootstrap variant of SHADE and show in Theorem 3 that the asymptotic behavior of the estimates matches that of the SHADE itself, facilitating the creation of provable confidence intervals.

To consolidate our findings, we provide empirical evidence underscoring the efficacy of the SHADE estimator, in regression and real-world settings.

2 Problem setup and preliminaries

Formalizing the interactions between control and latent parameters common in machine learning penalties, we utilize the broad and expressive class of *hidden or composite convex* loss functions, characterized as below.

Definition 1 (Latent Convex Function) *A loss function ℓ admits latent or hidden structure if*

1. *The control variables $\theta \in \Theta$ are mapped faithfully to a closed, compact, convex set of latent variables $x \in \mathcal{X} \subseteq \mathbb{R}^d$; the map is Lipschitz smooth $\chi : \Theta \to \mathcal{X}$ with no critical points and $\mathrm{cl}(\chi(\Theta)) = \mathcal{X}$*

2. *The loss factors through the latent space x as*

$$\ell(\theta) = f(\chi(\theta))$$

for a strongly convex Lipschitz smooth function $f : \mathcal{X} \to \mathbb{R}$.

For this class of penalty, any stationary point corresponds to the global optima [Fatkhullin et al., 2023]. To motivate this notion, we provide a couple of illustrative examples.

Example 2.1: Consider equipping the classical linear model with a latent parameter $y_i = w_i^T \chi(\theta)$, where χ is a preconfigured multi-layer perceptron (MLP). Minimizing the squared error loss over a finite dataset yields an objective of

$$\ell(\theta) = \sum_{i=1}^{n} (y_i - w_i^T \chi(\theta_i))^2 = \sum_{i=1}^{n} (y_i - w_i^T x_i)^2$$

Example 2.2: We now turn to a more complex example in convex reinforcement learning. The standard reinforcement learning setting hinges on a Markov decision process, $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, H, \rho, \gamma)$, where $\mathcal{S}, \mathcal{A}$ represent the state and action spaces respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a state transition kernel returning a distribution over states, ρ is the initial state distribution, and γ is the time discount factor. Our goal is to ascribe actions to a distribution over states, through a policy function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We additionally denote $\mathbb{P}_{\rho,\pi}$ to be the probability measure induced by an initial distribution ρ and a policy π. We define a *state-occupancy measure* as follows.

$$\lambda^{\pi}(s, a) = \sum_{t=1}^{\infty} \gamma^t \mathbb{P}_{\rho,\pi}(s_t = s, a_t = a)$$

Denote $\mathcal{U} = \{\lambda^{\pi}; \pi \in \Pi\}$ to be the set of state-occupancy measures. Then given a convex cost function $H : \mathcal{U} \to \mathbb{R}$, our goal is to minimize

$$\min_{\pi \in \Pi} \ell(\pi) = H(\lambda^{\pi})$$

This loss function is not necessarily convex in $\ell(\cdot)$, yet the cost $H(\cdot)$ exhibits convexity in terms of the state-occupancy measures. This latent convex characterization subsumes many reinforcement learning strategies, including pure exploration learning, when $H(\lambda^{\pi})$ represents the negative entropy of the policy π, and imitation learning, where $H(\lambda^{\pi})$ represents the KL divergence between the expert policy and the current. Note in this paradigm, we can only interact with the representation of the loss through the representation.

Given that $\ell(\theta)$ is statistically incomputable, we assume only the stochastic loss function $L(\theta; \Omega)$ is given. As such, we quantify the qualitative definition given above and impose certain criteria common in the literature [Hazan, 2012, Lan, 2020].

Assumption 1 (Loss Function) *The loss function $\ell(\theta)$ is latent convex, and is the expectation of the noisy empirical loss $L : \Theta \times \Omega \to \mathbb{R}$, over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

- *$L(\theta; \omega)$ is differentiable and $c-$Lipschitz in θ for almost all ω.*
- *The gradients of $L(\theta; \omega)$ have bounded $p-$th moments, for some $p > 2$, i.e. $\sup_{\theta \in \Theta} \mathbb{E}[\| \nabla L(\theta; \omega)\|^p] \leq M^p$*
- *The Jacobian of the representation mapping $\mathrm{Jac}_{\chi}(\theta)$ has a bounded spectra; i.e. $\sigma_{\min}(\mathrm{Jac}_{\chi}(\theta)) > 0$ and $\sigma_{\max}(\mathrm{Jac}_{\chi}(\theta)) < \infty$*

Remarks: Combining the parts of Assumption 1 implies $\ell(\theta)$ is a Lipschitz smooth differentiable function and $\nabla L(\theta, \omega)$ is an unbiased estimator of $\nabla \ell(\theta)$. Moreover, the second assumption is common in the literature regarding "sandwich" covariance estimators [Liu et al., 2023].

Notation: To streamline notation, we denote the parameters in the control space θ and write $x = \chi(\theta)$, for the induced latent map. When the representation mapping χ is clear from context, we write $\mathbf{J}(\theta) := \mathrm{Jac}(\chi(\theta))$. Moreover, for brevity, we denote the gradient concerning multiple datapoints

$$\nabla L(\theta; \Omega) := \nabla \frac{1}{|\Omega|} \sum_{s \in \Omega} L(\theta; \omega_s)$$

A similar substitution is made for the monotonic loss $\nabla f(x; \Omega) = \nabla \frac{1}{|\Omega|} \sum_{s \in \Omega} f(x; \omega_s)$. The term $(\cdot)^+$ refers to the Moore-Penrose pseudoinverse.

$\phi-$**Mixing:** Furthermore, we assume the data comes from a $\phi-$mixing process [Ibragimov, 1959]. Formally, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{A}, \mathcal{B}$ are sub-sigma-algebras of $\mathcal{F}$, the mixing coefficient for $\mathcal{A}, \mathcal{B}$ is defined

$$\phi_P(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}, P(B)>0} |P(B|A) - P(B)|$$

For a sequence of data $\{\omega_s\}_{s=1}^{\infty}$, we define $\mathcal{F}_a^b$ to be the $\sigma-$algebra generated by $\{\omega_s\}_{s=a}^{b}$ and $\phi_\Omega(t)$

$$\phi_\Omega(t) = \sup_{s \geq 1} \phi_P(\mathcal{F}_1^s, \mathcal{F}_{s+t}^{\infty}).$$

The expression $\phi_\Omega(t)$ can be thought of as the maximum amount of information gained, here in the form of a conditional probability, by knowing data from t or more steps in the past. For a sequence to qualify as $\phi-$mixing, the value of $\phi_\Omega(t)$ must converge to zero as t diverges to infinity. Examples of $\phi-$mixing sequences include Markov decision processes [Bradley, 2005], low-rank vector autoregressive processes [Rémillard et al., 2012], and Gaussian processes [Samson, 2000]. As an example of the power of this condition, we present the following proposition.

Proposition 1 (Markov Dependence) *Let $X = (X_k)_{k=0}^{\infty}$ be a Markov chain. Then if $\phi_\Omega(n) < 1$ for some $n \geq 1$, then $\phi_\Omega(n) \to 0$ at least exponentially fast as $n \to \infty$.*

This fact is shown in the appendix.

Covariance Estimation: To adequately conduct statistical inference tasks, e.g. creating confidence intervals, the covariance structure of the parameters needs to be estimated. However, given that the data is correlated, this often contains auto-correlated components and is difficult to determine. To remedy these issues, we leverage a version of the bootstrap SGD algorithm, introduced by Fang et al. [2018], outlined below is a computationally efficient method for elucidating the asymptotic covariance and is detailed as follows.

$$\theta_t^{\bullet} = \theta_{t-1}^{\bullet} - \gamma_t U_t \nabla L(\theta_{t-1}^{\bullet}, \omega_t) \qquad \text{(Bootstrap SGD)}$$

The U_t values are random variables with mean and variance one. Under i.i.d. data, the authors demonstrated the convergence of the Polyak-Ruppert average of the bootstrap estimates to the same limiting distribution as the averaged SGD iterates. If $\theta^{\bullet}$ represents the bootstrap SGD iterate, then the following holds.

$$\frac{1}{\sqrt{n}}(\sum_{k=1}^{n} \theta^{SGD} - \theta_0), \frac{1}{\sqrt{n}}(\sum_{k=1}^{n} \theta_k^{\bullet} - \sum_{k=1}^{n} \theta_k)|\omega_1, ..., \omega_n \xrightarrow{\mathbb{L}} \varphi$$

3 Latent Estimation

Throughout this section, we introduce the stochastic hidden averaging data estimator (SHADE), our parameter estimator for inference in latent convex regimes. By leveraging gradient preconditioning and mini-batching, our goal is to overcome (*i.*) the pitfalls of traditional stochastic gradient descent estimators within non-convex loss landscapes and, (*ii.*) the idealistic expectation that the data is created from an i.i.d. process.

Gradient Preconditioning: We begin by addressing the first issue by building the parameter estimator from a *natural gradient flow*. Given a function $\ell : \Theta \to \mathbb{R}$, and a class of metric tensors, here a set of parameterized positive semidefinite matrices $\mathbf{P}_\theta, \theta \in \Theta$, the natural gradient flow is given by the steepest descent in the geometry induced by the preconditioner matrix.

$$\dot{\theta} = -\mathbf{P}_\theta^{-1} \nabla \ell(\theta)$$

Commonly used preconditioners include the identity matrix, which captures the standard geometry of Euclidean spaces and yields the vanilla gradient flow, and the Fisher information matrix which can be characterized as follows. If p_θ is a probability measure, under suitable regularity conditions

$$\dot{\theta} := -\mathbf{P}_\theta^{-1} \nabla \ell(\theta) \text{where } \mathbf{P}_\theta := -\mathbb{E}_{x \sim p(\theta)}[\nabla^2 \log(p_\theta(x))]$$

In this same vein, given the strongly convex nature of the penalty function relative to the output of the representation map $\ell(\theta) = f(\chi(\theta))$, Sakos et al. [2023] propose the preconditioned hidden gradient flow, which uses a metric that captures the complex geometry of the representation mapping.

$$\dot{\theta} = -\mathbf{P}_\theta \nabla \ell(\theta), \mathbf{P}_\theta = [\text{Jac}(\chi(\theta))^T \text{Jac}(\chi(\theta))]^+$$

The term $(\cdot)^+$ refers to the Moore-Penrose pseudoinverse. Via showing the L_2 energy function $E(\theta) = \frac{1}{2}\|\chi(\theta) - x^*\|^2$ is a Lyapunov function, the authors successfully demonstrated the flows converges despite the lack of separability arguments for the representation mapping. Below we demonstrate why such a preconditioner is necessary to achieve near optimal The discretized flow, the preconditioned hidden gradient descent, detailed below, moreover, was shown to converge polynomially in the latent convex regime. Thus as a starting point for building our estimator, we leverage the iterates from a preconditioned hidden gradient descent process to ensure the consistency of our estimates.

