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

December 10, 2023

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 on individuals, like drug discovery and vaccines 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 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.

However, while optimizing prediction accuracy, such as $\mathbb{E}_{(\omega_1,\omega_2)}[(y_\theta - y_{\theta^*})^2]$, remains vital, 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 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 as 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, 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 larged 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 estimate 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

[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.

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.

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

Is it possible to desiging 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 phase retrieval [Duchi and Ruan, 2017], Kalman smoothing [Aravkin et al., 2013], to neural regression [Specht, 1991] — the objective often arises from merging a strongly convex function (typically a penalty function like squared-mean or negative loglikelihood) with a non-convex differentiable map. In these contexts, the map's outcomes are defined by a low-dimensional space of *control variables* compressed in lower dimensional but semantically expressive manifold, akin to the bias and weight parameters found in neural networks. Then through this map, the *control variables* are expanded into the much larger space of latent variables within the penalty function. From an optimization standpoint, this composition "hides" the strongly convex geometry of the penalty function, and creates a robust low-dimensional model leading to highly non-convex optimization landscapes.

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.

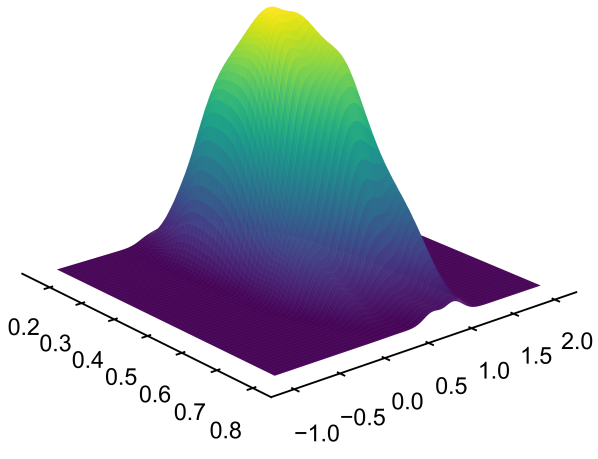
In this paper we aim to offer an *affirmative answer* to the outlined challenges. For this purpose, we assert that the input steaming data only needs to comply with the $\phi-$mixing property [Ibragimov, 1959]—a broader form of time series dependence encompassing classical Markovian dependencies.

Further, 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 in-

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



troduction, 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 in the latent space through a central limit theorem. Leveraging this asymptotic behavior, we formulate a bootstrap variant of SHADE, facilitating the creation of provable confidence intervals.

To consolidate our findings, we provide empirical evidence underscoring the efficacy of the SHADE estimator, particularly in regression settings.

2 Problem setup and preliminaries

Throughout the sequel, we focus on parameter inference of a $m-$dimensional model over a convex set of *control variables* $\theta \in \Theta := \mathbb{R}^m$ subject to a smooth differentiable *loss function* $\ell : \Theta \to \mathbb{R}$, the expectation of a stochastic loss $L(\theta; \omega)$ over a distribution spanning the dataset sample set Ω.

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

Finding the global minima for a general non-convex function like the above is NP-Hard [Danilova et al., 2022, Hochba, 1997], and in general there is no tool for computing the optima[Gower et al., 2019]. In light of these difficulties, our work focuses on optimizing the expressive class of *latent* convex functions defined as follows.

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 $f : \mathcal{X} \to \mathbb{R}$.

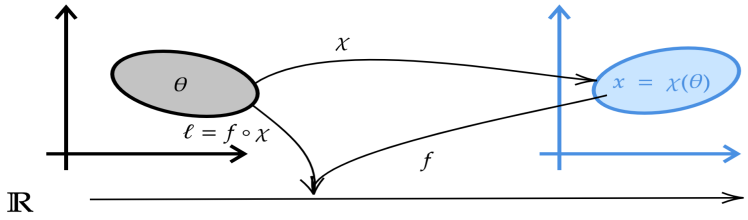
To motivate this notion, we present two examples of latent convex functions. For a schematic representation cf. Figure 2

Example 2.1: Consider the problem of logistic regression in one dimension, where

$$\ell(\theta) := \sum_{s \in I} (\mathrm{sigmoid}(\theta_s \cdot \beta) - a_s).$$

This can be cast as a hidden convex problem where $f(x) = \sum_{s \in I} \exp(x_s - a_s)$, and $\chi(\theta) = \mathrm{sigmoid}(\theta_s \cdot \beta)$

Figure 2: Schematic Representation of Latent Convex Loss



Example 2.2: A more complex example involves equipping the classical linear model with a latent parameter $y_i = w_i^T \chi(\theta)$, over a finite dataset, where χ is a preconfigured multi-layer perceptron (MLP). Minimizing the squared error loss yields an obejective of

$$\ell(\theta) = \|y - W\chi(\theta)\|^2 = \sum_{i=1}^{n} y_i - x_i^T \chi(\theta) = \|y - Wx\|^2$$

Due to the intractability of the true loss in empirical risk minimization problems, we assume only the stochastic loss function $L(\theta; \Omega)$ is given. As such, we impose certain criteria common in the literature [Hazan, 2012, Lan, 2020].

Assumption 1 (Loss Function) *The loss function $\ell(\theta)$ is the expectation of a random composite convex function $L : \Theta \times \Omega \to \mathbb{R}$, 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 $\text{Jac}_\chi(\theta)$ has a bounded spectra; i.e. $\sigma_{\min}(\text{Jac}_\chi(\theta)) > 0$ and $\sigma_{\max}(\text{Jac}_\chi(\theta)) < \infty$*

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 $\nabla \ell(\theta)$.

$\phi-$Mixing: In the sequel, 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

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

Given 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 value of $\phi_\Omega(t) \to 0$ as $t \to \infty$, as the dependency between points weakens with time. 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 $n \to \infty$.*

*Proof.*The broad class of Markov chains (with not necessarily countable state space) satisfy the mixing condition above, and the proof is contained in Davydov [1973] and is a variation on the result of Doeblin 1937. □

3 Latent Estimation

We now present the averaging parameter estimator for learning in latent convex problems, tackling the following theoretical limitations in stochastic optimization literature: (i) the literature has focused on the idealistic structure of a strictly convex underlying objective, (ii) the data is created from an i.i.d. process.

We address the restriction of convex loss functions by introducing SHADE (Stochastic Hidden Averaging Data Estimator), a novel mini-batch parameter estimator with asymptotic theoretical guarantees, which uses mini-batch sampling to mitigate the temporal dependence of the data. The asymptotic results established for SHADE are extended to the Polyak-Ruppert average by exploiting the duality of the latent and control spaces.

Estimation in Iterative Problems Connecting our approach with the previous literature, we begin with a discussion of parameter estimation in a strongly convex setting. A computationally efficient way solution is the stochastic gradient descent algorithm [Robbins and Monro, 1951], which updates the parameter estimates in an online fashion.

$$\theta_T = \theta_{T-1} - \gamma_T \nabla L(\theta_{T-1}; \omega_T)$$

The seminal work of Polyak [1990b] and Ruppert [1988] showed the time average of the iterates

$$\hat{\theta}_T^{\text{ASGD}} = T^{-1} \sum_{t=1}^{T} \theta_t, \qquad \text{(Polyak-Ruppert Average)}$$

enjoys an $\sqrt{n}$-rate of consistency and asymptotic normality, opening the door for many advances in the field seeing great success in a variety of contexts, such as Q-Learning [Li et al., 2023], derivative-free methods [Jin et al., 2021], and over Riemannian manifolds [Tripuraneni et al., 2018].

