Statistical Inference in Latent Convex Objectives with Stream Data

Rohan Chauhan RMCHAUHA@UCI.EDU
University of California, Irvine
Emmanouil-Vaselios Vlatakis-Gkaragkounis VLATAKIS@WISC.EDU
University of Wisconsin-Madison
Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley

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 [6]. Optimizing the parameters of these data-driven models is foundational in the discipline of statistics, and it is well known that in the simplified case of the aforementioned loss $L(\theta, \omega)$ is strongly convex with respect to θ, then θ^* can be estimated with the ubiquitous stochastic gradient descent algorithm and the resulting estimator $\hat{\theta}$ enjoys consistency and asymptotic normality results from the work of Polyak and Judditsky [25]. Yet extending this line of work into a framework that includes both dependent, streaming data and non-convex losses is by no means trivial, implying classical methods that have worked in decades past often struggle with the complexities of modern machine learning workflows.

Many of these objectives are characterized by *hidden structure*, where the model parameters are interfaced through an intricate and often intractable representation mapping. This paradigm has been used in diverse fields such as generative modeling [18], policy gradient methods within reinforcement learning [17], and multi-agent games [34] among others, leading to a composition that ”hides” the strongly convex geometry of the penalty function, yielding non-convex, but tractable objectives.

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

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

Prior work in this area have been shown to be challenging and fruitful areas of inquiry. The landmark paper of Polyak and Juditsky [27] demonstrating the asymptotic statistical behaviour of iterates of a stochastic descent process, has led to the establishment of similar statistical statements in wide ranging areas from policy gradients to functional analysis. Extending these results, the recent paper Liu et al. [22], showed the consistency and normality of estimators built on the SGD framework despite $\phi-$mixing weakly dependent data.

Optimization in the face of a non-convex loss has proven a challenging task, and little is known of asymptotic statistical behaviour estimators trained on a non-convex losses.

Our Results & Techniques. We aim to tackle the above challenges, by utilizing inference methods that are statistically sound in the presence of streaming, mixing data and under the non-convex class of latent convex losses. To this end, we assert the input steaming data only needs to comply with the $\phi-$mixing property [14], which subsumes a broad class of mixing processes, including the $\alpha-$(strong) and $\beta-$mixing conditions among others.

Inspired by the streaming demands of large-scale inference systems and sequential data applications, and the ubiquity of latent representations with modern deep learning, we introduce SHADE, an innovative gradient-based estimator that capitalizes on the representation function linking control and latent variables. To our knowledge, the asymptotic statistical results concerning SHADE render it the first set of gradient-based estimators that are *consistent and asymptotically normal* under latent convex losses, for both mixing and i.i.d data.

Aligning with the criteria outlined in our introduction, Theorem 2 confirms the consistency of our estimator and Theorem 4 demonstrates the SHADE estimator's asymptotic normality through a tailored central limit theorem. In addition, leveraging recent advances in bootstrap covariance estimation under mixing data [5, 11], we formulate a bootstrap variant of SHADE and show in Theorem 5 that the asymptotic behavior of the estimates matches that of the SHADE itself. Consolidating our findings, we provide empirical evidence underscoring the efficacy of the SHADE estimator in real-world settings.

2. Problem setup and preliminaries

To tackle the above challenges, we aim to estimate the true parameters $\theta^* \in \mathbb{R}^d$ of a d-dimensional model, through minimization of an objective function, expressed as the expected loss over a *stationary* distribution $\Pi(\omega)$ spanning the dataset sample space Ω, also known as the population risk.

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

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

In the sequel, we assume the true loss $\ell(\theta) = \mathbb{E}[L(\theta;\omega)]$ admits hidden structure by virtue of being a latent convex function.

Definition 1 (Latent Convex Function) *A function $\ell : \Theta \to \mathbb{R}$ admits the following composition*

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

with its components satisfying the following constraints.

1. *$\chi : \Theta \to \mathcal{X}$, a Lipschitz smooth injective mapping with no critical points onto a closed and convex set $\mathcal{X}$, with* $\mathrm{cl}(\chi(\Theta)) = \mathcal{X}$.

2. *$f : \mathcal{X} \to \mathbb{R}$, is a Lipschitz smooth and strongly convex function.*

Directly computing the expected loss $\ell(\theta)$ is often intractable; as such, we assume only the stochastic loss function $L(\theta; \Omega)$ defined over parameter space, and the data distribution is given. We impose the following conditions on the latent convex objective below. [13, 20].