Algorithm 1 Pre-Conditioned Hidden Gradient Descent

Input: Data $\{\omega_s\}_1^\infty$, learning rate γ_t, block size B_t, intial value θ_0.
for $t = 1$ **to** T **do**
  Construct index set I_t based on block size B_t.
  Compute the Jacobian of χ at θ_{t-1}, $\text{Jac}(\chi(\theta_{t-1})) := \mathbf{J}_{\theta_{t-1}}$.
  Compute $\mathbf{P}(\theta_{t-1}) := [\mathbf{J}_{\theta_{t-1}}^T \mathbf{J}_{\theta_{t-1}}]^+$
  Compute $V_{t-1} := \nabla L(\theta_{t-1}; \Omega_t)$
  Update $\theta_t = \theta_{t-1} - \gamma_{t-1} \mathbf{P}(\theta_{t-1}) V_t$
end for
return θ_T

Comparison With Gradient Descent Based Estimators: In the literature of convex optimization, Fatkhullin et al. [2023] have recently shown that stochastic gradient descent methods converge deterministically to the global optima of latent convex penalties, like the aforementioned preconditioned hidden gradient descent. The authors additionally show that for certain classes of smooth convex functions, up to logarithmic factors, sample complexities, and convergence rates known for convex and strongly convex objectives can be achieved for this particular class. Yet for still others, in comparison to the convergence of preconditioned hidden gradient descent, the rates lag, signaling the power of the Jacobian. As an example, we observe the following proposition.

Proposition 2 (Convergence Rate of Stochastic Descent in Hidden Convex Objectives) *Let $\ell(\theta)$ be a composite convex function. Fix $\epsilon > 0$. Then, we aim to have a bound for T such that*

$$\mathbb{E}[E(\theta_T; x^*)] = \frac{1}{2}\|\chi(\theta_T) - x^*\|^2 \leq \epsilon$$

If $\ell(\theta) = f(\chi(\theta))$ is merely composite convex, i.e. f is merely convex, then $\mathbb{E}[E_T] \leq \epsilon$ when $T = \tilde{\mathcal{O}}(\epsilon^{-3})$ if θ is an iterate of stochastic gradient descent, and $T = \tilde{\mathcal{O}}(\epsilon^{-2})$ if it is an iterate of PHGD.

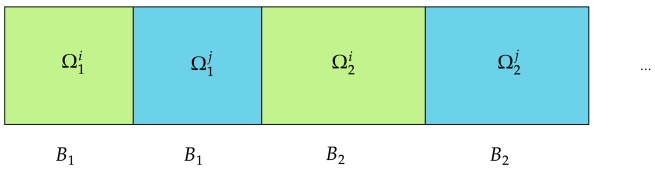
This fact is shown in the appendix and highlights the advantages PHGD has over traditional gradient descent, with the bound PHGD attains matching that of the merely convex case for stochastic gradient descent. Moreover, in practical settings, as we will examine in the experiments, SGD often struggles to robustly estimate the true parameters in comparison to PHGD.

Mini-Batching: The complex nature of $\phi-$mixing sequences yields that limiting distributions of bootstrap estimators may not share the same structure as those built from empirical risk minimization solutions. Indeed, Liu et al. [2023] point out that the results from Fang et al. [2018] do not extend to the mixing cases, as the two quantities can differ greatly asymptotically. To counteract this, they propose a mini-batch estimator inspired by the independence block trick from Yu [1994] which takes advantage of the decreased correlation of temporally separated data. The authors partition the dataset $\{\omega\}_t^\infty$ into non-overlapping blocks with the following indices, where B_t is the block size at time t.

$$I_t = \{2\sum_{k=1}^{t-1} B_k, ..., 2\sum_{k=1}^{t-1} B_k + B_t\}$$
$$J_t = \{2\sum_{k=1}^{t-1} B_k + B_t + 1, ..., 2\sum_{k=1}^{t} B_k\}$$

Parameter estimates are found by averaging two Polyak-Ruppert SGD estimators built off of observations from non-overlapping blocks.

Figure 2: Visualization of data blocks



$$\bar{\theta} = \frac{1}{2T}\sum_{t=1}^{T}(\theta_t^i + \theta_t^j) \text{where } \theta_t^k = \theta_{t-1}^k - \gamma_t \nabla L(\theta_{t-1}^k; \Omega_t^k), \text{and } \Omega_t^k = \{\omega_i | i \in K_t\}$$

This step along with the properties of $\phi-$mixing sequences imply the following approximation.

$$\mathbb{E}[\sum_{s\in\Omega_{t+1}} \nabla L(\theta_{t+1}; \omega_s)|\theta_t] \approx \nabla \ell(\theta_t)$$

This falls in line with the existing stochastic optimization literature which has shown both theoretically, in terms of sample complexity [Ma et al., 2022], and practically, with the ever larger scope of datasets.

Stochastic Hidden Data Averaging Estimator: Leveraging insights from both mini-batch estimation and preconditioned descent schemes, we arrive at the stochastic hidden data averaging estimator (SHADE). From the parameter estimates generated from PHGD, we use the mini-batch ideas from above to derive the following estimator.

$$\hat{\theta}_{SHADE} = \frac{1}{2T}\sum_{t=1}^{T}(\theta_t^i + \theta_t^j), \text{where } \theta_t^k = \theta_{t-1}^k - \gamma_t \mathbf{P}_\theta \nabla L(\theta_{t-1}^k; \Omega_t^k) \tag{SHADE}$$

By using the preconditioned descent algorithm we can ensure the consistency of our estimator and can confidently conduct inferential tasks.

4 Asymptotic analysis and results

In this section, we present our main results regarding the asymptotic behaviour of SHADE, and the bootstrap estimates, showing they are asymptotically consistent and normal. To this end, the section begins with a discussion of our model of $\phi-$mixing data sequence and its interplay with the batch size. The proofs are deferred to the appendix.

Model 1 (Descent Parameters) *Our model of the asymptotic nature of the correlated, streaming data is characterized as follows. We impose rate limiting conditions on the $\phi-$mixing data sequence $\{\omega_s\}_{s=1}^{\infty}$ and associated batch size B_t.*

- *The learning rate of the algorithm satisfies $\gamma_t = (\gamma_0 + t)^{-\rho}$*
- *The mixing coefficients satisfy $\sum_{t=1}^{\infty}\sqrt{\phi(t)} < \infty$ and that $\sum_{t=1}^{\infty}\phi^{1-2/p}(t) < \infty$ and $\phi^{1-1/p}(B_t) < \infty$*
- *Our batch size increases to infinity as t increases and that $\sum_{t=1}^{\infty} t^{-\rho}\phi^{1/2-1/p}(B_t) < \infty$*
- *The mixing conditions, batch size and learning rate satisfy the following $\phi^{1/2-1/p}(B_t) = O(t^{-\rho})$, $\sum_{j=1}^{t}\phi^{1/2-1/p}(B_t), \sum_{j=1}^{t} j^{-\rho} = o(\sqrt{\sum_{j=1}^{t} B_j^{-1}})$, $t^{\rho} = o(\sum_{j=1}^{t} B_j^{-1})$*

Remarks: These conditions provide some rate limitations for both the learning rate and the mixing coefficients. The first assumption requires that the learning rate satisfies, $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$ and that $\sum_{t=1}^{\infty}\gamma_t = \infty$; this assumption is widely used in the literature [Polyak, 1990b], [Fang et al., 2018], [Sakos et al., 2023], [Liu et al., 2023]. The remaining assumptions are the so-called algebraic mixing conditions and ensure the covariance matrix is well-defined [Fan and Yao, 2003], and imply the batch size increases to infinity asymptotically. The final assumption controls higher-order error terms.

The batch size diverges asymptotically to ensure the estimator stays $\sqrt{n}$-consistent. Indeed, as we will see in the sequel, our estimates have a convergence rate of $\sqrt{\sum_{t\leq T} B_t^{-1}}/T$ given $2T$ batches. Note that if we replace B_t with a constant, the rate degenerates to a constant speed.

Convergence Results: We are now in a position to state the main results regarding the asymptotic behavior of the SHADE estimator. To streamline the presentation, we begin with results regarding consistency before analysis of the asymptotic distribution.

Theorem 1 (Consistency) *Given the assumptions in part 2, we conclude that the SHADE estimator satisfies*

$$\hat{\theta}_{SHADE} \xrightarrow{a.s.} \theta^*.$$

This theorem ensures the SHADE estimator is strongly consistent and follows the results found in Polyak and Juditsky [1992a] and Liu et al. [2023]. This proof relies on the Robbins-Siegmund theorem [Robbins and Siegmund, 1971], a result which shows that if Lyapunov functions of a non-negative martingale converge to zero, the overall function converges to the optimum. By showing the L_2 energy function

$$E(\theta; x^*) = \frac{1}{2}\|\chi(\theta) - x^*\|^2$$

is asymptotically zero, we additionally show the iterates of the stochastic process converge to the optima.

Asymptotic Normality: We proceed with characterizations of the asymptotic distributions of our estimators. In advance of discussing the covariance structure, we define the autocorrelation coefficients.

Definition 2 (Autocorrelation Coefficients) *Let $L(\theta; \Omega)$ be a stochastic loss function satisfying the assumptions laid out in Assumption 1. Then we define*

$$r(t) = \mathbb{E}[\nabla L(\theta_{k+t}; \Omega_{k+t}) \nabla L(\theta_k; \Omega_k)^T]$$

Theorem 2 (Asymptotic Normality of SHADE) *Given the assumptions above, the following correspondence occurs*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^{T} B_t^{-1}}}(\hat{\theta}_{SHADE} - \theta^*) \to \mathcal{N}(0, \mathbf{J}(\theta^*)^+ \Sigma [\mathbf{J}(\theta^*)^+]^T)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, $G = \nabla^2 f(x^*)$ and $\mathbf{J}(\theta^*)$ is the Jacobian evaluated at the optimal θ value.

This result mirrors the central limit theorem found in Polyak and Juditsky [1992a], Liu et al. [2023], and Mou et al. [2020], and uses the variance structure found in [Fan and Yao, 2003] It is also important to note that this theorem subsumes central limit theorem attained under i.i.d. data, as in this case, the autocorrelation coefficient $r(k) = 0$, and yields the sandwich estimator found in Polyak [1990b] and Su and Zhu [2018].

$$T^{-1/2}(\hat{\theta}_{SHADE} - \theta^*) \to \mathcal{N}(0, \mathbf{J}(\theta^*)^+ G^- 1 r(0) G^{-1} [\mathbf{J}(\theta^*)^+]^T)$$

Moreover, the asymptotic normality results can be extended to the bootstrap estimates,by showing the sequence satisfies the Lindeberg condition [Brown, 1971].

Theorem 3 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

$$\frac{T}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}(\hat{\theta}_T^* - \hat{\theta}_{SHADE}) | \mathcal{D} \xrightarrow{\mathbb{L}} \mathcal{N}(0, \hat{\Sigma})$$

where $\mathcal{D} = \{\omega_i | i \in I_t \cup J_t\}$ and represents the data used in the empirical risk minimization process, and $\hat{\Sigma} = \mathbf{J}(\theta^*)^+ G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}[\mathbf{J}(\theta^*)^+]^T$ is the covariance matrix in Theorem 2

This theorem provides the guarantee that the asymptotic distributions of the bootstrap estimates will match that of SHADE. This ensures the consistency of these estimates and from this we can construct confidence regions and find standard errors.

Given that the proofs of Theorems 1-3 are quite involved, we defer proofs to the appendix. To show a few of the ideas that aided in the construction of the results, we now show a few pivotal lemmas (without proof).