To reduce the variance of the gradient updates, we assume the stream arrives in chunks of data, or mini-batches [Liu et al., 2023, Qian and Klabjan, 2020]. Note that by optimizing with respect to a batch instead of a singular piece of data, the autocorrelation between gradient terms decreases heuristically.

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

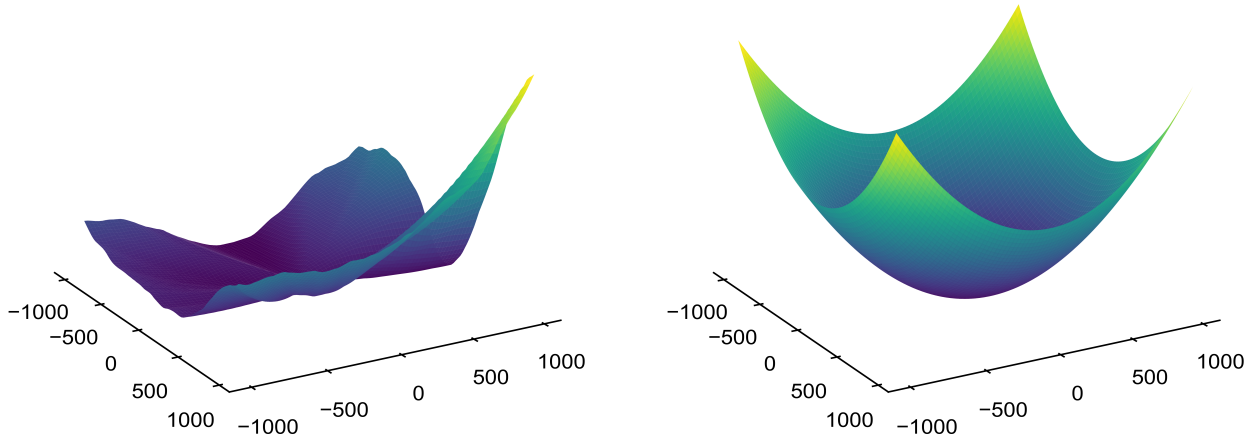
Moreover, Ma et al. [2022] found that within mixing contexts, minibatches decreased the regret bound and sped up convergence. Proceeding, we portion the data stream $\{\omega_s\}_1^{\infty}$ into non overlapping blocks $\Omega_1, ...\Omega_T, ...$, of size $B_t.$, such that $\Omega_T = \{\omega_s : \sum_{t=1}^{T} B_t \leq s \leq \sum_{t=1}^{T+1} B_t\}$.

Descent in Latent Convex Problems In the literature of hidden (latent) games, Sakos et al. [2023] introduce an extension of stochastic gradient descent which enjoys theoretical guarantees of swift convergence to the global optima of latent convex objectives. Taking inspiration from the preconditioning schemes of natural gradient descent [Amari, 1998], and its extension in the context of latent monotone games [Mladenovic et al., 2021], the authors precondition the gradient term, in such aa manner to ensure the dynamics converge. These dyanmics are discretized and introduced as the preconditioned hidden gradient descent augmented below to support minibatching.

Algorithm 1 Batched 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

The preconditioner $\mathbf{P}(\theta)$ is a Riemannian metric over the control space Θ, and the gradient defined relative to this metric captures the natural geometry induced by χ, and restores the strongly convex geometry of the penalty function. In light of this duality between latent and control spaces, we seek a parameter estimator in the latent space $\mathcal{X}$, which takes advantage of the strongly monotone penalty function f. Examining the gradient dynamics in the latent space, the Taylor expansion of $\chi(\theta_t) = x_t$



(a) Loss in Control Space

(b) Loss in Latent Space

Figure 3: Comparison of Loss Profiles Generated by a Hidden Regression, Note the ridges and imperfections are smoothed out

yields

$$\begin{aligned}\chi(\theta_{t+1}) &= \chi(\theta_t - \gamma_t \mathbf{P}(\theta_t) g_t) \\ &= x_t - \gamma_t \nabla f(x_t; \Omega_t) + \gamma_t^2 O(\|\theta_{t+1} - \theta_t\|^2)\end{aligned}$$

This correspondence mimimcs the famed stochastic gradient descent process, with the addition of an asymptotically zero second-order term. Taking inspiration from the Polyak-Ruppert estimator, the time averaged latent iterates yields the SHADE estimator.

$$T^{-1} \sum_{t=1}^{T} \chi(\theta_t) = T^{-1} \sum_{t=1}^{T} x_t \tag{SHADE}$$

Bootstrap Estimation To create confidence intervals, we estimate the covariance structure between model parameters via the random resampling bootstrap from Fang et al. [2018]. This online method remedies the limitations traditional plug-in methods have in high dimensional and correlated data workflows [Chen et al., 2020, Liu et al., 2022].

Multiple estimates are created via perturbing the descent process with random scalings U_t with $\mathbb{E}[U_t] = 1$, and $\mathrm{Var}(U_t) = 1$.

$$\theta_t^{(k)*} = \theta_{t-1}^{(k)*} - U_t^{(k)} \gamma_{t-1} \mathbf{P}(\theta_{t-1}^{(k)*}) \hat{g}_{t-1}$$

Bootstrap versions of the previous estimators are created by taking the time average of the iterates.

$$\hat{x}_T^{\text{SHADE}*} = \sum_{i=1}^{T} x_T^{(*)} = \bar{x}^*$$

This estimator, like the non-scaled counterparts, enjoys robust theoretical guarantees, namely asymptotic convergence and normality. In addition, computationally efficient confidence intervals for functions of the parameter estimates $f(x^*)$ are created via extracting the quantiles of the samples $f(\hat{x}_T^{*(1)}), ..., f(\hat{x}_T^{*(k)})$, as seen in Algorithm 2

Notation: To streamline notiation, 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 with respect to 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 $\nabla f(x; \Omega) = \nabla \frac{1}{|\Omega|} \sum_{s \in \Omega} f(x; \omega_s)$.

4 Asymptotic analysis and results

In this section, we present our main results regarding the asymptotic behaviour of SHADE, the Polyak-Ruppert averaging estimator and the bootstrap estimates, showing they are asymptotically consistent and normal. To this end, the section begins with detailing a few rate limitations intrinsic to the model; the main results are then presented in Section 4.2 with the tools aiding in analysis. The proofs are deferred to the appendix.