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

- *$L(\theta; \omega)$ is differentiable and $c-$Lipschitz in θ for almost all ω.*
- *For a constant $p > 2$, and a function $M(\cdot)$, the gradients of $L(\theta; \omega)$ have bounded p-th moments,*

$$\sup_{\theta \in \Theta} \| \nabla L(\theta; \omega) \|^p \leq M^p(\omega), \; \mathbb{E}[M^p(\omega)] < \infty$$

- *The singular values of the Jacobian of the latent map χ at θ, $\mathrm{Jac}_\chi(\theta) = \mathbf{J}(\theta)$ have bounded spectra*

$$\sigma^2_{\min}(\mathbf{J}(\theta)) \leq \sigma(\mathbf{J}(\theta)\mathbf{J}(\theta)^T) \leq \sigma^2_{\max}(\mathbf{J}(\theta))$$

where $\sigma_{\min}, \sigma_{\max} \in (0, \infty)$

Notation: The parameters in the control space are denoted $\theta \in \Theta$ and their latent representations $x = \chi(\theta) \in \mathcal{X}$. We refer to its Jacobian evaluated at θ as $\mathbf{J}(\theta) := \mathrm{Jac}(\chi(\theta))$. For brevity, we denote the gradient concerning multiple data points with the following expression.

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

3. Estimation in latent objectives

To ensure the statistical soundness of the empirical risk relaxation, two core criteria for the estimator θ_t must be met (See Wasserman [35]). Namely, the *consistency* of the estimator θ_t, which ensures it converges in probability to the true parameter value as the sample size tends to infinity and the existence of a *central limit theorem* (CLT), which identifies an asymptotic limiting distribution, is pivotal for crafting confidence intervals for θ_t.

Building on these prerequisites, [33] demonstrated that under mild regularity conditions, ERM solutions exhibit asymptotic normality, meaning $\sqrt{n}(\theta_t - \theta^*)$ weakly converge to a normal distribution.

To ensure an optimally $\sqrt{n}$-consistent estimator for solutions to stochastic approximation objectives, Polyak [28] and Ruppert [31] independently proposed the averaged SGD (ASGD) iterate:

$$\hat{\theta}^{\text{ASGD}}_n = n^{-1} \sum_{t=1}^{n} \theta_t, \text{ where } \theta_t = \theta_{t-1} - \gamma_t \nabla L(\theta_{t-1}; \omega_t) \qquad \text{(Polyak-Ruppert averaging scheme)}$$

for a diminishing learning rate $\gamma_t \propto 1/t^\rho, \; \rho \in (0.5, 1)$.

Capitalizing on the hidden structure of latent convex objectives, Sakos et al. [32] introduced the preconditioned hidden gradient algorithm inspired by gradient flows on Riemmanian manifolds, extending the convergence rates of the averaged stochastic gradient descent iterate to latent convex objectives.

$$\theta_t = \theta_{t-1} - \gamma_{t-1}\mathbf{P}(\theta_{t-1})\,\nabla\,L(\theta_{t-1};\Omega_t),\ \text{where}\ \mathbf{P}(\theta) = [\mathbf{J}(\theta)^T\mathbf{J}(\theta)]^+ \tag{PHGD}$$

This algorithm enjoys an improved rate of convergence on latent strongly convex functions compared to the conventional Polyak-Ruppert average, and with this in mind, we base our estimator on iterates of this process.

4. Main results

In this section, we introduce the notion of ϕ-mixing sequences and present our main results regarding the asymptotic behaviour of SHADE.

4.1 Correlated Data Streams Formally, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the mixing coefficient for $\mathcal{A}, \mathcal{B}$, where $\mathcal{A}, \mathcal{B}$ are sub-sigma-algebras of $\mathcal{F}$, is defined

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

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

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

This sequence $\{\phi_\Omega(t)\}_{t=1}^\infty$ is $\phi-$mixing if $\phi_\Omega(t) \downarrow 0$ asymptotically. Heuristically, $\phi_\Omega(t)$ is the maximum amount of information gained by knowing data from t or more steps in the past.

Given our dataset $\{\omega_s\}_{s=1}^\infty$, if we partition this sequence into blocks of sufficient size $\{\Omega_t\}_{t=1}^\infty = \{\omega_s\}_{s=1}^\infty$, neighboring blocks will be almost independent due to the weaker correlation of data points temporally separated. Furthermore, if the block is sufficiently small, they behave similarly to the original $\phi-$mixing sequence. With this in mind, we divide our data into two sets of non-overlapping blocks $\Omega_t^a = \Omega_{2t}$, $\Omega_t^b = \Omega_{2t+1}$, and run PHGD on the two sets of blocks.

$$\hat\theta_t^k = \hat\theta_{t-1}^k - \gamma_t\mathbf{P}(\hat\theta_{t-1}^k)\nabla L(\hat\theta_{t-1}^k;\Omega_t^k), k\in\{a,b\}$$

From these two streams, we use the Polyak-Ruppert average the estimates, yielding our stochastic hidden averaging data estimator.

$$\hat\theta_t = \frac{1}{2T}\sum_{t=1}^T(\theta_t^a + \theta_t^b) \tag{SHADE}$$

We now stipulate some mild conditions on the relationship between the batch size, learning rate and mixing coefficients.