Lemma 1 (Representation Gap) *If the assumptions on the singular values in part three hold, then the following inequality holds, where* $\sigma_{\min}, \sigma_{\max}$ *are the smallest and largest singular values respectively.*

$$\sigma_{\min}||\theta - \theta^*|| \leq ||\chi(\theta) - \chi(\theta^*)|| \leq \sigma_{\max}||\theta - \theta^*||.$$

This lemma controls the 'distortion' of the distance metric between the latent and control spaces by the representation mapping and performs a key role in the analysis. To take advantage of the simpler geometry of the latent space, we characterize the evolution of the latent iterates in the following manner.

Lemma 2 (Descent Equality) *Let* $x_t = \chi(\theta_t)$*, we then have that for iterates of PHGD*

$$x_{t+1} = x_t - \gamma_t U_t \nabla f(x_t; \Omega) + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2).$$

This lemma shifts problems of convergence and consistency into the strongly convex latent space, where the analysis of martingale sums simplifies dramatically.

Importantly, the asymptotic results show that the SHADE estimator is (i.) consistent and (ii.) asymptotically normal, despite the impediment of mixing data, ensuring the soundness of our statistical estimates. This demonstrates that the latent convex structure of our penalty can be exploited to conduct robust inference.

6. Conclusion

This paper proposed a new statistical estimator with strong asymptotic convergence guarantees for inference problems with a latent convex penalty with mixing data. Our estimator extends the Polyak-Ruppert averaging scheme to iterate of the preconditioned gradient descent and is successful in modeling the latent/hidden variables that arise in neural networks and can thus model numerous AI applications. These results emerge from the interplay of non-convex optimization, online learning and statistical estimation and open the door for future work.

References

Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012.

Aleksandr Y. Aravkin, James V. Burke, and Gianluigi Pillonetto. Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: modeling, computation, and theory. *J. Mach. Learn. Res.*, 14(1):2689–2728, 2013.

Gunjan Arora, Jayadev Joshi, Rahul Shubhra Mandal, Nitisha Shrivastava, Richa Virmani, and Tavpritesh Sethi. Artificial intelligence in surveillance, diagnosis, drug discovery and vaccine development against covid-19. *Pathogens*, 10(8):1048, 2021.

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

Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

Bruce M Brown. Martingale central limit theorems. *The Annals of Mathematical Statistics*, pages 59–66, 1971.

Elynn Y. Chen, Rui Song, and Michael I. Jordan. Reinforcement learning with heterogeneous data: Estimation and inference. *CoRR*, abs/2202.00088, 2022.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. 2020.

National Research Council et al. *Frontiers in massive data analysis*. National Academies Press, 2013.

Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*, volume 20. Springer, 2003.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 2018.

Spilios D. Fassois and Fotis P. Kopsaftopoulos. *Statistical Time Series Methods for Vibration Based Structural Health Monitoring*, pages 209–264. Springer Vienna, Vienna, 2013.

Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic optimization under hidden convexity, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2011.

Elad Hazan. A survey: The convex optimization approach to regret minimization. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 287–304. MIT Press, 2012.

IA Ibragimov. Some limit theorems for stochastic processes stationary in the strict sense. In *Dokl. Akad. Nauk SSSR*, volume 125, pages 711–714, 1959.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer, New York, NY, USA, 2nd edition, 2003. ISBN 9780387008943. doi: 10.1007/b97441. URL https://link.springer.com/book/10.1007/b97441.

G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Series in the Data Sciences. Springer International Publishing, 2020. ISBN 9783030395681. URL https://books.google.com/books?id=7dTkDwAAQBAJ.

Ruiqi Liu, Xi Chen, and Zuofeng Shang. Statistical inference with stochastic gradient methods under $\phi-$mixing data. *arXiv preprint arXiv:2302.12717*, 2023.

Shaocong Ma, Ziyi Chen, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Data sampling affects the complexity of online sgd over dependent data. In *Uncertainty in Artificial Intelligence*, pages 1296–1305. PMLR, 2022.

Reid McMurry, Patrick Lenehan, Samir Awasthi, Eli Silvert, Arjun Puranik, Colin Pawlowski, AJ Venkatakrishnan, Praveen Anand, Vineet Agarwal, John C O'Horo, et al. Real-time analysis of a mass vaccination effort confirms the safety of fda-authorized mrna covid-19 vaccines. *Med*, 2(8): 965–978, 2021.

Andjela Mladenovic, Iosif Sakos, Gauthier Gidel, and Georgios Piliouras. Generalized natural gradient flows in hidden convex-concave games and gans. In *International Conference on Learning Representations*, 2021.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Steve Pincus and Rudolf E. Kalman. Irregularity, volatility, risk, and financial market time series. *Proceedings of the National Academy of Sciences*, 101(38):13709–13714, 2004.

Boris T. Polyak. New stochastic approximation type procedures. *Automation and Remote Control*, 51 (7):98–107, Jul 1990a.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, Jul 1992a. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992b.

Boris Teodorovich Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika*, (7):98–107, 1990b.

Bruno Rémillard, Nicolas Papageorgiou, and Frédéric Soustra. Copula-based semiparametric models for multivariate time series. *Journal of Multivariate Analysis*, 110:30–42, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Iosif Sakos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Panayotis Mertikopoulos, and Georgios Piliouras. Exploiting hidden structures in non-convex games for convergence to nash equilibrium. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=05P1U0jk8r.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Bianca Schroeder and Garth A Gibson. A large-scale study of failures in high-performance computing systems. *IEEE transactions on Dependable and Secure Computing*, 7(4):337–350, 2009.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

Nathan Srebro and Ambuj Tewari. Stochastic optimization for machine learning. *ICML Tutorial*, 2010.

Weijie Su and Yuancheng Zhu. Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, page 3, 2018.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Yu Wang, Guanqun Cao, Shiwen Mao, and R. Mark Nelms. Analysis of solar generation and weather data in smart grid with simultaneous inference of nonlinear time series. In *2015 IEEE Conference on Computer Communications Workshops, INFOCOM Workshops, Hong Kong, China, April 26 - May 1, 2015*, pages 600–605. IEEE, 2015.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Ryozo Yokoyama. Moment bounds for stationary mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 52(1):45–57, 1980.

Bin Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94 – 116, 1994.

A Background and Preliminaries

A.1 Background in dynamical systems:

Our analysis combines tools from dynamical systems, probability theory, and stochastic algorithms. We begin with a discussion of asymptotic stability and Lyapunov's theorem, culminating in the introduction of a Lyapunov function for the preconditioned hidden gradient descent. Dynamical systems theory plays a role in laying the groundwork for the Robbins-Siegmund theorem, which is used to prove the asymptotic consistency of our estimators.

We define $\mathbf{f} : D \to \mathbb{R}^n$ to be a local Lipschitz map from a subset $D \subset \mathbb{R}^n$. In this section, we consider dynamical systems of the form

$$\dot{\mathbf{x}} = f(\mathbf{x}) \tag{A.1}$$

If $f(\mathbf{x}^*) = 0$, we denote $\mathbf{x}^*$ to be a fixed point. We can characterize stability in the following manner.

Definition A.1 (Stability Properties) *The dixed point $\mathbf{x} = 0$ of Equation A.1 is*

- ***stable*** *if, $\forall \epsilon > 0$, $\exists \delta > 0$ such that*

$$\|\mathbf{x}(0)\| < \delta \to \|\mathbf{x}(t)\| < \epsilon, \forall t \geq 0$$

- ***unstable*** *if it is not* ***stable***
- ***asymptotically stable*** *if it is* ***stable*** *and δ can be chosen such that*

$$\|\mathbf{x}(0)\| < \delta \to \lim_{t \to \infty} \|\mathbf{x}(t)\| = 0$$

The Lyapunov theorem is useful for proving asymptotic stability and can be seen as a precursor to the Robbins-Siegmund theorem used in section B.

Theorem A.1 *Let $\mathbf{x} = 0$ be a fixed point for Equation A.1, and let $D \subset \mathbb{R}^n$ contain 0. Let $V : D \to \mathbb{R}$ be a continuously differentiable function such that*

$$V(0) = 0 \text{ and } V(\mathbf{x}) > 0 \text{ in } D/\{0\}, \dot{V}(\mathbf{x}) \leq 0 \text{ in } D$$

then $\mathbf{x} = 0$ is stable. Furthermore, if

$$\dot{V}(\mathbf{x}) < 0 \text{ in } D/\{0\}$$

then $\mathbf{x} = 0$ is asymptotically stable.

In Sakos et al. [2023], the authors demonstrated the L_2 energy function satisfies the conditions of the $V(\cdot)$ function above.

Lemma A.1 *Let $E(\theta; x^*)$ be the L_2 energy function.*

$$E(\theta; x^*) = \frac{1}{2}\|\chi(\theta) - x^*\|^2$$

Then $E(\theta; x^*)$ satisfies the Lyapunov theorem above

Proof. This is Proposition 1 in Sakos et al. [2023] □

A.2 Gradient notation:

Here we characterize the gradient of the latent convex loss $\ell(\theta) = f(\chi(\theta))$, into a sum of 3 martingales. This is similar to the analysis done in Polyak [1990a] and allows tight second-moment gradient bounds to be formed. The minibatch estimators $\hat{\theta}^i, \hat{\theta}^j$ have similar functions within the SHADE estimator, so without loss of generality it is sufficient to study $\hat{\theta}^i$. For ease of notation, we omit the superscripts in the sequel.

Both preconditioned hidden gradient descent and its bootstrap counterpart can be subsumed into the following characterization.

$$\theta_{t+1} = \theta_t - \gamma_{t+1} U_t \mathbf{P}(\theta_t) V_t \tag{A.2}$$

U_t are random variables with the following assumption.

Assumpsion A.1 *The i.i.d. random variables U_t are independent from the data process $\{\omega_k\}_{k=1}^{\infty}$. In addition, $\mathbb{E}[U_t] = 1$ and $\mathbb{E}[U_t^p] < \infty$ for the p introduced in Assumption 1.*

It is clear to see that when U_t is the identity we recover the preconditioned hidden gradient descent, and when $\mathbb{E}[U_t] = 1$ and $\mathrm{Var}(U_t) = 1$, we obtain the bootstrap variant. We use the following notation for the analysis of the behaviour of the gradient.

Definition A.2 *The gradient of the loss function can be characterized as follows.*

- $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\Omega_1, ..., \Omega_t]$
- $h(x) = \nabla f(x)$
- $e_t = \mathbb{E}_{T-1}[\nabla f(x_{T-1}; \Omega_T)] - \nabla f(x_{T-1})$
- $\zeta_t = U_t \nabla f(x_{T-1}; \Omega_T) - \mathbb{E}_{T-1}[\nabla f(x_{T-1}; \Omega_T)]$
- $\hat{g}_t = \nabla f(x; \Omega) = \frac{1}{|\Omega|} \sum_{s \in \Omega} \nabla f(x; \omega_s))$

Thus we can rewrite the descent procedure as the following.

$$\theta_{t+1} = \theta_t - \gamma_{t+1} U_t \mathbf{P}(\theta_t) \mathbf{J}(\theta_t)^T (h(\chi(\theta_t) + e_t + \zeta_t) \tag{A.3}$$

A.3 Moment inequalities:

We introduce a few key moment inequalities for $\phi-$mixing sequences that will be of use when bounding sums of martingale sequences. The second point in Model 1 allows the use of a theorem from Yokoyama [1980], which gives sharp bounds on sums of martingale iterates.