Algorithm 2 Bootstrap Batched PHGD

Input: Data $\{\omega_t\}_t^{\infty}$, learning rate γ_t, block size B_t, intial value $\theta_0, \theta^{(k)*}$, for $j \in \{1, ..., k\}$, siginificance level $\alpha \in [0, 1]$.
for $t = 1$ **to** T **do**
  Update $\theta_t = \theta_{t-1} - \gamma_{t-1}\mathbf{P}(\theta_{t-1})\hat{H}_{t-1}$
  Update $\bar{\theta}_t = \frac{t-1}{t}\bar{\theta}_{t-1} + \frac{1}{t}\theta_t$
  for $j = 1$ **to** k **do**
    Generate the random weight $V_t^{(j)}$
    Update $\theta_t^{(j)*} = \theta_{t-1}^{(j)*} - \gamma_{t-1}\mathbf{P}(\theta_{t-1})\hat{H}_{t-1}$
    Update $\bar{\theta}_t^{(j)*} = \frac{t-1}{t}\bar{\theta}_{t-1}^{(j)*} + \frac{1}{t}\theta_t^{(j)*}$
  end for
end for
Let $\hat{\Sigma} = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=t}^{T}(\theta_T^{(j)*} - \bar{\theta}_T^*)(\theta_T^{(i)*} - \bar{\theta}_T^*)^T$
Let l_α, u_α be the $\alpha/2$, and $1 - \alpha/2$ quantiles of $\theta_T^{(1)*}, ..., \theta_T^{(k)*}$.
return point estimator θ_T, empirical confidence interval (l_α, u_α), and empirical covariance matrix $\hat{\Sigma}$.

Model 1 (Descent Parameters) *We impose conditions on the nature of our $\phi-$mixing sequence $\{\omega_s\}_{s=1}^{\infty}$, where our batch size is 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.

Covergence Results: We are now in a position to state the main results regarding the asymptotic behavior of both the Polyak-Ruppert and SHADE estimators. To streamline the presentation, we begin with results regarding consistency before analysis on the asymptotic distribution. We begin with analyzing the consistency of the estimators in question.

Theorem 1 (Consistency of Shade) *Given the assumptions in part 2 and Model 1, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\chi(\theta_t) \xrightarrow{a.s.} \chi(\theta^*).$$

This theorem ensures SHADE estimator is consistent and mirrors 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] which shows that if Lyapunov functions of a non-negative martingale converge to zero, the overall function converges to the optimum.

This theorem is extended to the Polyak-Ruppert estimator via the following lemma. The singular value bound on the Jacobian of the latent mapping restricts norm of its image. Concretely, this is

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 creates a duality gap between the representatives and the parameters themselves. Combining the two above we claim a corollary.

Corollary 1 (Consistency) *Given the assumptions in part 2, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t \xrightarrow{a.s.} \theta^*.$$

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 SHADE estimator satisfies the following*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^{T} B_t^{-1}}}(\bar{x}_T - x^*) \to \mathcal{N}(0, \Sigma)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, and $G = \nabla^2 f(x^*)$

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] This result can also be extended to the Polyak-Ruppert average via the delta method, a theorem which establishes central limit theorems for functions of random variables.

Lemma 2 (Delta Method) *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*

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

Because the latent mapping χ is an *faithful* representation, it is possible to create an inverse function χ^{-1}. Using the Delta Method and the implicit function theorem [Lee, 2003], we can claim the following theorem for the Polyak-Ruppert average.

Corollary 2 (Asymptotic Normality of $\bar{\theta}$) *Given the assumptions above, the following correspondence occurs*

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

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.*

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. The*

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

where $\mathcal{D} = \bigcup_{t=1}^{T} I_t$ *and represents the data used in the empirical risk minimization process, and* Σ *is the covariance matrix in Theorem 2*

This theorem provides the statistical justification for bootstrap covariance estimation and provides a practical manner for robust estimation. The algorithm is described in *.

Importantly, the convergence results of Theorems 1-3 mirror the results found in strongly convex problems with i.i.d. date. This takeaway shows that *when it exists, a hidden convex structure can be exploited to the greatest possible degree, without any loss in inference ability relative to standard, non-hidden convex problems.*

5. Experiments

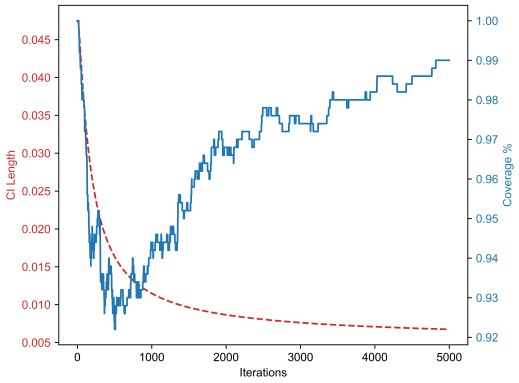
This section shows the applicability of the (SHADE) and Polyak-Ruppert estimators in a few scenarios of theoretical interest. Technical details and additional experimental results are deferred to the appendix. For each of the proceeding scenarios, we seek to find a point estimate and produce a confidence region around this estimate.

To this end, we utilize the bootstrap methods of covariance estimation developed by Fang et al. [2018]. Via creating bootstrap estimates, quantities of use can be estimated via the Algorithm 2

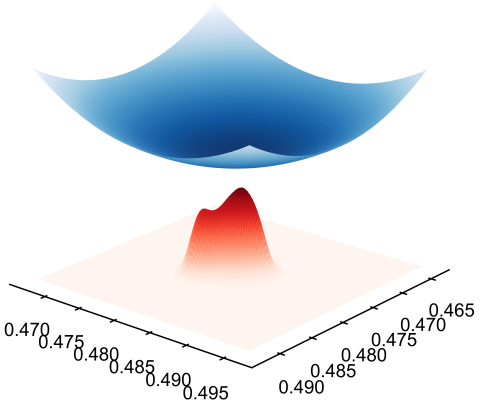
Latent Linear Model We begin with estimation in a latent linear model as seen in Example 2.2, where the optimal regression parameters are controlled via a preconfigured differentiable MLP.

$$L(\theta; \Omega) = \|y - W\chi(\theta)\|^2$$

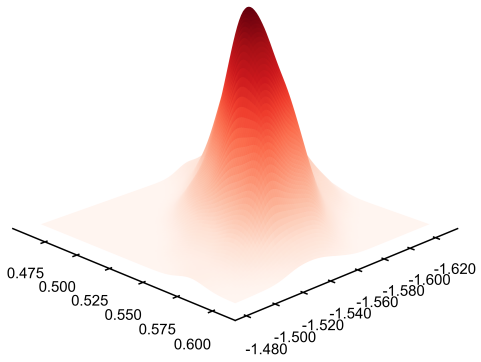
Each MLP acts as the representation map χ, for which each input θ is output the latent variable $\chi(\theta) = x$ which lies in the compact space $\mathcal{X} = [0, 1]^d$. The following figure demonstrates the empirical distribution and loss of the Polyak-Rupert estimator over 500 individual runs. Notably, the peak in the distribution perfectly aligns with the minima of the loss curve, indicating that preconditioned hidden dynamics indeed find the optimal parameters.



(a) Confidence Interval Length and Empirical Coverage Rate



(b) Empirical Distribution and Loss of SHADE in Latent Space



(c) Empical Distribution of SHADE projected onto a hyperplane

Figure 4: The confidence interval length and empirical coverage rates over 500 bootstrap samples for the latent linear models

Medical As an example of the efficacy of our methods to real-world inference problems, we consider determining whether hospital patients have diabetes using the [Dua et al., 2017] dataset. To achieve this, a logistic regression model with a latent regression mapping of key medical features, is used to predict a binary response variable.

$$L(\theta; \Omega) = \log(1 + \exp(-y \cdot W\chi(\beta))).$$

Further experimental details can be found in the supplement. A key insight about this model is the latent map's ability to sparsely encode response vectors and encode the structure, and still retain

prediction accuracy. The plots below show the distribution of the model parameters along multiple planes of interest.