Descent Parameters *The following conditions on the data sequence $\{\omega_s\}_{s=1}^\infty$, batch size B_t and learning rate γ_t are set, where $\alpha\beta(1/2 - 1/p) = \rho$:*

- *The learning rate of the algorithm is of the form $\gamma_t = (\gamma_0 + t)^{-\rho}$, $\rho \in (0.5, 1)$*
- *The batch size satisfies $B_t = \lceil t^\alpha\rceil$ where $\alpha \in (0, 1)$*
- *The mixing coefficients satisfy $\phi(t) = t^{-\beta}$ where $\beta > 2$*

4.2 Asymptotic Behavior We wish to demonstrate our estimators are statistically sound, i.e. they are *consistent* and *asymptotically normal*. To this end, we first show that SHADE is asymptotically consistent.

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

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

This theorem ensures the SHADE estimator is strongly consistent and follows the results found in Polyak and Juditsky [26] and Liu et al. [22]. This proof relies on the Robbins-Siegmund theorem [30], a celebrated result that relates the convergence of a Lyapunov function of a non-negative martingale implies the overall martingale converges as well.

Theorem 3 (Asymptotic Normality of SHADE) *Given the assumptions above, the following theorem occurs*

$$\frac{T}{\sqrt{\sum_{t=1}^{T} B_t^{-1}}}(\hat{\theta}_T - \theta^*) \xrightarrow{d.} \mathcal{N}(0, \Sigma)$$

where $r(t) = \mathbb{E}[\nabla L(\theta_{k+t}; \Omega_{k+t}) \nabla L(\theta_k; \Omega_k)^T]$, $H = \nabla^2 f(x^*) \mathbf{J}(\theta^*)$, $V = (2r(0) + 4\sum_{k \geq 1} r(k))$, *and* $\Sigma = H^+ V [H^+]^T$.

This result mirrors the central limit theorem found in Polyak and Juditsky [26], Liu et al. [22], and Mou et al. [24], and uses the "sandwich" covariance structure found in [10]

These asymptotic normality results can be extended to the bootstrap estimates, by showing a version of the Lindeberg condition holds for bootstrap iterates. [4].

Theorem 4 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

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

where $\mathcal{D} = \{\omega_i | i \in \Omega_t^a \cup \Omega_t^b\}$ *and represents the data used in the empirical risk minimization process, and* $\hat{\Sigma} = \Sigma$ *is the covariance matrix in Theorem 4*

This theorem provides the theoretical guarantee that the asymptotic distributions of the bootstrap estimates will match that of SHADE. This ensures the consistency of these estimates and provides a base to construct confidence regions.

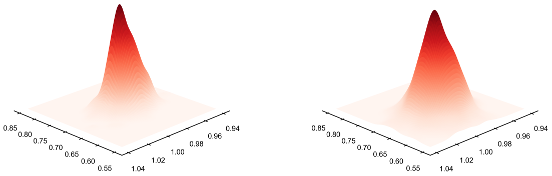
5. Experiments

This section shows the applicability of the estimators discussed earlier in a host of applications. In each of the proceeding examples, we define an estimation problem, in which the loss function is latent convex. The latent parameters are interfaced via a set of control parameters through a pre-configured neural network which acts as the representation map $\chi(\theta)$.

Latent Linear Model: The first model discussed is the hidden linear model which minimizes the following non-convex penalty.

$$L(\theta;\Omega) = \|y - X\chi(\theta)\|^2,\ \omega = (y, \boldsymbol{x}^T)^T$$

The data points are $\phi-$mixing and generated via a autoregressive process. In each experiment, there are 200 bootstrap samples, 10000 training steps, and 500 trials. The learning rate $\gamma_t = (t+10)^{-0.66}$, and the batch size $B_t = t^{0.3}$, and confidence intervals are made at the 95% significance level, the joint parameter estimate capturing the true model 90% of the time. The estimates can be seen below.



(a) Distribution of SHADE

(b) Distribution of PR

Figure 1: Parameter Distribution of Estimators in Latent Linear Model

We plot the empirical parameter distributions of the models in Figure 1. The SHADE-based estimator has a tighter empirical variance, hinting at increased asymptotic efficiency.

Detection of Fake LLM Texts: As an example of the efficacy of our methods to real-world problems, we consider the scenario of determining whether academic text was artificially generated from a large-language model, i.e. ChatGPT etc. To this end, we utilize the dataset from which contains thousands of examples of machine-generated and authentic papers from scientific fields. To simplify the analysis, the abstract and introduction sections are encoded into a vector embedding in $\mathbb{R}^{384}$, which is then passed into a latent logistic regression.

$$L(\theta;\Omega) = \log(1 + \exp(-y \cdot \psi(x) \cdot \chi(\theta)))$$

When the control parameters lie in $\mathbb{R}^d$, then we see that estimators based on the SHADE paradigm, completely outclass those using traditional methods on this non-traditional loss function.

Table 1: Accuracy of artificial text detection.

d	SGD Acc.	SHADE Acc.
25	0.550	0.535
50	0.560	0.580
100	0.588	0.622

6. Conclusion

This paper proposed a novel statistical estimator that is asymptotically consistent and normal, under a non-convex penalty and online, mixing data. This estimator synergizes the Polyak-Ruppert averaging scheme, mini-batch sampling and conditioned gradient descent and is successful despite the demands of modern machine learning systems. We furthermore demonstrated its effectiveness in a few real-world scenarios. These results emerge from the interplay of non-convex optimization, online learning and statistical estimation and open the door for future work.