Lemma A.2 (Moment Inequality) *Let $\{X_t\}_{t=1}^{\infty}$ be a stationary sequence with ϕ-mixing coefficients bounded by the operator $\phi(t)$ defined analogously to above. In addition, we require that $\sum_{t=1}^{\infty} \sqrt{\phi(t)} < \infty$, $\mathbb{E}[X_t] = 0$ and that $\mathbb{E}[\|X_t^k\|] < \infty$ for some value $k > 2$. Then $\mathbb{E}(|\sum_{i=1}^{T} X_t|^k) \leq C_k T^{k/2}$*

Proof. This result is theorem 3 in Yokoyama [1980] □

Lemma A.3 *Let (X, Y) and $(X, \tilde{Y})$ be random vectors where Y and $\tilde{Y}$ with the same marginal distributions and m be an arbitrary constant. Then we claim*

$$\| \mathbb{E}[h(X,Y)|X] - \mathbb{E}[h(X,\tilde{Y})|X]\| \leq m\phi(X,Y) + \frac{\mathbb{E}[\|h(X,Y)\|^p|X]}{m^{p-1}} + \frac{\mathbb{E}[\|h(X,\tilde{Y})\|^p|X]}{m^{p-1}}$$

Proof. This moment inequality is Lemma S.5 from Liu et al. [2023] □

A.4 Template inequality:

We now introduce a template inequality that relates the evolution of the energy Lyapunov function, $E(\theta; x^*)$ throughout time. This plays an important role in our analysis because it allows us to relate the error in the game's control space with the evolution of the algorithm's quasi-Lyapunov function in the game's latent space.

Proposition A.1 (Template inequality) *Let $\ell(\theta) = \mathbb{E}[L(\theta; \omega)]$ be a composite convex loss function. Then for all $\tilde{x} \in \mathcal{X}$ the iterates of (PHGD) satisfy the template inequality, where $E_t = E(\theta_t; x^*) = \|\chi(\theta_t) - x^*\|^2$:*

$$E_{t+1} \leq E_t - \gamma_t \langle h(x_t), x_t - \tilde{x}\rangle + \gamma_t \alpha_t + \gamma_t^2 \psi_t$$

where $\alpha_t = \sum_{i \leq N} \langle \mathbf{J}(\theta_t)^T \hat{g}_t, x_t - \tilde{x}\rangle$ and $\psi_t = \kappa \|V_t\|^2$, for some constant $\kappa > 0$.

Proof. This is an extension from the template descent inequality from Sakos et al. [2023], which can be achieved via linearity in argument. This claims is created via creating bounds on the Taylor expanded potential function. □

A.5 Analysis of convergence rates:

We conclude with a comparison of the convergence rates of stochastic gradient descent and preconditioned hidden gradient descent on latent convex objectives. The PHGD iterates achieve the (up to a logarithmic factor) the sharp convergence bounds of convex objectives, in contrast to iterates of SGD which have a sub-optimal rate. Moreover, in practical settings, stochastic gradient descent often fails to robustly converge, where objectives are not quite convex.

Proposition 2 (Convergence Rate of Stochastic Descent in Hidden Convex Objectives) *Let $\ell(\theta)$ be a composite convex function. Fix $\epsilon > 0$. Then, we aim to have a bound for T such that*

$$\mathbb{E}[E(\theta_T; x^*)] = \frac{1}{2}\|\chi(\theta_T) - x^*\|^2 \leq \epsilon$$

If $\ell(\theta) = f(\chi(\theta))$ is merely composite convex, i.e. f is merely convex, then $\mathbb{E}[E_T] \leq \epsilon$ when $T = \tilde{\mathcal{O}}(\epsilon^{-3})$ if θ is an iterate of stochastic gradient descent, and $T = \tilde{\mathcal{O}}(\epsilon^{-2})$ if it is an iterate of PHGD.

Proof. In preparation for the subsequent steps we define

$$\Lambda_T = \ell(\theta_T) - \ell(\theta^*) + \frac{\rho}{2}\|\theta_T - \theta^*\|^2$$

In their recent review of the role of stochastic gradient descent in latent convex objectives Fatkhullin et al. [2023] showed that when $T = \tilde{\mathcal{O}}(\frac{\beta D_{\mathcal{U}}}{\kappa^2}\frac{1}{\epsilon} + \frac{\beta D_{\mathcal{U}} M^2}{\kappa^4}\frac{1}{\epsilon^3})$, $\mathbb{E}[\Lambda_T] < \epsilon$. Here, κ is the Lipschitz coefficient of the latent map χ, and $D_{\mathcal{U}}$ is a constant stemming from the fact that $f(\cdot)$ is strongly convex and thus satisfies the Kurdyka-Lojasiewicz (KL) condition [Karimi et al., 2016].

$$D_{\mathcal{U}} \geq 2\beta(f(x) - f(x^*))$$

From the nature of $f(\cdot)$ we can deduce that

$$\ell(\theta_T) - \ell(\theta^*) = f(\chi(\theta_T)) - f(\chi(\theta^*)) \geq \frac{\beta}{2}\|x_T - x^*\|^2$$

Thus from the Lipschitz properties of $\chi(\cdot)$, it follows that

$$\Lambda_T \geq (\frac{\beta + \kappa\rho}{2})\|x_T - x^*\|^2 = (\frac{\beta + \kappa\rho}{2})E_T$$

So it follows that the desired claim is shown for SGD. To proceed, we define g to be a convex function, and the restricted merit function as follows.

$$\text{Gap}_{\mathcal{C}}(\hat{x}) = \sup_{x \in \mathcal{C}} \langle g(x), \hat{x} - x \rangle$$

The affine function $g(x) = x - \hat{x}$, indeed satisfies the condition of being merely convex. Thus now turning attention to PHGD, through Theorem 1 of Sakos et al. [2023] demonstrated in the hidden merely convex case that the averaged iterate satisfies the following convergence rate

$$\mathbb{E}[\text{Gap}_{\mathcal{C}}(\chi(\bar{\theta}_T))] = \tilde{\mathcal{O}}(t^{-1/2})$$

Inverting the convergence rate to achieve the sample complexity achieves the desired result. □

B Asymptotic consistency

Our goal in this appendix is to prove Theorem 1 which we restate below for convenience.

Theorem 1 (Consistency) *Given the assumptions in part 2, we conclude that the SHADE estimator satisfies*

$$\hat{\theta}_{SHADE} \xrightarrow{a.s.} \theta^*.$$

Our proof strategy is comprised of the following steps.

1. In Lemma A.3, we show the second moment of these martingale terms are then bounded convex analysis; in turn, in Lemma A.4 we establish a bound on the entire second moment in terms of the $\phi-$mixing coefficients and the energy function.

2. From these bounds and the Lyapunov properties of the energy function, the Robbins-Siegmund theorem [Robbins and Siegmund, 1971], the consistency of the averaged latent iterates is shown

3. Applying ideas from real analysis, we can relate the latent and control spaces together, allowing the consistency of SHADE to be shown.

In the sequel, these notions are made precise via a series of intermediate results.

B.1 Martingale Bounds

As a starting point, we aim to create a second moment bound on different parts of the loss gradient. To deal with the difficulty of a latent convex objective, we create these bounds in the latent space, rather than the control space, allowing us to use all the machinery of convex functions. We use the moment bounds in the previous section and the independence block trick to relate gradient terms to the energy function with the final goal of relating our estimators with the energy function.

Lemma B.1 *We seek the following bounds on elements of the descent, where v_t is a quantity with finite first moment:*

(i) $\mathbb{E}\,\|e_t\|^2 \leq \phi^{1-2/p}(B_{t-1})v_t$

(ii) $\mathbb{E}\,\|[\hat{g}_t - h(x_t)\|^2 \leq \frac{C}{B_t}(1 + \|x_t - x^*\|^2)$

(iii) $\mathbb{E}_t(\|\zeta_t\|^2) \leq \phi^{1-2/p}(B_{t-1})v_t + \frac{C}{B_t}(1 + \|x_{t-1} - x^*\|^2)$

Proof. (i.) The auxiliary term $\tilde{e}_t$ is equal to $\mathbb{E}_t[\nabla f(x_t;\Omega_t)] - h(x_t)$, where $\tilde{\omega}_s$ is both independent from and i.i.d. to our ω time series. Via independence, this value compresses into Gaussian noise and has mean zero. Using lemma 5,

$$\mathbb{E}_t(\|e_t\|^2) = |\,\mathbb{E}_t(\|e_t\|^2) - \mathbb{E}_t(\|\tilde{e}_t\|^2)| \tag{B.1}$$

$$\mathbb{E}_t(\|e_t\|^2) \leq m\phi(B_t) + \frac{\mathbb{E}_t[\|e_t\|^p]}{m^{p/2-1}} + \frac{\mathbb{E}_t[\|\tilde{e}_t\|^p]}{m^{p/2-1}} \tag{B.2}$$

So because our variable m is arbitrary, we define it to be $\phi^{-2/p}(B_t)$. So (A.2) is

$$\mathbb{E}_t(\|e_t\|^2) \leq \phi^{1-2/p}(B_t) + \phi^{1-2/p}(B_t)\,\mathbb{E}_t[\|e_t\|^p] + \phi^{1-2/p}(B_t)\,\mathbb{E}_t[\|\tilde{e}_t\|^p]$$

We now seek a bound on the p-th moment of e_t. We note that

$$\mathbb{E}^{1/p}(\|e_t\|^p) = \mathbb{E}^{1/p}[\|\nabla f(x_t;\Omega_t) - h(x_t)\|^p] \tag{B.3}$$

$$\leq \mathbb{E}^{1/p}[\|\nabla f(x_t;\Omega_t)\|^p] + \sup_{x\in\mathcal{X}}\|h(x)\| \tag{B.4}$$

$$\leq \sup_{\omega\in\Omega_t}\mathbb{E}^{1/p}[M^p(\omega)] + C \tag{B.5}$$

An analogous result holds for $\tilde{e}_t$.

$$\mathbb{E}^{1/p}(\|\tilde{e}_t\|^p) \leq \mathbb{E}^{1/p}[M^p(Z)] + C \tag{B.6}$$

Defining v_{1t}

$$v_{1t} = 1 + \mathbb{E}_t[\|e_t\|^p] + \mathbb{E}_t[\|\tilde{e}_t\|^p] \tag{B.7}$$

So, v_{1t} has finite mean

$$\mathbb{E}[v_t] \leq 2\,\mathbb{E}^{1/p}[M^p(\omega^*)] + C \tag{B.8}$$

The bound in Lemma A.3 finishes the claim. □

Proof.(ii.) Using the definition of g_t found in the method section

$$\hat{g}_t - h(x_t) = \nabla f(x_t; \Omega_t) - h(x_t) \tag{B.9}$$

The sum of these variables satisfies condition of Lemma A.1, and so we conclude that

$$\mathbb{E}\,\|\nabla f(x_t; \Omega_t) - h(x_t)\|^2 \leq \frac{C}{B_t} \leq \frac{C}{B_t}(1 + \|x_t - x^*\|^2) \tag{B.10}$$

So we have shown the desired claim. □

Proof.[(iii.)] Analogously part (i.), we define $\tilde{\zeta}_t$ as follows.

$$\tilde{\zeta}_t = U_t \nabla f(x_{t-1}; \tilde{\Omega}_t) - \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \tilde{\Omega}_t)] \tag{B.11}$$