Machine Generated Text As a final show of the power of latent encodings, we present a hidden logistic regression that determines whether academic texts are authentic or generated by an LLM. The input documents are encoded into a 384-dimensional vector via a transformer architecture, and model parameters are hidden via a differentiable MLP. We present classification rates at different sizes of the control manifold.

References

Mohamed Hesham Ibrahim Abdalla, Simon Malberg, Daryna Dementieva, Edoardo Mosca, and Georg Groh. A benchmark dataset to distinguish human-written and machine-generated scientific papers. *Information*, 14(10):522, September 2023. ISSN 2078-2489. doi: 10.3390/info14100522. URL http://dx.doi.org/10.3390/info14100522.

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.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

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.

Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pages 79–163. Springer, 2022.

Yurii Aleksandrovich Davydov. Mixing conditions for markov chains. *Teoriya Veroyatnostei i ee Primeneniya*, 18(2):321–338, 1973.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Wolfgang Doeblin. Sur les propriétés asymptotiques de mouvement régis par certains types de chaines simples. *Bulletin mathématique de la Société roumaine des sciences*, 39(1):57–115, 1937.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

John C. Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *CoRR*, abs/1705.02356, 2017.

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

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.

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.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Dorit S Hochba. Approximation algorithms for np-hard problems. *ACM Sigact News*, 28(2):40–52, 1997.

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

Yanhao Jin, Tesi Xiao, and Krishnakumar Balasubramanian. Statistical inference for polyak-ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.

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.

John M Lee. Introduction to smooth manifolds, 2003.

Xiang Li, Wenhao Yang, Jiadong Liang, Zhihua Zhang, and Michael I Jordan. A statistical analysis of polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2261. PMLR, 2023.

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

Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Xiangyang Ji, Antoni Chan, and Rong Jin. Improved fine-tuning by better leveraging pre-training data. *Advances in Neural Information Processing Systems*, 35:32568–32581, 2022.

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.

OpenAI. Gpt-4 technical report, 2023.

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.

Xin Qian and Diego Klabjan. The impact of the mini-batch size on the variance of gradients in stochastic gradient descent. *arXiv preprint arXiv:2004.13146*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

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.

Donald F. Specht. A general regression neural network. *IEEE Trans. Neural Networks*, 2(6):568–576, 1991. doi: 10.1109/72.97934. URL https://doi.org/10.1109/72.97934.

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

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Nguyen Minh Tien and Cyril Labbé. Detecting automatically generated sentences with grammatical structure similarity. *Scientometrics*, 116(2):1247–1271, 2018.

Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan. Averaging stochastic gradient descent on riemannian manifolds. In *Conference on Learning Theory*, pages 650–687. PMLR, 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 Asymptotic consistency

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

Theorem 1 (Consistency of Shade) *Given the assumptions in part 2 and Model 1, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\chi(\theta_t) \xrightarrow{a.s.} \chi(\theta^*).$$

Corollary 1 (Consistency) *Given the assumptions in part 2, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t \xrightarrow{a.s.} \theta^*.$$

Our proof strategy is comprised of the following steps.

1. We will show the gradient term V_t can be decomposed into the sum of martingales, in a similar fashion to Polyak and Juditsky [1992a] and Liu et al. [2023].
2. In Lemma A.3 , we show the second moment of these martingale terms are then bounded via martingale theory; in turn, in Lemma A.4 the entire second moment of the graident term is shown to be finite.
3. From these bounds and the Lyapunov properties of the energy function, the Robbins-Siegmund theorem [Robbins and Siegmund, 1971], the consistency of the shade estimator is shown. Moreover, because the the Jacobian of the latent mapping χ is of full rank, the image of the latent map can be bounded, and an extension to Corollary One can be made.

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

A.1 Gradient Decomposition To aid in the analysis of the gradient moments, we propose the following gradient decomposition into a sum of martingales. Indeed notice that the gradient of the convex function of the latent states $f(\cdot)$ can be expressed as the population loss at a parameter value and the martingale difference of the empircal and true losses.

$$\begin{aligned}\nabla f(x_{T-1};\Omega_T) &= \mathbb{E}[\nabla f(x_{T-1};\Omega_T)] + (\nabla f(x_{T-1};\Omega_T) - \mathbb{E}[\nabla f(x_{T-1};\Omega_T)]) \\ &= \nabla f(x_{T-1}) + (\nabla f(x_{T-1};\Omega_T) - \nabla f(x_{T-1}))\end{aligned}$$

Taking this one step further, the gradient can be expressed as the difference between the empirical value, and the expected value given $T-1$ timesteps.

$$\begin{aligned}\nabla f(x_{T-1};\Omega_T) &= \mathbb{E}[\nabla f(x_{T-1};\Omega_T)|\Omega_1,...,\Omega_{T-1}] + (\nabla f(x_{T-1};\Omega_T) - \mathbb{E}[\nabla f(x_{T-1};\Omega_T)|\Omega_1,...,\Omega_{T-1}]) \\ &= \mathbb{E}_{T-1}[\nabla f(x_{T-1};\Omega_T)] + (\nabla f(x_{T-1};\Omega_T) - \mathbb{E}_{T-1}[\nabla f(x_{T-1};\Omega_T)])\end{aligned}$$

Combining the above yields that

$$\begin{aligned}\nabla f(x_{T-1};\Omega_T) &= \nabla f(x_{T-1}) + (\mathbb{E}_{T-1}[\nabla f(x_{T-1};\Omega_T)] - \nabla f(x_{T-1})) + (\nabla f(x_{T-1};\Omega_T) - \mathbb{E}_{T-1}[\nabla f(x_{T-1};\Omega_T)]) \\ &= \nabla f(x_{T-1}) + e_T + \zeta_T\end{aligned}$$

The terms e_T and ζ_T are discrete time martingales, [Durrett, 2019], and will allow for the use of the Robbins-Siegmund theorem. Moreover, for simplicity, we denote the batched gradient

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

and $\hat{g}_t = \nabla f(x_t;\Omega_t)$, and $h(x) = \nabla f(x)$. To subsume both the case of bootstrap descent and ordinary descent, we overload the scaling coefficient U_t. It can be both a determinstic identity constant, or a random scalar with mean and variance one. Concretely, it is

$$\theta_{t+1} = \theta_t - \gamma_{t+1} U \mathbf{P}(\theta_t) V_t = \theta_t - \gamma_{t+1} U_t \mathbf{P}(\theta_t)\mathbf{J}(\theta_t)^T (h(\theta_{t+1}) + e_t + \zeta_t)$$

A.2 Gradient Bounds In advance of instutiting the gradient bounds, we present some supplementary lemmas. To start we present a useful moment inequality.