References

[1] 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.

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

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

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

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

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

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

[8] 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.

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

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

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

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

[13] 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.

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

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

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

[17] Arbaaz Khan, Ekaterina Tolstaya, Alejandro Ribeiro, and Vijay Kumar. Graph policy gradients for large scale robot control. In *Conference on robot learning*, pages 823–834. PMLR, 2020.

[18] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[19] 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.

[20] 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.

[21] John M Lee. Introduction to smooth manifolds, 2003.

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

[23] James Martens. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1):5776–5851, 2020.

[24] 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.

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

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

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

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

[29] 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.

[30] 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.

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

[32] 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.

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

[34] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Appendix A. Background

A.1. Motivating examples

We begin the appendix with a collection of examples demonstrating the ubiquity of the latent convex objectives.

Example A.1: Consider the estimation problem of predicting a noisy binary response variable $y \in \{0, 1\}$ given data w. To simplify the analysis, we assume y_i generated by the process $y_i = \sigma(\theta^* x_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau^2)$, where σ is the sigmoid function. Minimizing the expectation of the squared logistic loss yields the following.

$$\ell(\theta) = \mathbb{E}[(y_i - \sigma(\theta x_i))^2] = \mathbb{E}[\big(y_i - (1 + \exp(-\theta x_i))^{-2}\big)^2]$$

Expanding out this objective in terms of the definition of y and the sigmoid function yields the following expression.

$$\ell(\theta) = \mathbb{E}[\big((1 + \exp(-\theta x_i))^{-2} - (1 + \exp(-\theta x_i))^{-2}\big)^2] + \tau^2$$

Taking $\chi(\theta) = \log(\theta)$, yields the following simplification.

$$\ell(\theta) = \mathbb{E}[\big((1 + (\theta^*)^{-x_i})^{-2} - (1 + (\theta)^{-x_i})^{-2}\big)^2] + \tau^2$$

This simplifies to a sum of moment generating function of normal random variables, which is indeed a convex function.

Example A.2: We now turn to a more complex example in convex reinforcement learning. Its standard framework hinges on a Markov decision process, $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, H, \rho, \gamma)$, where $\mathcal{S}, \mathcal{A}$ represent the state and action spaces respectively, $\mathcal{P}$ the state transition kernel returning a distribution over states, ρ the initial state distribution, and γ the time discount factor. Our goal is to ascribe actions to a distribution over states, through a policy function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which induces a probability measure over states and actions $\mathbb{P}_{\rho,\pi}(s_t, a_t)$ given an initial distribution. By defining a *state-occupancy measure* as follows:

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

and denoting $\mathcal{U} = \{\lambda^\pi; \pi \in \Pi\}$ to be the set of state-occupancy measures, our goal is to minimize

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

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

As a final related example, we show the applicability of the Jacobian condition within the context of neural networks.

Proposition A.1 *Suppose h is a shallow one-layer hidden neural network taking d_0 input nodes to d_1 hidden nodes to d_2 output nodes.*

$$h : \mathbb{R}^{d_0} \to \mathbb{R}^{d_2},\ h(x) = V\phi(Wx)$$

where ϕ is twice-differentiable activation function (i.e. Sigmoid, GELU). Then for all $w^* \in \mathbb{B}(w_0, \rho_\phi)$

$$\sigma_{\min}(\phi(W_0X)) \leq \sigma_{\min}(\nabla\, \phi^*(W0)) \leq \sigma_{\max}(\nabla\, \phi^*(W_0)) \leq \sup_a |\dot{\phi}(a)|\sigma_{\max}(X)\sigma_{\max}(V) + \sigma_{\min}(\phi(W_0X))$$

where

$$\rho_\phi = \frac{\sigma_{\min}(\phi(W_0X))}{2\sqrt{2}\sigma_{\max}(X)(\sup_a \sigma_{\min}|\dot{\phi}(a)| + \sup_{a,V} |\ddot{\phi}(a)|\sigma_{\max}(V))}$$

As can be seen, proposition shows that within a certain radius, the Jacobian of shallow neural networks have bounded spectra, implying our condition above captures this rich problem set.

A.2. Background in dynamical systems

Our analysis combines tools from dynamical systems, probability theory, and stochastic algorithms. To this end, we begin by laying out a useful introduction to the different versions of stability. We then recall the celebrated theorem of Lyapunov and conclude by introducing a Lyapunov function tailored for the preconditioned hidden gradient dynamics.

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

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

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

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

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

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

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

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

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

Theorem A.2 *Let $\mathbf{x} = 0$ be a fixed point for Equation A.1, and let $D \subset \mathbb{R}^n$ contain 0. Let $V : D \to \mathbb{R}$ be a continuously differentiable function such that*

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

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

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

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

In Sakos et al. [32], the authors demonstrated the L_2 energy function satisfies the conditions of the $V(\cdot)$ function above.