As mentioned earlier, $\{\tilde{\omega}\}_{s=1}^{\infty}$ is identical to the process in $(i.)$. So applying Lemma A.3 yields

$$|\,\mathbb{E}_t(\|\zeta_t\|^2) - \mathbb{E}_t(\|\tilde{\zeta}_t\|^2)| \leq m\phi(B_t) + \frac{\mathbb{E}_t[\|\zeta_t\|^p]}{m^{p-1}} + \frac{\mathbb{E}_t[\|\tilde{\zeta}_t\|^p]}{m^{p-1}} \tag{B.12}$$

In the same vein as (i.), we can bound $\mathbb{E}[\|\zeta_t\|^p]$ and by extension $\mathbb{E}[\|\tilde{\zeta}_t\|^p]$ by the constant C_p. If we set $m = \phi^{-2/p}(B_t)$ This leads to

$$|\,\mathbb{E}_t(\|\zeta_t\|^2) - \mathbb{E}_t(\|\tilde{\zeta}_t\|^2)| \leq \phi^{1-2/p}(B_t)[1 + \mathbb{E}_t[\|\zeta_t\|^p] + \mathbb{E}_t[\|\tilde{\zeta}_t\|^p]] \tag{B.13}$$

Expanding the terms in above we see that

$$\mathbb{E}_t\,\|\tilde{\zeta}_t\|^2 \leq 2(\nabla f(x_{t-1}; \Omega_t) - \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \Omega_t)] + \|e_t\|^2) \tag{B.14}$$

Combining statements (i.) and (ii.) yields the following inequality, where $v_{2t} = (1 + \mathbb{E}_t[\|\zeta_t\|^p] + \mathbb{E}_t[\|\tilde{\zeta}_t\|^p])$

$$\mathbb{E}_{t-1}(\|\zeta_t\|^2) \leq \phi^{1-\frac{2}{p}}(B_{t-1})v_{2t} + CB_t^{-1}(1 + \|x_{t-1} - t^*\|^2) \tag{B.15}$$

Here v_{2t} bounded above by our constant C. To show the desired result, we take $v_t = v_{1t} + v_{2t}$ □

Given we have established bounds for $e_t, h(x) and \zeta_t$ in terms of the energy function, we now extend the argument for the entire gradient term.

Lemma B.2 *We now show the following bound on the second moment of the gradient, where* $\Delta_t = x_t - x^*$.

$$\begin{aligned}\mathbb{E}_t[\|\gamma_t V_t\|^2] \leq (C\gamma_t^2\tilde{\sigma}_{\max}^2 + C\gamma_t^2\tilde{\sigma}_{\max}^2 B_t^{-1})\|\Delta_{t-1}\|^2 \\ + C\gamma_t^2\tilde{\sigma}_{\max}^2(1+B_t^{-1}) + 4\gamma_t^2\tilde{\sigma}_{\max}^2\phi^{1-2/p}(B_t)v_t + 4C\gamma_t\tilde{\sigma}_{\max}\sqrt{\phi^{1-2/p}(B_t)v_t} \\ + 2\gamma_t^2\tilde{\sigma}_{\max}^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{CB_t^{-1}(1+C^2)}\end{aligned}$$

Proof.

$$\begin{aligned}\|\gamma_t V_t\|^2 &= \|\gamma_t \mathbf{J}(\theta_t)^T[h(x_t) + e_t + \zeta_t]\|^2 \\ &= \gamma_t^2\|\mathbf{J}(\theta_t)^T h(x_t)\|^2 + \gamma_t^2\|\mathbf{J}(\theta_t)^T e_t\|^2 + \gamma_t^2\|\mathbf{J}(\theta_t)^T\zeta_t\|^2 \\ &+ 2\gamma_t h(x_t)^T[\mathbf{J}(\theta_t)\mathbf{J}(\theta_t)^T]^+ e_t + 2\gamma_t h(x_t)^T[\mathbf{J}(\theta_t)\mathbf{J}(\theta_t)^T]^+\zeta_t + 2\gamma_t e_t^T[\mathbf{J}(\theta_t)\mathbf{J}(\theta_t)^T]^+\zeta_t \\ &\leq \gamma_t^2\|\mathbf{J}(\theta_t)^T h(x_t)\|^2 + \gamma_t^2\|\mathbf{J}(\theta_t)^T e_t\|^2 + \gamma_t^2\|\mathbf{J}(\theta_t)^T\zeta_t\|^2 \\ &+ 2\gamma_t\|h(x_t)\mathbf{J}(\theta_t)^+\|\|\mathbf{J}(\theta_t)^+ e_t\| + 2\gamma_t\|h(x_t)\mathbf{J}(\theta_t)^+\|\|\mathbf{J}(\theta_t)^+\zeta_t\| + 2\gamma_t\|e_t\mathbf{J}(\theta_t)^+\|\|\mathbf{J}(\theta_t)^+\zeta_t\|\end{aligned}$$

We can then harness conditional expectation to yield that

$$\begin{aligned}\mathbb{E}_{t-1}[\|\gamma_t V_t\|^2] \leq \gamma_t^2\|\mathbf{J}(\theta_t)^+ h(x_t)\|^2 + \gamma_t^2\,\mathbb{E}_{t-1}\,\|\mathbf{J}(\theta_t)^+ e_t\|^2 + \gamma_t^2\,\mathbb{E}_{t-1}\,\|\mathbf{J}(\theta_t)^+\zeta_t\|^2 \\ + 2\gamma_t^2\|\mathbf{J}(\theta_t)^+ h(x_t)\|\mathbb{E}_{t-1}\|\mathbf{J}(\theta_t)^+ e_t\| + 2\gamma_t^2\,\mathbb{E}_{t-1}[\|\mathbf{J}(\theta_t)^+ e_t\|\|\mathbf{J}(\theta_t)^+\zeta_t\|]\end{aligned}$$

We can then use the rates from Lemma A.3 in order to bound the value of $\|\Delta_t\|^2$. In addition, because $\mathbf{J}(\theta_t)^+$ has a bounded spectra, we can write that $\|\mathbf{J}(\theta_t)^+ v\| \leq \tilde{\sigma}_{\max}\|v\|$, where $\tilde{\sigma}_{\max}$ is the largest singular value of the Moore-Penrose pseudo-inverse of the Jacobian, or the inverse of the smallest positive singular value.

$$\begin{aligned}\mathbb{E}_{t-1}[\|\gamma_t V_t\|^2] &\leq \gamma_t^2\tilde{\sigma}_{\max}^2\cdot C + 2\gamma_t^2\tilde{\sigma}_{\max}^2\phi^{1-2/p}(B_t)v_t + \frac{C\tilde{\sigma}_{\max}^2\gamma_t^2}{B_t}(1+\|\Delta_t\|^2) \\ &+ 2C\gamma_t\tilde{\sigma}_{\max}\sqrt{\phi^{1-2/p}(B_t)v_t} + 2C\gamma_t^2\tilde{\sigma}_{\max}^2\sqrt{\phi^{1-2/p}(B_t)v_t} \\ &+ 2\gamma_t^2\tilde{\sigma}_{\max}^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{\phi^{1-2/p}(B_t)v_t + \frac{C}{B_t}(1+\|\Delta_t\|^2)} \\ &\leq (1 + C\gamma_t^2\tilde{\sigma}_{\max}^2 + C\gamma_t^2\tilde{\sigma}_{\max}^2 B_t^{-1})\|\Delta_t\|^2 \\ &+ C\gamma_t^2\tilde{\sigma}_{\max}^2(1+B_t^{-1}) + 4\gamma_t^2\tilde{\sigma}_{\max}^2\phi^{1-2/p}(B_t)v_t + 4C\gamma_t\tilde{\sigma}_{\max}\sqrt{\phi^{1-2/p}(B_t)v_t} \\ &+ 2\gamma_t^2\tilde{\sigma}_{\max}^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{CB_t^{-1}(1+C^2)}\end{aligned}$$

□

B.2 Consistency Theorem:

We now satisfy the first requirement of an estimator to conduct robust estimation, consistency. Given the bound generated in the previous parts, and the rate limits established in Model 1, we show the convergence of the energy function to zero through a Lyapunov argument and the template inequality. Consequently, the estimator built off of the latent iterates $\bar{x}_T = T^{-1}\sum_{t=1}^{T} x_t$ is also asymptotically consistent. To extend this to the control space, we show a duality gap in Lemma B.4, that relates how far the parameter values get stretched to the singular values of the Jacobian.

Theorem B.1 (Robbins-Siegmund) *If $(V_t)_{t\geq 1} = V(X_t)_{t\geq 1}, (\psi_t)_{t\geq 1}, (\alpha_t)_{t\geq 1}, (U_t)_{t\geq 1}$ be four nonnegative $(\mathcal{F}_t)_{t\geq 1}$-adapted processes such that*

$$\sum_{t\geq 1} \psi_t \leq \infty, \sup_{\omega\in\Omega} \prod_{n\geq 1} (1+\alpha_n(\omega)) \leq \infty$$

Then if $\forall n \in N$

$$\mathbb{E}[V_t|F_{t-1}] \leq V_{t-1}(1+\alpha_{t-1}) + \psi_{t-1} - U_{t-1}$$

Then we claim that $V_n \xrightarrow{a.s.} V_\infty$, $\sup_{n\geq 0}\mathbb{E}[V_n] < \infty$.

In the proceeding section, the function $V(\cdot)$ will be "Lyapunov". If the algorithm indeed satsifies the above inequality, then with a suitable function $V(\cdot)$, the algorithm itself can be shown to be convergent.

As shown earlier in part A.1, the elements e_T and ζ_T are martingale terms. Thus when analyzing the sequence in question,

$$\theta_{T+1} = \theta_T - \gamma_{T+1}\hat{g}_t = \theta_T - \gamma_{T+1}(h(\theta_T) + e_{T+1} + \zeta_{T+1})$$

We can apply the Robbins-Siegmund theorem to the previous martingale sequence to derive the following.

Corollary B.1 *Let θ_n be defined in the sequence above, and let $V(\cdot)$ be a Lyapunov function. Then if $\mathbb{E}[\|\hat{g}\|^2|\mathcal{F}_{n-1}] \leq C\gamma_n^2(1+V(\theta_{n-1}))$ then*

$$\hat{\theta}_n - \hat{\theta}_{n-1} \xrightarrow{a.s.} 0$$

Proof. This proof is an adaptation of Theorem 5.3 in [Kushner and Yin, 2003]. □

We next show a useful lemma relating the latent and control spaces.

Lemma 1 (Representation Gap) *If the assumptions on the singular values in part three hold, then the following inequality holds, where $\sigma_{\min}, \sigma_{\max}$ are the smallest and largest singular values respectively.*

$$\sigma_{\min}||\theta - \theta^*|| \leq ||\chi(\theta) - \chi(\theta^*)|| \leq \sigma_{\max}||\theta - \theta^*||.$$

Proof.