Lemma A.1 (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.2 *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] □

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. □

In advance of the proceeding lemma, define an alternative dataset $\{\tilde{\omega}_s\}_1^{\infty}$, i.i.d. to the original dataset $\{\tilde{\omega}_t\}_1^{\infty}$. From this define the alternative martingale

$$\tilde{\zeta}_T = \nabla L(\theta_T;\tilde{\Omega}_T) - \mathbb{E}_{T-1}[\nabla L(\theta_T;\tilde{\Omega}_T)] = \nabla L(\theta_T;\tilde{\Omega}_T) - \nabla \ell(\theta_{T-1})$$

Lemma A.3 *We seek the following bounds on elements of the descent, where v_t has a 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,

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

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

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

$$\mathbb{E}_t(\|e_t\|^2) \leq \phi^{1-2/p}(B_t) + \phi^{1-2/p}(B_t)\,\mathbb{E}_t[\|e_t\|^p] + \phi^{1-2/p}(B_t)\,\mathbb{E}_t[\|\tilde{e}_t\|^p] \tag{A.3}$$

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

$$\mathbb{E}^{1/p}(\|e_t\|^p) = \mathbb{E}^{1/p}[\|\nabla f(x_t;\Omega_t) - h(x_t)\|^p] \tag{A.4}$$

$$\leq \mathbb{E}^{1/p}[\|\nabla f(x_t;\Omega_t)\|^p] + \sup_{x\in\mathcal{X}}\|h(x)\| \tag{A.5}$$

$$\leq \sup_{\omega\in\Omega_t}\mathbb{E}^{1/p}[M^p(\omega)] + C \tag{A.6}$$

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

$$\mathbb{E}^{1/p}(\|\tilde{e}_t\|^p) \leq \mathbb{E}^{1/p}[M^p(Z)] + C \tag{A.7}$$

Defining v_{1t}

$$v_{1t} = 1 + \mathbb{E}_t[\|e_t\|^p] + \mathbb{E}_t[\|\tilde{e}_t\|^p] \tag{A.8}$$

So, v_{1t} has finite mean

$$\mathbb{E}[v_t] \leq 2\,\mathbb{E}^{1/p}[M^p(\omega^*)] + C \tag{A.9}$$

The bound (A.3) finishes the claim. □

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

$$\hat{g}_t - h(x_t) = \nabla f(x_t;\Omega_t) - h(x_t) \tag{A.10}$$

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

$$\mathbb{E}\,\|\nabla f(x_t;\Omega_t) - h(x_t)\|^2 \leq \frac{C}{B_t} \leq \frac{C}{B_t}(1 + \|x_t - x^*\|^2) \tag{A.11}$$

So we have shown the desired claim. □

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

$$\tilde{\zeta}_t = U_t \nabla f(x_{t-1}; \tilde{\Omega}_t) - \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \tilde{\Omega}_t)] \tag{A.12}$$

As mentioned earlier, $\{\tilde{\omega}\}_{s=1}^{\infty}$ is identical to the process in $(i.)$. So applying Lemma A.2 yields

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

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

$$|\,\mathbb{E}_t(\|\zeta_t\|^2) - \mathbb{E}_t(\|\tilde{\zeta}_t\|^2)| \leq \phi^{1-2/p}(B_t)[1 + \mathbb{E}_t[\|\zeta_t\|^p] + \mathbb{E}_t[\|\tilde{\zeta}_t\|^p]] \tag{A.14}$$

Expanding the terms in above we see that

$$\mathbb{E}_t\,\|\tilde{\zeta}_t\|^2 \leq 2(\nabla f(x_{t-1}; \Omega_t) - \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \Omega_t)] + \|e_t\|^2) \tag{A.15}$$

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

$$\mathbb{E}_{t-1}(\|\zeta_t\|^2) \leq \phi^{1-\frac{2}{p}}(B_{t-1})v_{2t} + CB_t^{-1}(1 + \|x_{t-1} - t^*\|^2) \tag{A.16}$$

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

Lemma A.4 *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.

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

We can then harness conditional expectation to yield that

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

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 of the Jacobian.

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

□

A.3 Robbins Siegmund Theorem: The main theoretical workhorse for showing this consistency of both the SHADE and the Poylak-Ruppert average estimator is the Robbins-Siegmund theorem [Robbins and Siegmund, 1971]. The theorem is an extension of the Doob martingale convergence theorem and is detailed as follows.

Theorem A.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 A.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]. □

A.4 Putting Together the Pieces We are now in a position to show Theorem 1.

Theorem 1 (Consistency of Shade) *Given the assumptions in part 2 and Model 1, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\chi(\theta_t) \xrightarrow{a.s.} \chi(\theta^*).$$

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.

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

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

$$E_{t+1} \leq E_t - \gamma_t\langle h(x_t), x_t - \tilde{x}\rangle + \gamma_t\alpha_t + \gamma_t^2\psi_t \leq E_t + \gamma_t\alpha_t + \kappa\gamma_t^2\psi_t \tag{A.17}$$

We have that

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

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

$$\sum_{t\geq 1}(C\gamma_t^2\tilde{\sigma}_{\max}^2 + C\gamma_t^2\tilde{\sigma}_{\max}^2 B_t^{-1}) < \infty \tag{A.18}$$

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

$$\sigma_{\min}\|\theta_t - \theta^*\| \leq \|\Delta_t\| \leq \sigma_{\max}\|\theta_t - \theta^*\| \tag{A.19}$$

Thus this implies that $\|\theta_t - \theta^*\|^2$ goes to zero almost surely and thus we have shown the desired claim. □

A.5 Consistency of Polyak-Ruppert Average It is possible to extend the consistency result of Theorem 1, to the Polyak-Ruppert average via the following lemma. Because the Jacobian of the latent mapping χ is *faithful*, the norm of the image of the image can be bounded in an intuitive manner.

Lemma A.5 *If χ is a faithful representation satisfying A2 (iii), we have that*

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

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

$$\lim_{t\to 0} \frac{|\chi(\theta) - \chi(\theta + tv)|}{t} = \mathbf{J}(\theta)(v) \tag{A.1}$$

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

$$||\frac{|\chi(\theta) - \chi(\theta + tv)|}{t} - \mathbf{J}(\theta)(v)|| < \varepsilon \tag{A.2}$$

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

$$-\varepsilon + \sigma_{\min} \leq -\varepsilon + ||\mathbf{J}(\theta)(v)|| < ||\frac{|\chi(\theta) - \chi(\theta + tv)|}{t}|| < \varepsilon + ||\mathbf{J}(\theta)(v)|| \leq \varepsilon + \sigma_{\max} \tag{A.3}$$

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

$$-\varepsilon + \sigma_{\min} < \frac{||\chi(\theta) - \chi(\theta^*)||}{||\theta - \theta^*||} < \varepsilon + \sigma_{\max} \tag{A.4}$$

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. □

Combining lemma A.5 and Theorem 1, we can conclude the following Corollary.

Corollary 1 (Consistency) *Given the assumptions in part 2, we conclude that the iterates of the batched pre-conditioned gradient descent satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \theta_t \xrightarrow{a.s.} \theta^*.$$

Proof. Noticing that Theorem 1 implies that $\|x_t - x^*\|$ goes to zero almost surely, applying lemma A.5 yields that $\|\theta_t - \theta^*\|$ also converges to zero. Thus the desired claim is shown. □

B Proof of Asymptotic Normality

The goal in this appendix is to prove Theorem 2, Corollary 2, and Theorem 3. For covenience, we restate the results below.

Theorem 2 (Asymptotic Normality of SHADE) *Given the assumptions above, the SHADE estimator satisfies the following*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^{T} B_t^{-1}}}(\bar{x}_T - x^*) \to \mathcal{N}(0, \Sigma)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, *and* $G = \nabla^2 f(x^*)$

Corollary 2 (Asymptotic Normality of $\bar{\theta}$) *Given the assumptions above, the following correspondence occurs*

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

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. The*

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

where $\mathcal{D} = \bigcup_{t=1}^{T} I_t$ *and represents the data used in the empirical risk minimization process, and* Σ *is the covariance matrix in Theorem 2*

The proof strategy will comprise of the following steps.