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

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

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

Proof This is Proposition 1 in Sakos et al. [32]. ■

A.3. Notation

We now introduce the notation used in the remainder of the appendix. A key step in the analysis of the asymptotic behaviour of estimators, is decomposing the gradient of the latent convex loss $\ell(\theta) = f(\chi(\theta))$ into the sum of three non-negative martingales. This facilitates the creation of robust second-moment bounds, and is similar to the framework created by Liu et al. [22], Polyak and Juditsky [27].

Both preconditioned hidden gradient descent and its bootstrap counterpart are subsumed by the following descent process.

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

The random variables U_t satisfy the following assumption.

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

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

Definition A.3 *In the proceeding sections, we use the following notation for different segments of the gradient.*

- $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\Omega_1, ..., \Omega_t] = \mathbb{E}[\cdot|\mathcal{D}_t]$
- $h(x) = \nabla f(x) := \mathbb{E}[\nabla f(x; \Omega)]$
- $e_t := \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \Omega_t)] - \nabla f(x_{t-1})$
- $\zeta_t := U_t \nabla f(x_{t-1}; \Omega_t) - \mathbb{E}_{t-1}[\nabla f(x_{t-1}; \Omega_t)]$
- $\hat{g}_t := \nabla f(x_{t-1}; \Omega_t) = \frac{1}{|\Omega_t|}\sum_{s\in\Omega_t} \nabla f(x_{t-1}; \omega_s)$
- $V_t := \mathbf{J}(\theta_t)^T(h(\chi(\theta_t)) + e_t + \zeta_t) = \nabla L(\theta_{t-1}; \omega_t)$
- $G := \nabla^2 f(x^*)$*, the Hessian of χ at θ^**

Using this notation, we rewrite the descent procedure as the following.

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

The SHADE estimator, is built off of the iterates of two PHGD processes θ_t^a, θ_t^b utilizing non-overlapping data blocks Ω_t^a, Ω_t^b. We define $\hat{g}_t^a, \hat{g}_t^b = \nabla f(x_{t-1}^a; \Omega_t^a), \nabla f(x_{t-1}^b; \Omega_t^b)$, respectively. The averaging estimators

$$\hat{\theta}^a, \hat{\theta}^b = T^{-1} \sum_{t=1}^{T} \theta_t^a, T^{-1} \sum_{t=1}^{T} \theta_t^b$$

serve similar functions within the SHADE estimator, so without loss of generality it is sufficient to study $\hat{\theta}^a$. For ease of notation, we omit the superscripts unless necessary in the sequel.

A.4. Background on PHGD

We now motivate the development of PHGD, discussing the gradient flow from which it was developed, and its connection to Riemannian geometry. As mentioned previously, the convergence rates of stochastic gradient descent on latent convex functions often lag behind those on conventional convex functions [12]. To mitigate this issue and accelerate the convergence rate of traditional gradient descent in non-convex loss functions, we introduce the notion of *natural gradient flows*. Given a function $\ell : \Theta \to \mathbb{R}$, and a class of positive semi-definite matrices, $\mathbf{P}_\theta, \theta \in \Theta$, the natural gradient flow is given by the steepest descent in the geometry induced by the matrix $\mathbf{P}_\theta$.

$$\dot{\theta} = -\mathbf{P}_\theta^{-1} \nabla \ell(\theta) \qquad \text{(Natural Gradient Flow)}$$

Some examples of these preconditioning matrices include $\mathbf{P}_\theta = I$, yielding standard gradient flow, and the Fisher information matrix, which flows in the steepest direction based on the metric induced by the Fisher information[1] within the space of distributions. [23]

$$\mathbf{P}_\theta^{\text{Fisher}} := -\mathbb{E}_{x \sim p(\theta)}[\nabla^2 \log(p_\theta(x))] \qquad \text{(Fisher matrix)}$$

The hidden nature of latent convex functions leads to an inherent geometry induced by the representation mapping. Indeed, by conditioning the gradient relative to the Riemmaninan metric $g_{ij} = \delta_{ij} \chi_i'(\theta_i)$, Sakos et al. [32] propose the preconditioned hidden gradient flow.

$$\mathbf{P}_\theta^{\text{Hidden}} := [\mathbf{J}(\theta)^T \mathbf{J}(\theta)]^+ \qquad \text{(Hidden Preconditioner)}$$

This flow has numerous convergence guarantees that exceed that of traditional gradient descent, including increased sample complexity efficiency on latent objectives. With this in mind, we will use the flow's discretization as the means of optimizing latent convex functions. For ease of notation, we denote P_θ^{Hidden} as $P(\theta)$.

$$\theta_t = \theta_{t-1} - \gamma_{t-1} \mathbf{P}(\theta_{t-1}) \nabla L(\theta_{t-1}; \Omega_t)$$

As a show of the improved convergence of PHGD, we present the following proposition.

Proposition A.2 (Sample Complexity of Stochastic Descent in Hidden Convex Objectives) *Let $\ell(\theta) = f(\chi(\theta))$ be a merely latent convex function , i.e. $\nabla^2 f(x) = 0$. Fix $\epsilon > 0$ and set the step size $\gamma_t = (t + \gamma_0)^{-\rho}$, $\rho \in (0.5, 1]$. Then define the L^2 energy Lyapunov function,*

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

Then the following statments hold.