We first prove the upper bound. This amounts to showing that our function χ is Lipschitz. Recall that by the definition of the Jacobian, given a vector v with norm equal to one, we have that

$$\lim_{t\to 0} \frac{|\chi(\theta) - \chi(\theta + tv)|}{t} = \mathbf{J}(\theta)(v) \tag{B.16}$$

So for all ε, there exists a δ such that if $t < \delta$, we have that

$$||\frac{|\chi(\theta) - \chi(\theta + tv)|}{t} - \mathbf{J}(\theta)(v)|| < \varepsilon \tag{B.17}$$

Then using the fact that $|||a|| - ||b||| < ||a - b||$, we claim that

$$-\varepsilon + \sigma_{\min} \leq -\varepsilon + ||\mathbf{J}(\theta)(v)|| < ||\frac{|\chi(\theta) - \chi(\theta + tv)|}{t}|| < \varepsilon + ||\mathbf{J}(\theta)(v)|| \leq \varepsilon + \sigma_{\max} \tag{B.18}$$

The second inequality is due to the fact that our Jacobian has a bounded spectra. We note that this holds true for all v on the unit ball. Thus if we replace the value $t \cdot v$ with $\theta^* - \theta$ where $||\theta - \theta^*|| < \delta$, we can conclude that

$$-\varepsilon + \sigma_{\min} < \frac{||\chi(\theta) - \chi(\theta^*)||}{||\theta - \theta^*||} < \varepsilon + \sigma_{\max} \tag{B.19}$$

Thus we have shown our function is bi-Lipschitz within the unit ball. To extend this result, for any given pair of points x, y we construct a set of unit balls intersecting on the edges and use the triangle inequality, tending ε towards zero to show our function is $\sigma_{\max}$-Lipschitz. Multiplying the denominator of the fraction yields the desired result. □

Theorem 1 (Consistency) *Given the assumptions in part 2, we conclude that the SHADE estimator satisfies*

$$\hat{\theta}_{SHADE} \xrightarrow{a.s.} \theta^*.$$

Proof. We define the quantity $\Delta_t = \bar{x}_t - x^*$, which serves as a precursor to a Lyapunov function $\|\Delta_t\|^2$ is very similar to the $E(\theta; x^*)$ found in Vlatakis-Gkaragkounis et al. We begin by utilizing Lemma A.2.

$$E_{t+1} \leq E_t - \gamma_t \langle h(x_t), x_t - \tilde{x} \rangle + \gamma_t \alpha_t + \gamma_t^2 \psi_t \tag{Lemma A.3}$$

Note because F is a strongly monotone operator (Assumption 2), we claim that $\gamma_t \langle h(x_t), x_t - \tilde{x} \rangle \geq \mu E_t$. Likewise, from lemma A.3, $\mathbb{E}[\alpha_t]$ is equal to zero. Lastly, via the rate limitations found in both assumptions one and two, $\sum_{t \geq 1} \gamma_t^2 < \infty$ and $\sum_{t \geq 1} \phi^{1-2/p}(B_t) < \infty$. So

$$E_{t+1} \leq E_t - \gamma_t \langle h(x_t), x_t - \tilde{x} \rangle + \gamma_t \alpha_t + \gamma_t^2 \psi_t \leq E_t + \gamma_t \alpha_t + \kappa \gamma_t^2 \psi_t \tag{B.20}$$

We have that

$$\begin{aligned}
\mathbb{E}_t[E_{t+1}] &\leq \mathbb{E}_t[E_t - \gamma_t \mu E_t + \gamma_t \alpha_t + \gamma_t^2 \psi_t] \\
&= E_t + \gamma_t^2 \, \mathbb{E}_t[\psi_t] - \gamma_t \mu E_t \\
&\leq \sum_{i=1}^{n} (1 + C\gamma_t^2 \tilde{\sigma}_{\max}^2 + C\gamma_t^2 \tilde{\sigma}_{\max}^2 B_t^{-1}) \|x_t - x^*\|^2 \\
&+ C\gamma_t^2 \tilde{\sigma}_{\max}^2 (1 + B_t^{-1}) + 4\gamma_t^2 \tilde{\sigma}_{\max}^2 \phi^{1-2/p}(B_t) v_t + 4C\gamma_t \tilde{\sigma}_{\max} \sqrt{\phi^{1-2/p}(B_t) v_t} \\
&+ 2\gamma_t^2 \tilde{\sigma}_{\max}^2 \sqrt{\phi^{1-2/p}(B_t) v_t} \sqrt{C B_t^{-1} (1 + C^2)} - \gamma_t \mu \|x_t - x^*\|^2
\end{aligned}$$

We then note that because $\sum_{t \geq 1} \gamma_t^2 < \infty$ and that $\sum_{t \geq 1} \phi^{1-2/p} < \infty$, the non-energy parts of the expression sum over all times t to a finite value. In addition, we have that

$$\sum_{t \geq 1} (C\gamma_t^2 \tilde{\sigma}_{\max}^2 + C\gamma_t^2 \tilde{\sigma}_{\max}^2 B_t^{-1}) < \infty \tag{B.21}$$

Thus we can apply the Robbins-Siegmund theorem to show that E_t converges to zero as t tends to infinity. To show convergence of θ note that from Lemma B.3

$$\sigma_{\min}\|\theta_t - \theta^*\| \leq \|\Delta_t\| \leq \sigma_{\max}\|\theta_t - \theta^*\| \tag{B.22}$$

Thus this implies that $\|\theta_t - \theta^*\|^2$ goes to zero almost surely and the result for the SHADE estimator follows swiftly. □

C Proof of Asymptotic Normality

For convenience, we restate Theorem 2, and Theorem 3 below. For ease of notation, we define G to be the Hessian matrix at the optimal latent state x^*, $\nabla^2 f(x^*)$.

Theorem 2 (Asymptotic Normality of SHADE) *Given the assumptions above, the following correspondence occurs*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^{T} B_t^{-1}}}(\hat{\theta}_{SHADE} - \theta^*) \to \mathcal{N}(0, \mathbf{J}(\theta^*)^{+}\Sigma[\mathbf{J}(\theta^*)^{+}]^T)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, $G = \nabla^2 f(x^*)$ *and* $\mathbf{J}(\theta^*)$ *is the Jacobian evaluated at the optimal* θ *value.*

Theorem 3 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

$$\frac{T}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}(\hat{\theta}_T^* - \hat{\theta}_{SHADE})|\mathcal{D} \xrightarrow{\mathbb{L}} \mathcal{N}(0, \hat{\Sigma})$$

where $\mathcal{D} = \{\omega_i | i \in I_t \cup J_t\}$ *and represents the data used in the empirical risk minimization process, and* $\hat{\Sigma} = \mathbf{J}(\theta^*)^{+}G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}[\mathbf{J}(\theta^*)^{+}]^T$ *is the covariance matrix in Theorem 2*

This section establishes the second criterion needed to reliably use empirical risk minimization, that of asymptotic normality. Here we leverage the strongly convex geometry of the latent space to simplify analysis. We show that asymptotic normality in the latent space implies normality in the control space as well via the use of the delta method. Additionally, we analyse the behaviour of bootstrap PHGD, and show that its asymptotic distribution is identical to PHGD. To this end, we use the following arguments.

1. In Lemma 2, we establish an inequality characterizing the evolution of the latent iterates over time.
2. In Lemma C.1-C.2 we equate the time-averaged parameter estimates with scaled gradients of the loss function. This is similar to the strategy undertaken by Polyak [1990a] and Liu et al. [2023]. To achieve this, we exploit the strongly convex nature of the objective in the latent space to create sharp bounds.
3. In Lemma C.3, we show that the sum of the autocorrelation coefficients is finite and conclude via the Lindeberg condition that the sum of gradients exhibits asymptotic normality.
4. In Theorem 3, the bootstrap estimates are shown to match the distribution of PHGD. Leveraging the requirements in Assumption 3 and a theorem from Kuczmaszewka, we can apply Theorem 2 to achieve the desired result.

C.1 Latent Descent Inequality: Similar to the equation governing stochastic gradient descent, we derive the following for the latent iterates.

Lemma 2 (Descent Equality) *Let $x_t = \chi(\theta_t)$, we then have that for iterates of PHGD*

$$x_{t+1} = x_t - \gamma_t U_t \nabla f(x_t; \Omega) + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2).$$

Proof. We show this lemma via Taylor's theorem. Recall that

$$\begin{aligned}
x_{t+1} &= \chi(\theta_{t+1}) \\
&= \chi(\theta_t - \gamma_t U_t \mathbf{P}_t V_t) \\
&= \chi(\theta_t) - \gamma_t \mathbf{J}_{\theta_t} \mathbf{P}_t U_t V_t + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= \chi(\theta_t) - \gamma_t \mathbf{J}_{\theta_t} \mathbf{P}_t \mathbf{J}_{\theta_t}^T U_t \nabla f(x_t; \Omega_{t-1}) + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= x_t - \gamma_t U_t \hat{g}_t + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2)
\end{aligned}$$

So the desired claim has been shown. □

C.2 Establishment of Asymptotic Normality:

So far our analysis has dwelt on bounding terms of the martingale loss gradient. To proceed, we establish a correspondence between these terms and the parameter estimates through lemmas found in Polyak [1990a]. By unrolling the iteration sequence into a product of terms dictated by the hessian, we extend our gradient bounds to the iteration parameters themselves.

We begin with a few key propositions regarding the evolution of the descent process.

Proposition C.1 *Let G be a positive definite matrix, and define the following.*

$$D_j^j = I$$

$$D_j^t = (I - \gamma_{t-1}G)D_j^{t-1} = ... = \prod_{k=j}^{t-1}(I - \gamma_k G)$$

$$\bar{D}_j^t = \gamma_j \sum_{i=j}^{t-1} D_j^i$$

Then we have that

(i) There are constants $C > 0$ such that $\|\bar{D}_j^t\| \leq C$

(ii) $\lim_{t\to\infty} \frac{1}{t}\|\bar{D}_j^t - G^{-1}\| = 0$

(iii) $\|D_j^t\| \leq \exp(\lambda_G \sum_{k=j}^{t-1}(k+\gamma)^{-\rho})$, *where λ_G is the largest eigenvalue of G.*

(iv) Let $\{a_j\}_{j=0}^{\infty}$ be a positive and non-increasing sequence, such that $\sum_{j\geq 0} alpha_j = \infty$, and $t^\rho / \sum_{j=1}^t a_j \to 0$, then $\lim_{t\to\infty} \sum_{j=1}^t a_j \|\bar{D}_j^t - G^{-1}\|/(\sum_{j=1}^t a_j) = 0$

Proof. The proof for (i) and (ii) can be found in Polyak and Juditsky [1992a], (iii) can be found in Chen et al. [2020] and (iv) is in Liu et al. [2023]. □

Lemma C.1 *Under assumptions 1-2 and Model 1, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(\zeta_j) = \frac{1}{T}\sum_{t=1}^{T} G^{-1}U_t(\nabla f(x^*;\Omega_t)) + R_n$$

where $\mathbb{E}\|R_n\|^2 = O(\frac{\sum_{t=1}^T B_j^{-1}}{T^2})$

Proof. This is Lemma S.21 in Liu et al. [2023]. □

With these lemmas, we can craft the following correspondence.