1. We first create a descent inequality for the latent iterates x_t based on the Taylor expansion of the latent map. The modified geometry in the latent space will simplify analysis.
2. In lemma Lemma B.3 we aim to find an expression for the error of SHADE in terms of scaled gradients. This is done via the nested Hessian equality from Polyak [1990a].
3. In lemma Lemma B.5, we show via the Lindeberg condition that the sum of gradients exhibits asymptotic normality.

B.1 Latent Descent Inequality: In a similar vein to the overall descent, we establish the following result.

Lemma B.1 *Let* $x_t = \chi(\theta_t)$, *we then have that*

$$x_{t+1} = x_t - \gamma_t U_t \hat{g}_t + \gamma_t^2 U_t^2 O(\|\theta_t - \theta_{t-1}\|^2). \tag{B.1}$$

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. □

B.2 Towards Asymptotic Normality: This section establishes an asymptotic normality claim for sums of graidents of the monotonic loss function f. To begin, we recall a result from Polyak and Judditsky that characterizes the hessian term throughout the evolution of the descent process. We define G to be the Hessian matrix at the optimal latent state x^*, $\nabla^2 f(x^*)$.

Proposition B.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 **?** and (iv) is in Liu et al. [2023]. This result is crucial in transforming the expression $x_T - x^*$ into one containing the gradient of the f. □

Lemma B.2 *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 are able to craft the following correspondence.

Lemma B.3 *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.

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

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 B.1 that 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 B.2 again, we have that $\|Q_j^T\| < C$. Then applying

Lemma A.3, we have that

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

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

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

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

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

□

B.3 Establishment of Asymptotic Normality: We are now in a position to lay out cursory asymptotic normality claims. Throughout this section, we will refer to $r(t)$ the autocorrelation coefficient redefined 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 to do not grow too large in a a $\phi-$mixing sequence.

Lemma B.4 *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. □

Now, we outline the primary lemma to aid in proving normality.

Lemma B.5 *Given the assumptions in part 3 we claim that*

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

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 $\nabla f(x^*; \Omega_t) = \hat{g}_t^*$ We see that the second moment of this expression can be expressed as

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

We now aim to bound these expressions via our assumptions on the $\phi-$mixing nature of these sequences. When $t > k+1$, we have that I_t and S_{k+1} 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. Thus we conclude that

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

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 $v_s = \nabla_{x_t} f(x^*; \omega_s)$ and $r(t) = \mathbb{E}[\nabla_{x_{i,k}} f(x^*; \omega_s) \nabla_{x_{i,t+k}} f(x^*; \omega_s)]$. 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,

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

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

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

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

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

If we define the above as ν_T we know that this value is finite. Thus we can have

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

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

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

Thus combining this all together we satisfy the Lindeberg condition.

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

The final equality is due to lemma. □

B.4 Proof of Results: We begin via establishing Theorem 2.

Theorem 2 (Asymptotic Normality of SHADE) *Given the assumptions above, the SHADE estimator satisfies the following*

$$\frac{T}{\sqrt{\sum_{t\geq 1}^T B_t^{-1}}}(\bar{x}_T - x^*) \to \mathcal{N}(0, \Sigma)$$

where $\Sigma = G^{-1}(2r(0) + 4\sum_{k\geq 1} r(k))G^{-1}$, *and* $G = \nabla^2 f(x^*)$

Proof. Given lemma B.4, we have that

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

Then applying Lemma 5 we then can conclude the desired claim. □

We extend this proof to the Polyak-Ruppert average via the delta method, restated here for convinience.

Lemma 2 (Delta Method) *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*

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

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

The following corollary can be directly claimed from these results.

Corollary 2 (Asymptotic Normality of $\bar{\theta}$) *Given the assumptions above, the following correspondence occurs*

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

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.*

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. The*

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

where $\mathcal{D} = \bigcup_{t=1}^{T} I_t$ *and represents the data used in the empirical risk minimization process, and* Σ *is the covariance matrix in Theorem 2*

Proof. Define $v_t = \hat{g}_t^*$. Then we claim that

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

Thus 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 4 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 lemma (?) 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

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

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

$$\leq \frac{C \sum_{t=1}^{T} \|v_t\|^4}{\epsilon^2 \lambda_{\min}(V_T) (\sum_{j=1}^{T} B_j^{-1})^2} \tag{3}$$

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

$$\mathbb{P}(g_T(\beta) > \delta, \forall \beta \in B) \tag{4}$$

$$\leq \frac{C \sum_{t=1}^{T} \|v_t\|^4}{\delta \epsilon^2 \lambda_{\min}(V) (\sum_{j=1}^{T} B_j^{-1})^2} + \mathbb{P}(\lambda_{\min}(V) < \lambda_{\min}(V)/2) \tag{5}$$

$$\leq \frac{C \sum_{t=1}^{T} B_t^{-2}}{\delta \epsilon^2 \lambda_{\min}(V) (\sum_{j=1}^{T} B_j^{-1})^2} + \mathbb{P}(\lambda_{\min}(V) < \lambda_{\min}(V)/2) \to 0 \tag{6}$$

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 and theorem 2, the claim is proven. □

C Experiments

This section shows the applicability of the estimators discussed earlier in a host of applications. We elaborate on the examples given in section 5, as well as present novel scenarios that further demonstrate the method's usefulness. The section begins with a overview on how the inference is conducted, and proceeds to detail each of the applications.

In each example, we define a latent convex estimation problem, in which the loss function is required to be latently convex. The model parameters are interfaced via a set of control variables through a pre-configured neural network which acts as the player's representation map $\chi(\theta)$. Attached to the parameter estimates are confidence regions generated by bootstrap algorithm described in section 3. The dimensions of the mapping vary with the datasets, and for the examples, we provide both point estimates and confidence regions.

We use the random bootstrap algorithm from Fang et al. [2018] to create an empirical estimate of the parameter covariance and cumulative distribution function. By taking the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical cumulative distribution, a $\alpha-$level confidence interval can be constructed. An addition confidence interval can be created with the empirical covariance, recalled here for reference

$$\hat{\Sigma} = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=t}^{T}(\theta_T^{(j)*} - \bar{\theta}_T^*)(\theta_T^{(i)*} - \bar{\theta}_T^*)^T,$$

By taking the sample variance of each of the parameters, a z-confidence interval can be constructed by finding

$$\hat{\theta}_i \in (\hat{\theta}_i - z_{\alpha/2}\hat{\Sigma}_{ii}, \hat{\theta}_i + z_{\alpha/2}\hat{\Sigma}_{ii})$$

$\hat{\Sigma}_{ii}$ is the ith element of the diagonal of the empirical covariance and $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal random variable.

Latent Linear Model: The first model discussed is the hidden linear model found in Example 2.2, which aims to minimize the loss, where $\omega \in \Omega$ is a tuple $(y, \vec{x}^T)^T$

$$L(\theta; \Omega) = \|y - X\chi(\beta)\|^2.$$

Here the control variables θ are two dimensional and are fed into the player's MLP given by the representation maps

$$\chi(\theta) = \alpha^{(2)} \cdot \text{CeLU}(\alpha^{(1)} \cdot \theta))$$

where $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{R}^{2\times 2}$ are randomly chosen. The activation function CeLU, is given below for reference.

$$\text{CeLU}(x) = \max\{0, x\} + \min\{0, \exp(x) - 1\}$$

In this example and the ones to follow, we set the bias of the fully-connected layers to zero to simplify notation. The output of the MLP $\chi(\theta)$ becomes the regression vector from which the linear model is based.