[1] For more information about the Riemannian metric induced by the Fisher information in parametric settings see [2, 21]

- *If θ is an iterate of preconditioned hidden gradient descent, $E_T \leq \epsilon$ after $T = \tilde{\mathcal{O}}(\epsilon^{-2})$*
- *If θ is an iterate of stochastic gradient descent, $E_T \leq \epsilon$ after $T = \tilde{\mathcal{O}}(\epsilon^{-3})$*

Proof In preparation for the subsequent steps we define

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

In their recent review of the role of stochastic gradient descent in latent convex objectives Fatkhullin et al. [12] showed that when $T = \tilde{\mathcal{O}}(\frac{\beta D_{\mathcal{U}}}{\kappa^2}\frac{1}{\epsilon} + \frac{\beta D_{\mathcal{U}} M^2}{\kappa^4}\frac{1}{\epsilon^3})$, $\mathbb{E}[\Lambda_T] < \epsilon$. Here, κ is the Lipschitz coefficient of the latent map χ, and $D_{\mathcal{U}}$ is a constant stemming from the fact that $f(\cdot)$ is strongly convex and thus satisfies the Kurdyka-Lojasiewicz (KL) condition [15].

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

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

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

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

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

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

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

The affine function $g(x) = x - \hat{x}$, indeed satisfies the condition of being merely convex. Thus now turning attention to PHGD, through Theorem 1 of Sakos et al. [32] demonstrated in the hidden merely convex case that the averaged iterate satisfies the following convergence rate

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

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

A.5. Moment inequalities

We now introduce a couple of key moment inequalities for $\phi-$mixing sequences, unlocked by the restrictions made on the $\phi-$mixing coefficients, that will be rendered useful when establishing the asymptotic consistency of our estimators. of martingale iterates.

Lemma A.2 (Moment Inequality) *Let $\{X_t\}_{t=1}^{\infty}$ be a mean-zero stationary sequence with ϕ-mixing coefficients bounded by a function $\phi(t)$, where $\sum_{t=1}^{\infty}\sqrt{\phi(t)} < \infty$ 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 [36] ■

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

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

Proof This moment inequality is Lemma S.5 from Liu et al. [22] ■

We recall the template inequality from Sakos et al. [32], that characterizes the evolution of the L_2 energy Lyapunov function, $E(\theta; x^*)$ throughout the preconditioned hidden gradient descent process. This allows us to relate the model error in the control space with that in the latent space and plays a key role in our analysis of the asymptotic behaviour of the iterate sequence.

Proposition A.3 (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 $E_t = E(\theta_t; x^*) = \frac{1}{2}\|\chi(\theta_t) - x^*\|^2$:*

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

where $\alpha_t = \sum_{i \leq t} \langle \mathbf{J}(\theta_t)^T P(\theta_t) V_t - h(x), x_t - \tilde{x}\rangle$ and ψ_t is a random error seqeunce with $\sup_t \mathbb{E}[\psi_t] < \infty$.

Proof This is an extension of the template descent inequality from Sakos et al. [32], which can be achieved via linearity in argument. This claim is created by creating bounds on the Taylor expanded potential function. ■

A.6. Remarks about $\phi-$mixing

In this short section, we recite a proof of Davydov [7] that demonstrates that the $\phi-$mixing condition subsumes the strong (α) mixing condition, which characterizes Markov dependence.

We first prove a proposition exemplifying the power of the $\phi-$mixing condition.

Proposition A.4 (Markov Chains are $\phi-$mixing (Bradley [3] Theorem 3.3)) *Let $X = (X_k)_{k=0}^{\infty}$ be an ergodic and aperiodic (not necessarily stationary) Markov chain over a state space of $\mathbb{R}$. Then if $\phi_\Omega(n) < 1/2$ for some n , then $\phi_\Omega(n) \to 0$ at least exponentially fast as $n \to \infty$, implying that X is a $\phi-$ mixing sequence.*

Proof The broad class of Markov chains, even with not necessarily countable state space, satisfy the mixing condition above. The proof is contained in Davydov [7] and is a variation on the result of Doeblin (1937). ■

We conclude with a few additional assumptions on the $\phi-$mixing sequence for the analysis to come.

Assumption A.2 *Along with the requirements in the descent parameter, we enforce a few other rate limiting conditions.*

$$t^{\rho} = o\left(\sum_{j=1}^{t} B_j^{-1}\right), \sum_{j=1}^{t} \phi^{1/2-1/p}(B_j) = o\left(\sqrt{\sum_{j=1}^{t} B_j^{-1}}\right)$$

Appendix B. Proofs of asymptotic consistency

In this section, our goal is to prove Theorem 2 which we restate below for convenience.

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

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

Our proof strategy is comprised of the following steps.