Lemma C.2 *We have the following correspondence.*

$$\frac{1}{T}\sum_{t=1}^{T} x_t - x^* = \frac{1}{T}\sum_{i=1}^{T} G^{-1}U_t \nabla f(x^*;\Omega_t) + o_P(\sqrt{\frac{\sum_{t=1}^T B_t^{-1}}{T}})$$

Proof. We being by rearranging the gradient term.

$$x_{t+1} = x_t - \gamma_t[h(x_t) + e_t + \zeta_t] + \gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \tag{C.1}$$

We now aim to separate this into terms that can be bounded. If we define $\Delta_t = (x_t - x^*)$ we can see that

$$\begin{aligned}
\Delta_{t+1} &= \Delta_t - \gamma_t[h(x_t) - e_t - \zeta_t] + \gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= \Delta_t - \gamma_t G\Delta_t - \gamma_t(e_t + \zeta_t) - \gamma_t(h(x_t) - G\Delta_t) + \gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= (I - \gamma_t G)\Delta_t - \gamma_t(e_t + \zeta_t) - \gamma_t(h(x_t) - G\Delta_t) + \gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= [\prod_{j=1}^{t}(I - \gamma_j)]\Delta_0 + \sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(e_j + \zeta_j) \\
&+ \sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(h(x_t) - G\Delta_t) + \gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2)
\end{aligned}$$

Thus we now look at the average value of Δ_t. We see that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\Delta_t &= \frac{1}{T}\sum_{t=1}^{T}[\prod_{j=1}^{t}(I - \gamma_j)]\Delta_0 + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(e_j + \zeta_j) \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(h(x_t) - G\Delta_t) + \frac{1}{T}\sum_{t=1}^{T}\gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= \frac{1}{T}\sum_{t=1}^{T}[\prod_{j=1}^{t}(I - \gamma_j)]\Delta_0 + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(e_j) \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(\zeta_j) \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_t(h(x_t) - G\Delta_t) + \frac{1}{T}\sum_{t=1}^{T}\gamma_t^2 O(\|\theta_t - \theta_{t-1}\|^2) \\
&= S_1 + S_2 + S_3 + S_4 + S_5
\end{aligned}$$

We see from Lemma 2that S_1 is asymptotically $o_P(\frac{\sqrt{\sum_{j=1}^{T} B_j^{-1}}}{T})$. We now aim to bound S_2. We define $Q_j^T = \sum_{t=j}^{T}[\prod_{k=j+1}^{t}(I - \gamma_k G)]\gamma_j$. Using Lemma 2 again, we have that $\|Q_j^T\| < C$. Then applying Lemma B.1, we have that

$$\mathbb{E}(\|S_2\|) \leq \frac{1}{T}\sum_{j=1}^{T}\|Q_j^T\|\,\mathbb{E}[\|e_t\|] \leq \frac{C}{T}\sum_{j=1}^{T}\phi^{1/2-1/p}(B_j) = o(\frac{\sqrt{\sum_{j=1}^{T} B_j^{-1}}}{T}) \tag{C.2}$$

If we use Taylor expansion and the fact that the Jacobian has bounded spectra, we have that

$$\|h(x_t) - G\Delta_t\| = \|h(x^*) + G(x_t - x^*) - G(x_t - x^*) + O(\|x_t - x^*\|^2)\| \leq C \cdot \|x_t - x^*\|^2 \tag{C.3}$$

Thus armed with this we can

$$\begin{aligned}\mathbb{E}[S_4] &\leq \frac{1}{T}\sum_{\{j:\|\Delta_{j-1}\|^2\leq\delta\|\}}\mathbb{E}[\|\Delta_{j-1}\|^2] \\ &+ \frac{1}{T}\sum_{\{j:\|\Delta_{j-1}\|^2\leq\delta\|\}}\sum_{t=j}^{T}[\prod_{k=j+1}^{t}(I-\gamma_k G)]\gamma_t\,\mathbb{E}\,\|h(x_t)-G\Delta_t\| \\ &\leq \frac{\sum_{t=1}^{T}\gamma_t}{T}+\frac{1}{T} \\ &= o(\frac{\sqrt{\sum_{j=1}^{T}B_j^{-1}}}{T})\end{aligned}$$

This follows because $\Delta_t \to 0$ almost surely, thus the set of all j such that $\|\Delta_j\|^2 < \delta$ is finite almost surely. Thus the second term of the sequence will be dominated by the $\frac{1}{T}$ term.

Lastly looking at S_5 we see that because $\theta_t \to \theta^*$ almost surely, we have that this term vanishes on order $\frac{1}{T}$. Thus using lemma B.3, we conclude that

$$\frac{1}{T}\sum_{t=1}x_t - x^* = \frac{1}{T}\sum_{t=1}^{T}G^{-1}\,\nabla f(x^*;\Omega_t) + o_P(\sqrt{\frac{\sum_{t=1}^{T}B_t^{-1}}{T}}).$$

□

We now lay out asymptotic normality claims. Throughout this section, we will refer to $r(t)$ the autocorrelation coefficient defined as follows.

Definition 2 (Autocorrelation Coefficients) *Let $L(\theta;\Omega)$ be a stochastic loss function satisfying the assumptions laid out in Assumption 1. Then we define*

$$r(t) = \mathbb{E}[\nabla L(\theta_{k+t};\Omega_{k+t})\nabla L(\theta_k;\Omega_k)^T]$$

We now recall a result from Fan and Yao [2003], which ensures the autocorrelation coefficients do not grow too large in a $\phi-$mixing sequence.

Lemma C.3 *Under the assumptions in part 3, we conclude that if $r(t)$ is defined as above. Then*

$$\sum_{t\geq 1}\|r(t)\| < \infty.$$

Proof. In Theorem 2.20 of Fan and Yao [2003], the authors claim the sum of the autocorrelation coefficients $r(t)$ is bounded throughout time for an $\alpha-$mixing (strong mixing) sequence. Using the relationship between $\phi-$mixing and strong mixing sequences found in Bradley [2005], we can conclude the desired claim. □

We show asymptotic normality through use of the Lindeberg condition.

Theorem C.1 (Lindeberg Condition) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_k : \Omega \to \mathbb{R}^n, k \in \mathbb{N}$ be independent random variables. Suppose $\mathbb{E}[X_k] = \mu_k$ and $\text{Var}(X_k) = \sigma_k^2 < \infty$. Then if $s_n^2 = \sum_{k=1}^n \sigma_n^2$ and the sequence $\{X_k\}_{k=1}^\infty$ satisfies*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}[(X_k - \mu_k)^2 \cdot \mathbf{1}\{|X_k - \mu_k| > \epsilon s_n\}]$$

Then

$$Y_n = \frac{\sum_{k=1}^n (X_k - \mu_k)}{s_n} \to \mathcal{N}(0, 1)$$

Proof. The proof can be found in Brown [1971]. □

With this key tool, we now outline our proof of asymptotic normality.

Lemma C.4 *Given the assumptions in part 3, we claim that*

$$\frac{1}{\sum_{t=1}^T B_t^{-1}} \sum_{t=1}^{T} \hat{g}_t^{i*} + \hat{g}_t^{j*} \to \mathcal{N}(0, V).$$

when $V = 2r(0) + 4\sum_{k\geq 1} r(k)$

Proof. We only prove this fact for the one-dimensional case. A multivariate extension can be made via a Cramer-Wold device. For ease of notation, we have that $\nabla f(x^*; \Omega_t^i) = \hat{g}_t^{i*}, \nabla f(x^*; \Omega_t^j) = \hat{g}_t^{j*}$ We see that the second moment of this expression can be expressed as

$$\begin{aligned}
\mathbb{E}[\| \sum_{t=1}^{T} \hat{g}_t^{i*} + \hat{g}_t^{j*} \|^2] = & \\
&= \sum_{t=1}^{T} \mathbb{E}[\|\hat{g}_t^{i*} + \hat{g}_t^{j*}\|^2] + 2\sum_{t=1}^{t-1} \mathbb{E}[\|(\hat{g}_t^{i*} + \hat{g}_t^{j*}) \cdot \hat{g}_{t+1}^{i*} + \hat{g}_{t+1}^{j*}\|^2] \\
&+ 2\sum_{t=1}^{T-2} \sum_{k=t+2}^{T} \mathbb{E}[\|(\hat{g}_t^{i*} + \hat{g}_t^{j*}) \cdot (\hat{g}_k^{i*} + \hat{g}_k^{j*})\|^2] \\
&:= S_1 + S_2 + S_3
\end{aligned}$$

We now aim to bound these expressions via our assumptions on the $\phi-$mixing nature of these sequences. When $t > k + 1$, the index differences between I_t and J_k are at least B_{t-1} apart. Thus we can use lemma 4 to see that

$$\begin{aligned}
\mathbb{E}[\hat{g}_t^* \cdot \hat{g}_k^*] &\leq C_p \phi^{1-2/p}(B_{t-1})\, \mathbb{E}^{2/p}[|\hat{g}_t^* \cdot \hat{g}_k^*|^{p/2}] \\
&\leq C_p \phi^{1-2/p}(B_{t-1})\, \mathbb{E}^{1/p}[|\hat{g}_t^*|]\, \mathbb{E}^{1/p}[\hat{g}_k^*] \\
&\leq C_p \phi^{1-2/p}(B_{t-1}) B_t^{-1/2} B_k^{-1/2}
\end{aligned}$$

The last inequality results from lemma 3, bounding the moment of the gradient sequence. An analogous result holds for when $k > t + 1$

$$\mathbb{E}[\hat{g}_t^* \cdot \hat{g}_k^*] \leq C_p \phi^{1-2/p}(B_{t-1}) B_t^{-1/2} B_k^{-1/2}$$

Combining the above, we conclude.

$$S_3 \leq 2C \sum_{t=1}^{T-2} \sum_{k=t+2}^{T} C_p \phi^{1-2/p}(B_{t-1}) B_t^{-1/2} B_k^{-1/2} \leq 2C \sum_{t=1}^{T-2} B_t^{-1} \sum_{k=t+2}^{T} \phi^{1-2/p}(B_{k-1}) \tag{C.4}$$

Thus using Lemma S.7 in Liu et al. (about convergence of fractions of sequences), we see that this term is $o_P(\sum_{t=1}^{T} B_t^{-1})$.

We now define $v_s = \nabla f(x^*; \omega_s)$, $\mathcal{D}_t = I_t \cup J_t$ and $r(t) = \mathbb{E}[\nabla f(x^*; \omega_s) \nabla f(x^*; \omega_s)]$. Thus we see that

$$\begin{aligned}
2\sum_{t=1}^{t-1} \mathbb{E}[\|\hat{g}_t^* \cdot \hat{g}_{t+1}^*\|^2] &= 2\sum_{t=1}^{t-1} \frac{1}{B_t B_{t+1}} \mathbb{E}[(\sum_{s \in I_t} v_s)(\sum_{k \in S_{t+1}} v_k)] \\
&= 2\sum_{t=1}^{t-1} \frac{1}{B_t B_{t+1}} \sum_{s \in I_t} \sum_{k \in S_{t+1}} \mathbb{E}[v_s v_k] \\
&= 2\sum_{t=1}^{t-1} \frac{1}{B_t B_{t+1}} \sum_{s=0}^{2B_t - 1} \sum_{k=1}^{2B_{t+1}} r(s+k) \\
&= 2\sum_{t=1}^{t-1} \frac{1}{B_t B_{t+1}} \sum_{k=1}^{2B_{t+1}} \sum_{m=k}^{k+2B_t-1} r(m)
\end{aligned}$$

Thus by because $\lim_{t\to\infty} \sum_{k=1}^{t} \|r(k)\| < \infty$, we have that asymptotically,

$$\lim_{t\to\infty} \frac{1}{2B_{t+1}} \sum_{k=1}^{2B_{t+1}} \sum_{m=k}^{\infty} \|r(m)\| = 0 \tag{C.5}$$

Then we see via Liu lemma 6 that this term is asymptotically $o(\sum_{t\geq 1} B_t^{-1})$