Moreover, the data points are assumed to be drawn from a $\phi-$mixing stream, and are generated via a autoregressive process. The covariates x_t are created from the following vector autoregressive process, where ε_t is a standard normal random variable.

$$x_t = Px_{t-1} + \varepsilon_t \tag{C.1}$$

The transform $P = M\Lambda M^T$ is the product of an orthogonal matrix M and uniformly random diagonal matrix Λ. The unique minima of the loss function $\ell(\theta)$ is generated randomly, and in our case lies at

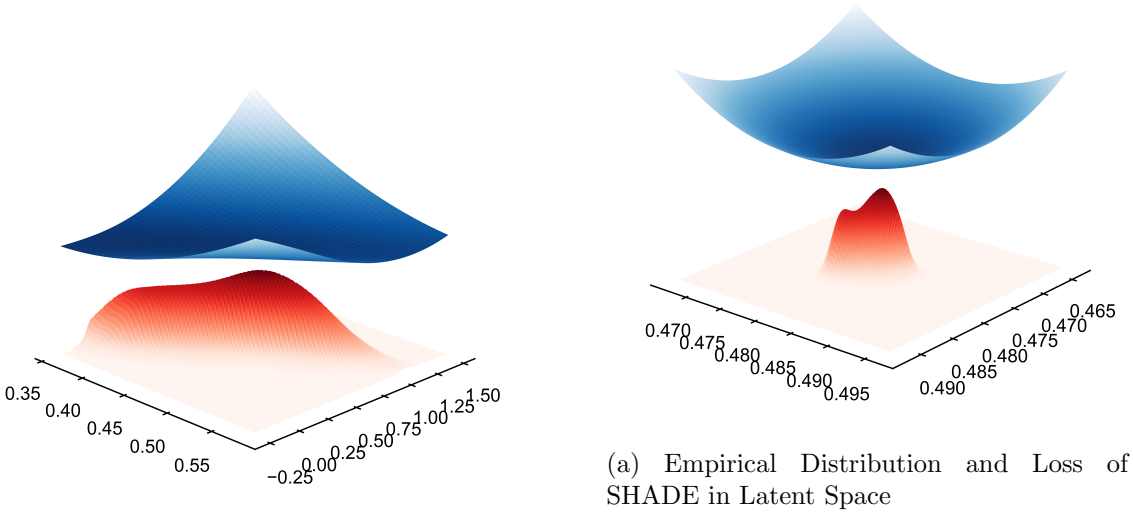
$$x^* = \chi(\begin{bmatrix}0.2 & -0.3\end{bmatrix})^T$$

In each experiment, there are 200 bootstrap samples created, 10000 training steps, and 500 trials completed. The learning rate is set to be $\gamma_t = (t+\gamma_0)^{-0.66}$, where $\gamma_0 = 10$ in line with [Ma et al., 2022], and the batch size is $t^{0.3}$.

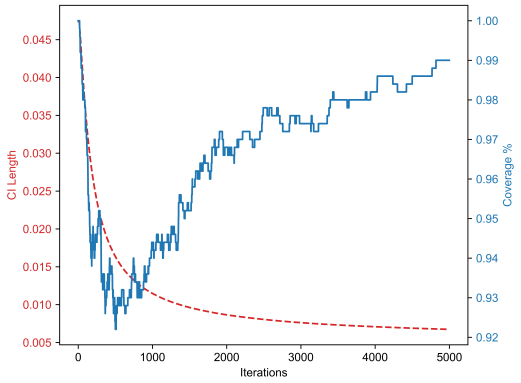
To visualize the normality of the estimator, we present plots of the distribution and the loss of the parameters over the control and latent spaces. The parameter space (θ_1, θ_2) are represented in the x and y axes, and the distribution of the SHADE and Polyak-Ruppert estimators are graphed through a kernel density estimation of the bootstrap samples. The loss is calculated over the average of thousands of samples generated through the same process as the training data. In addition, to show the consistency of the estimator, we plot the average size and empirical coverage rates of the confidence intervals generated by 500 bootstrap samples.

Measurement Est		
	Est θ	95 %CI
$\chi(\theta)_1$	0.451	(0.4504, 0.4516)
θ_2	0.506	(0.500, 0.512)

Figure 5: Estimation of linear regression



(a) Empirical Distribution and Loss of SHADE in Latent Space



(b) Confidence Interval Length and Coverage Percentage

Figure 6: The confidence interval length and empirical coverage rates over 500 bootstrap samples for the latent linear models

POWER Measurements: The next example is applying a latent linear model to the POWER dataset from the UCI Machine Learning Repository Dua et al. [2017]. The dataset consists of millions of measurements of electricity usage collected from a house over the course of 47 months, Removing missing values, each of the measurements are classified by when they occur, (a.) morning (00:00-11:59), (b.) afternoon (12:00-17:59), and evening (18:00-23:59). These three classes are one-hot encoded into data vectors from which the latent regression is performed.

$$y = \chi(\beta)^T x = \chi(\begin{bmatrix}\beta_{\text{morning}} & \beta_{\text{afternoon}} & \beta_{\text{evening}}\end{bmatrix})^T x \qquad \text{(Model A)}$$

As the power data arrives in a stream, the usage at points in time close to each other are highly correlated and acts as an empirical example of our methods. The latent mapping is defined

$$\chi(\theta) = \alpha^{(2)} \cdot \text{CeLU}(\alpha^{(1)} \cdot \theta))$$

The matrices $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{R}^{3\times 60}, \mathbb{R}^{60\times 3}$ are populated uniformly and subsequently normalized through He normalization [He et al., 2015]. As with the last example, we assume a learning rate of $\gamma_t = (t+10)^{-0.66}$ and a batch size $B_t = t^{0.3}$. The results shown are averaged over 400 bootstrap samples and 10000 training steps.

Figure 7: Estimation of linear regression

Measurement Est		
	Est.	95 %CI
Morning	0.7968	(0.7895, 0.8005)
Afternoon	1.0569	(1.0538, 1.0607)
Evening	1.8098	(1.8088, 1.8115)

Medical Experiments The applicability of our methods can be extended to testing in medical contexts. To this end, we use the diabetes dataset from machine learning repository [Dua et al., 2017] in order to predict whether hospital patients have diabetes through a logistic regression augmented with a latent mapping.