1. In Lemma B.1, we show the second moment of these martingale terms are then bounded through convex analysis; in turn, in Lemma B.2 we combine these claims to establish a bound on the entire second moment in terms of the $\phi-$mixing coefficients and the energy function.
2. From these bounds and the Lyapunov properties of the energy function, the Robbins-Siegmund theorem [30] implies the consistency of the averaged latent iterates.
3. Applying ideas from real analysis, we characterize the degree of distortion induced by traveling through the representation mapping in Lemma One, allowing the consistency of SHADE to be shown.

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

B.1. Martingale bounds

As a starting point, we instantiate a second moment bound on the different components of the gradient of our loss. Wrestling with the difficulty of a latent convex objective, we instantiate these bounds in the latent space $\mathcal{X}$, as opposed to the control space Θ, unlocking the use of the machinery of convex functions. These work in unison to create a second moment bound of the empirical gradient $\nabla L(\theta; \omega)$, in terms of distance to the true model.

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

(i) $\mathbb{E}\,\|e_t\|^2 \leq \phi^{1-2/p}(B_{t-1})v_t$
(ii) $\mathbb{E}\,\|\hat{g}_t - h(x_t)\|^2 \leq \frac{C}{B_t}(1+\|x_t - x^*\|^2)$
(iii) $\mathbb{E}_t(\|\zeta_t\|^2) \leq \phi^{1-2/p}(B_{t-1})v_t + \frac{C}{B_t}(1+\|x_{t-1}-x^*\|^2)$

Proof (i.) Define $\tilde{e}_t$ 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. By virtue of its independence to the original data sequence, this value compresses into mean zero Gaussian noise. Using Lemma A.3,

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

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

m being arbitrary, we define it $\phi^{-2/p}(B_t)$ yielding an inequality in terms of the $\phi-$mixing sequence terms.

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

The p-th moment of e_t can be bounded as follows.

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

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

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

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

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

We define v_{1t} below, and see it has finite mean.

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

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

The bound in Lemma A.3 completes the claim. ■

Proof (ii.) The definition of $\hat{g}_t$ found above implies

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

The term $\nabla f(x_t;\Omega_t)$ satisfies the martingale condition of Lemma A.2, and so we conclude that

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

■

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

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

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

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

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

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

Moreover, from the definition of $\tilde{\zeta}_t$, we obtain

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

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

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

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

We now present a descent equality from the viewpoint of the latent iterates $x_t = \chi(\theta_t)$ by way of Taylor's theorem. The result closely resembles traditional stochastic gradient descent.

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

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

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

Importantly, the asymptotic results show that the SHADE estimator and its bootstrap counterparts are (i.) *consistent* and (ii.) asymptotically normal, despite the impediment of mixing data. This leads to the creation of robust estimators and confidence intervals in practical settings, which we explore in the next section.

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. ■

With this descent lemma, we can bound the second moment of the energy function in terms of itself and other gradient terms.

Lemma B.3 *We now show the following bound, where* $\Delta_t = x_t - x^*$.

$$\begin{aligned}\mathbb{E}_{t-1}[\|\Delta_t\|^2] \leq{}& (1 + C\gamma_t^2 + CB_t^{-1}\gamma_t^2)\|\Delta_t\|^2 \\ &+ C\gamma_t^2(1 + B_t-1) + 4\gamma_t^2\phi^{1-2/p}(B_{t-1})v_t + 4C\gamma_t\sqrt{\phi^{1-2/p}(B_t)v_t} \\ &+ 2\gamma_t^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{CB_t^{-1}(1 + \|\Delta_t\|^2)} \\ &+ 2\gamma^3\mathbf{J}(\theta_t)V_t\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2) + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4)\end{aligned}$$

Proof From Lemma 2, we see that

$$\|\Delta_t\|^2 = \|\Delta_t - \gamma_t(h(x_t) + e_t + \zeta_t) + \mathcal{O}(\|\theta_t - \theta_{t-1}\|^2)\|^2$$

Thus expanding this out

$$\begin{aligned}\|\Delta_t\|^2 &= \|\Delta_t - \gamma_t(h(x_t) + e_t + \zeta_t) + \gamma_t^2\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2)\|^2 \\ &= \|\Delta_{t-1}\|^2 + \gamma^2\|h(x_t)\|^2 + \gamma^2\|e_t\|^2 + \gamma^2\|\zeta_t\|^2 + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4) \\ &\quad - 2\gamma_t\Delta_{t-1}^T h(x_t) - 2\gamma_t\Delta_{t-1}^T e_t - 2\gamma_t\Delta_{t-1}^T\zeta_t \\ &\quad + 2\gamma_t^2 h(x_t)^T e_t + 2\gamma_t^2 h(x_t)^T\zeta_t + 2\gamma_t^2 e_t^T\zeta_t \\ &\quad + 2\gamma_t^3(h(x_t) + e_t + \zeta_t)\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2) \\ &\leq \|\Delta_{t-1}\|^2 + \gamma_t^2\|h(x_t)\|^2 + \gamma_t^2\|e_t\|^2 + \gamma_t^2\|\zeta_t\|^2 + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4) \\ &\quad - 2\gamma_t\Delta_{t-1}^T h(x_t) - 2\gamma_t\|\Delta_{t-1}\|\|e_t\| - 2\gamma_t\|\Delta_{t-1}\|\|\zeta_t\| \\ &\quad + 2\gamma_t^2\|h(x_t)^T\|\|e_t\| + 2\gamma_t^2\|h(x_t)\|\|\zeta_t\| + 2\gamma_t^2\|e_t\|\|\zeta_t\| \\ &\quad + 2\gamma_t^3(h(x_t) + e_t + \zeta_t)\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2)\end{aligned}$$