In a similar vein we look at the term S_1. This term can also be represented as the following

$$\begin{aligned}
\sum_{t=1}^{t-1} \mathbb{E}[\|\hat{g}_t^* \cdot \hat{g}_k^*\|^2] &= \sum_{t=1}^{T} \frac{1}{B_t^2} \mathbb{E}[(\sum_{s \in I_t} v_s)(\sum_{k \in S_t} v_k)] \\
&= \sum_{t=1}^{T} \frac{1}{B_t^2} \sum_{s \in I_t} \sum_{k \in S_t} \mathbb{E}[v_s v_k] \\
&= \sum_{t=1}^{T} \frac{1}{B_t^2} \sum_{s \in I_t} \sum_{k \in S_t} \mathbb{E}[v_s v_k] \\
&= \sum_{t=1}^{T} \frac{1}{B_t^2} (2B_t r(0) + 2\sum_{k=1}^{2B_t} (2B_t - k) r(k)) \\
&= \sum_{t=1}^{T} \frac{2}{B_t} (r(0) + 2\sum_{k=1}^{2B_t} (1 - \frac{k}{2B_t}) r(k)) \\
&:= \sum_{t=1}^{T} \frac{2}{B_t} \beta_t
\end{aligned}$$

Thus because we have that $\sum_{k\geq 0} \|r(k)\| < \infty$. we can apply the dominated convergence theorem to claim that $\lim_{t\to\infty} \beta_t = r_0 + 2\sum_{k\geq 0} r(k)$. Thus putting this all together we can see that

$$\lim_{T\to\infty} \frac{S_1}{\sum_{t=1}^T B_t^{-1}} = \lim_{T\to\infty} 2\frac{\sum_{t=1}^T B_t^{-1}\beta_t}{\sum_{t=1}^T B_t^{-1}} = 2r(0) + 4\sum_{k\geq 1} r(k) \tag{C.6}$$

We now gear up to prove the normality theorem. Define the quantity $V_{T,t} = (\hat{g}_t^{i*} + \hat{g}_t^{j*})/\sqrt{\sum_{k=1}^T B_k^{-1}}$. We have just shown that

$$v_T^2 := \mathbb{E}[|\sum_{t=1}^T V_{T,t}|^2] \to 2r(0) + 4\sum_{k\geq 1} r(k) \tag{C.7}$$

Thus we know that $v_T^2 \geq c^2$ for some value $c > 0$.

$$\mathbb{E}[|V_{T,t}|^2 \mathbf{1}\{|V_{T,t}| > \varepsilon v_T\}] \leq \frac{(\varepsilon v_T)^2\, \mathbb{E}[|V_{T,t}|^p]}{(\varepsilon v_T)^p} \tag{C.8}$$

We get this via Markov's inequality. We then apply lemma three to arrive at

$$\frac{(\varepsilon v_T)^2\, \mathbb{E}[|V_{T,t}|^p]}{(\varepsilon v_T)^p} \leq \frac{C_p B_t^{-p/2}}{(c\varepsilon)^{p-2}(\sum_{j=1}^t B_j^{-1})} \tag{C.9}$$

Thus combining this all together we satisfy the Lindeberg condition.

$$\frac{1}{v_T^2}\sum_{t=1}^T \mathbb{E}[|V_{T,t}|^2 \mathbf{1}\{|V_{T,t}| > \varepsilon v_T\}] \leq \frac{C\sum_{t=1}^T B_t^{-p/2}}{\sum_{t=1}^T B_t^{-1}} \to 0 \tag{C.10}$$

The final equality is due to lemma. □

To establish Theorem 2, we recall a useful lemma.

Lemma C.5 *If $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ and $f : \mathbb{R}^d \to \mathbb{R}^k$, then via Taylor expansion we claim*

$$\sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, \text{Jac}_f(\mu)\Sigma\, \text{Jac}_f(\mu)^T)$$

Proof. This proof can be shown using the Central Limit theorem and Taylor's theorem. A full proof can be found [Keener, 2010]. □

Theorem 2 (Asymptotic Normality of SHADE) *Given the assumptions above, the following correspondence occurs*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^T B_t^{-1}}}(\hat{\theta}_{SHADE} - \theta^*) \to \mathcal{N}(0, \mathbf{J}(\theta^*)^+ \Sigma [\mathbf{J}(\theta^*)^+]^T)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, $G = \nabla^2 f(x^*)$ and $\mathbf{J}(\theta^*)$ is the Jacobian evaluated at the optimal θ value.

Proof. Given lemma B.4, we have that

$$T^{-1}\sum_{t=1}^{T} x_t^i - x^* = \frac{1}{T}\sum_{t=1}^{T} G^{-1}\hat{g}_t^{i*} + o_P(\sqrt{\frac{\sum_{t=1}^{T} B_t^{-1}}{T}}) \tag{C.11}$$

$$T^{-1}\sum_{t=1}^{T} x_t^j - x^* = \frac{1}{T}\sum_{t=1}^{T} G^{-1}\hat{g}_t^{j*} + o_P(\sqrt{\frac{\sum_{t=1}^{T} B_t^{-1}}{T}}) \tag{C.12}$$

Then applying Lemma C.4 we can conclude.

$$\frac{1}{2}T^{-1}(\sum_{t=1}^{T} x_t^i + x_t^j) - x^* \to \mathcal{N}(0, \Sigma)$$

Applying the delta method results in the desired claim. □

Proof. This proof can be shown using the Central Limit theorem and Taylor's theorem. A full proof can be found [Keener, 2010]. □

The following corollary can be directly claimed from these results.

C.3 Bootstrap Normality We conclude with an asymptotic normality proof for the bootstrap iterates. Formally we claim,

Theorem 3 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

$$\frac{T}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}(\hat{\theta}_T^* - \hat{\theta}_{SHADE})|\mathcal{D} \xrightarrow{\mathbb{L}} \mathcal{N}(0, \hat{\Sigma})$$

where $\mathcal{D} = \{\omega_i | i \in I_t \cup J_t\}$ *and represents the data used in the empirical risk minimization process, and* $\hat{\Sigma} = \mathbf{J}(\theta^*)^+ G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}[\mathbf{J}(\theta^*)^+]^T$ *is the covariance matrix in Theorem 2*

Proof. In this proof, we follow the outline of Theorem 3 from Liu et al. [2023]. Define $v_t = \hat{g}_t^{*i} + \hat{g}_t^{*j}$, and $\chi(\hat{\theta}_{SHADE}) = \bar{x}_T^*$. Then we claim that

$$\begin{aligned}\frac{T}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}(\bar{x}_T^* - \bar{x}_T) &= \frac{1}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}\sum_{t=1}^{T}\frac{1}{2}(U_t - 1)G^{-1}(\hat{g}_t^{*i} + \hat{g}_t^{*j}) + o_P(1) \\ &= \frac{1}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}\sum_{t=1}^{T}\frac{1}{2}(U_t - 1)G^{-1}v_t + o_P(1)\end{aligned}$$

To show asymptotic normality we observe the limiting behavior of $Y_T := \sum_{t=1}^{T}(U_t - 1)v_t/\sqrt{\sum_{t=1}^{T} B_t^{-1}}$. Define B to be the unit ball in $\mathbb{R}^d$, i.e. $\{w \in \mathbb{R}^d : \|w\| = 1\}$. Because the U_t random variables have unit mean and unit variance, we verify that

$$\mathbb{E}[|\beta^T Y_T|^2] = \beta^T\left(\frac{T}{\sum_{t=1}^{T} B_t^{-1}}\sum_{t=1}^{T} v_t v_T^T\right)\beta$$

Moreover, we cap the variance of $v_t v_t^T$ with Lemma 2 and assumption 3, we see that

$$\mathbb{E}[\|v_t v_t^T - \mathbb{E}(v_t v_t^T)\|^2] \leq \mathbb{E}[\|v_t\|^4] \leq C B_t^{-2}$$

as the sequence v_t is $\phi-$mixing. Furthermore, using the rate-limiting assumptions in assumption 3 we see that

$$\sum_{t=1}^{T} \frac{\mathbb{E}\,\|v_t v_t^T - \mathbb{E}(v_t v_T^T)\|^2}{(\sum_{j=1}^{t} B_j^{-1})^2} \leq C \sum_{t=1}^{\infty} \frac{B_t^{-2}}{(\sum_{j=1}^{t} B_j^{-1})^2} \lesssim \sum_{t=1}^{\infty} B_t^{-2} t^{-2\rho} < \infty.$$

Then using Corollary 1 from Kuczmaszewka we see that this value is consistent.

$$\frac{1}{\sum_{t=1}^{T} B_t^{-1}} \sum_{t=1}^{T} [v_t v_t^T - \mathbb{E}(v_t v_t^T)] \xrightarrow{a.s.} 0$$

So using the properties of strong convexity, the sample covariance matrix converges almost surely to the true covariance.

$$V_T := \frac{1}{\sum_{t=1}^{T} B_t^{-1}} \sum_{t=1}^{T} \mathbb{E}(v_t v_t^T) \to 2r(0) + 4 \sum_{k=1}^{\infty} r(k) := V$$

Thus we can conclude that $\beta^T V_T \beta$ converges uniformly to $\beta^T V \beta$ for all $\beta \in B$.

Likewise, looking at the Lindeberg condition, we can see

$$g_T(\beta) := \frac{1}{\beta^T V_T \beta \sum_{j=1}^{T} B_j^{-1}} \sum_{t=1}^{T} \mathbb{E}\left[|(U_t - 1)\beta^T v_t|^2 I\left(|(U_t - 1)\beta^T v_t| > \epsilon \beta^T V_T \beta \sum_{j=1}^{T} B_j^{-1}] \right) \right] \tag{C.13}$$

$$\leq \frac{\sum_{t=1}^{T} \mathbb{E}[(U_t - 1)\beta^T v_t|^4]}{\epsilon^2 (\beta^T V_T \beta)^2 (\sum_{j=1}^{T} B_j^{-1})^2} \tag{C.14}$$

$$\leq \frac{C \sum_{t=1}^{T} \|v_t\|^4}{\epsilon^2 \lambda_{\min}(V_T) (\sum_{j=1}^{T} B_j^{-1})^2} \tag{C.15}$$

So because the sample covariance converges almost surely to the true covariance. We claim that $\lim_{t\to\infty} \mathbb{P}(\lambda_{\min}(V_T) \geq \lambda_{\min}(V)/2) = 1$. Thus we conclude that

$$\mathbb{P}(g_T(\beta) > \delta, \forall \beta \in B) \tag{C.16}$$

$$\leq \frac{C \sum_{t=1}^{T} \|v_t\|^4}{\delta \epsilon^2 \lambda_{\min}(V) (\sum_{j=1}^{T} B_j^{-1})^2} + \mathbb{P}(\lambda_{\min}(V) < \lambda_{\min}(V)/2) \tag{C.17}$$

$$\leq \frac{C \sum_{t=1}^{T} B_t^{-2}}{\delta \epsilon^2 \lambda_{\min}(V) (\sum_{j=1}^{T} B_j^{-1})^2} + \mathbb{P}(\lambda_{\min}(V) < \lambda_{\min}(V)/2) \to 0 \tag{C.18}$$

The last convergence comes from lemma S.6 of Liu et al. Thus we have shown the Lindeberg condition is satisfied and can conclude that $Y_t | D_t \to \mathcal{N}(0, V)$. Thus applying Lemma B.5, Theorem 2, and the delta method the claim is proven. □