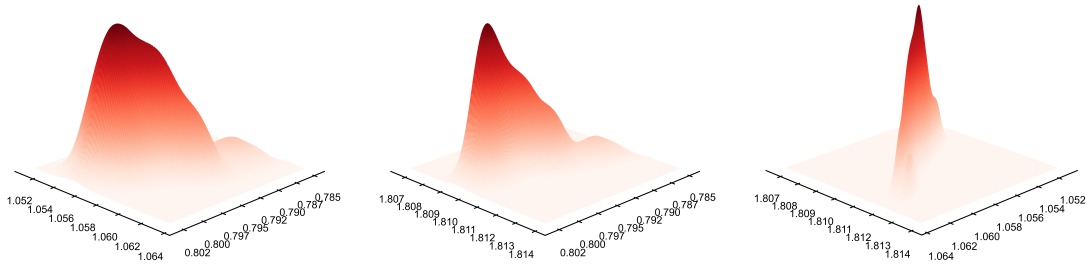
$$L(\theta; \Omega) = \log(1 + \exp(-yW\chi(\beta))$$

The dataset contains over 250,000 patients, from which data on age, vegetable consumption, income and key health indicators were collected. The target is a binary response variable y which indicates whether the patient is diabetic or predisposed to the condition. The scalar features are normalized into the normal interval via min-max scaling, where the minimum and maximum of a feature are set to be zero and one, the remaining data a linearly interpolated in between. The latent map in our case is a differentiable multi-layer perceptron with two hidden layers, that maps the control space $\mathbb{R}^{10}$ into the feature space $\mathcal{X} \subseteq \mathbb{R}^{21}$.

$$\chi(\theta) = \alpha^{(2)} \cdot \text{CeLU}(\alpha^{(1)} \cdot \theta))$$

The linear mappings $\alpha^{(1)} \in [-1, 1]^{10\times 21}, \alpha^{(2)} \in [-1, 1]^{21\times 21}$. and the CeLU function is defined above. The control space is smaller than the latent space in this scenario and represents a *sparse* encoding of the regression vector $\chi(\beta)$. The experiments are run 400 times with 10000 training steps and 300 bootstrap samples. In accordance with the literature, the learning rate is $\gamma_t = (t + 10)^{-0.5}$ and a batch size of $B_t = t^{0.3}$. Pictured below in Figure 9 are the distributions of linear combinations of model parameters, i.e. the distribution of $(c_1^T\chi(\beta), c_2^T\chi(\beta))$.

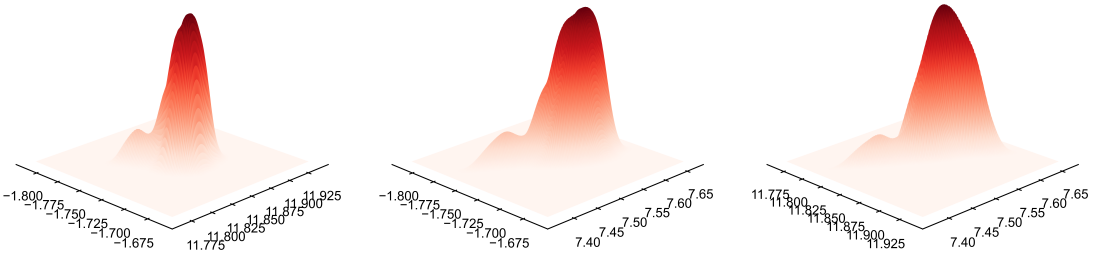
ChatGPT Detection



(a) Empirical Distribution Polyak-Ruppert Estimator of β_{morning} and $\beta_{\text{afternoon}}$

(b) Empirical Distribution Polyak-Ruppert Estimator of β_{evening} and $\beta_{\text{afternoon}}$

(c) Empirical Distribution Polyak-Ruppert Estimator of β_{morning} and β_{evening}



(d) Empirical Distribution of the SHADE Estimator of $\chi(\beta)_0$ and $\chi(\beta)_1$

(e) Empirical Distribution of the SHADE Estimator of $\chi(\beta)_1$ and $\chi(\beta)_2$

(f) Empirical Distribution of the SHADE Estimator of $\chi(\beta)_1$ and $\chi(\beta)_2$

Figure 8: The confidence interval length and empirical coverage rates over 500 bootstrap samples for the latent linear models

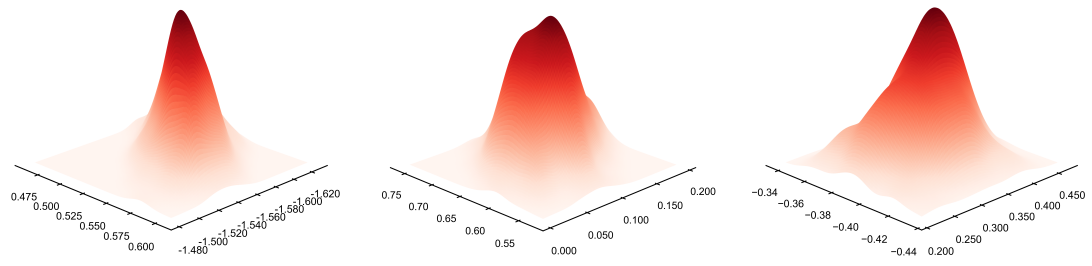


Figure 9: Distribution of Latent Model Parameters Along Hyperplanes.

LLM Generated Text As a final use of the efficacy of our methods, we detect whether academic papers were human-generated or created by large language models using the scientific papers dataset

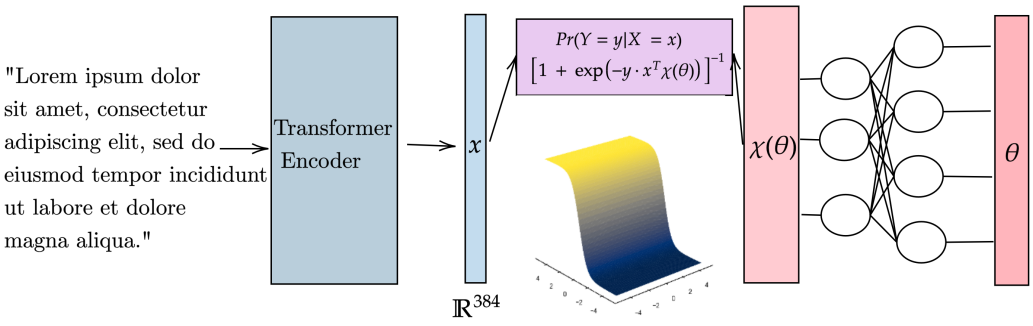


Figure 10: Schematic of ChatGPT Experiment

from Abdalla et al. [2023]. The dataset contains over 20,0000 examples of papers from human-generated sources such as arKiV, and machine generated sources, such as GPT-4, Galatica, and SCIgen [OpenAI, 2023, Taylor et al., 2022, Tien and Labbé, 2018], consisting of abstracts, introductions and conclusions. To classify the papers, we compress the abstract and introduction paragraphs into a vector embedding of 384 features through the sentence encoder from [Reimers and Gurevych, 2019], which under utilizes the transformer architecture from the BeRT model [Devlin et al., 2018]. The introductions and abstracts are naturally correlated with each other yielding mixing data for the model. A latent logistic regression is performed on the data, with a MLP taking various control spaces $\Theta \subseteq \mathbb{R}^{25}, \mathbb{R}^{50}, \mathbb{R}^{100}, \mathbb{R}^{200}$ into $\mathcal{X} \subseteq \mathbb{R}^{384}$.

$$\chi(\theta) = \alpha^{(2)} \cdot \text{CeLU}(\alpha^{(1)} \cdot \theta))$$

For a schematic representation cf Figure 10 The linear mappings $\alpha^{(1)} \in [-1, 1]^{50 \times 384}, \alpha^{(2)} \in [-1, 1]^{384 \times 384}$. This is another use of the latent regression enforcing a low rank contraints. Below we note show graphs of the empirical distributions of the parameters.