The conditional expectation of this expression given the data $\mathcal{D}_t := \bigcup_{k=1}^t(\Omega_k)$ yields the following.

$$\begin{aligned}\mathbb{E}_{t-1}[\|\Delta_t\|^2] \leq{}& \|\Delta_{t-1}\|^2 + \gamma_t^2\|h(x_t)\|^2 + \gamma_t^2\|e_t\|^2 + \gamma_t^2\,\mathbb{E}_{t-1}[\|\zeta_t\|^2] + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4) \\ &- 2\gamma_t\Delta_{t-1}^T h(x_t) - 2\gamma_t\|\Delta_{t-1}\|\|e_t\| \\ &+ 2\gamma_t^2\|h(x_t)^T\|\|e_t\| + 2\gamma_t^2\|e_t\|\,\mathbb{E}_{t-1}[\|\zeta_t\|] \\ &+ 2\gamma_t^3(h(x_t) + e_t + \mathbb{E}_{t-1}[\zeta_t])\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2)\end{aligned}$$

Using Lemma B.1, we can transform the inequality into the following.

$$\begin{aligned}\mathbb{E}_{t-1}[\|\Delta_t\|^2] &\leq \|\Delta_{t-1}\|^2 + \gamma_t^2 C(1 + \|\Delta_{t-1}\|^2) - 2\gamma_t C^{-1}\|\Delta_{t-1}\|^2 + 4\gamma_t^2\phi^{1-2/p}(B_{t-1})v_t \\ &\quad + C\gamma_t^2 B_t^{-1}(1 + \|\Delta_{t-1}\|^2) + 2C\gamma_t\sqrt{\phi^{1-2/p}(B_t)v_t} + 2C\gamma_t\sqrt{\phi^{1-2/p}(B_t)v_t} \\ &\quad + 2\gamma_t^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{\phi^{1-2/p}(B_t)v_t + \frac{C}{B_t}(1 + \|\Delta_t\|^2)} \\ &\quad + 2\gamma_t^3\mathbf{J}(\theta_t)V_t\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2) + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4) \\ &\leq (1 + C\gamma_t^2 + CB_t^{-1}\gamma_t^2)\|\Delta_t\|^2 \\ &\quad + C\gamma_t^2(1 + B_t-1) + 4\gamma_t^2\phi^{1-2/p}(B_{t-1})v_t + 4C\gamma_t\sqrt{\phi^{1-2/p}(B_t)v_t} \\ &\quad + 2\gamma_t^2\sqrt{\phi^{1-2/p}(B_t)v_t}\sqrt{CB_t^{-1}(1 + \|\Delta_t\|^2)} \\ &\quad + 2\gamma_t^3\mathbf{J}(\theta_t)V_t\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^2) + \gamma_t^4\,\mathcal{O}(\|\theta_t - \theta_{t-1}\|^4)\end{aligned}$$

■

B.2. Consistency preliminaries and proof

We now prove the first requirement of an estimator to conduct statistical estimation, consistency. Given the bound generated in the previous parts, and the rate limitations established in the data descent parameters, we demonstrate that the energy function converges to zero through an appeal to the Robbins-Siegmund theorem and the PHGD template inequality. As a consequence, the estimator built off of the latent iterates $\hat{x}_T = T^{-1}\sum_{t=1}^{T} x_t$ is asymptotically consistent. We then demonstrate a duality gap in Lemma B.4, that quantifies the distortion between the control space Θ and latent space $\mathcal{X}$ induced by the representation mapping. This allows us to extend the consistency result to the control space Θ, the SHADE estimator $\hat{\theta}$.

We begin by introducing the primary vehicle by which we demonstrate consistency, the Robbins-Siegmund theorem.

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

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

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 $V_n \xrightarrow{a.s.} V_\infty$ and $\sup_{n\geq 0}\mathbb{E}[V_n] < \infty$.

Proof The proof is found in Robbins and Siegmund [30]. ■

With a suitable function Lyapunov function $V(\cdot)$, a stochastic algorithm that satisfies the above inequality, can be shown to be convergent.

The gradient sub-elements e_t and ζ_t are martingale terms. When analyzing the preconditioned hidden gradient descent process, restated below for convenience, we can apply the Robbins-Siegmund theorem to demonstrate consistency.

$$\theta_{t+1} = \theta_t - \gamma_{t+1} V_t = \theta_t - \gamma_{t+1}\mathbf{J}(\theta_t)(h(\theta_t) + e_{t+1} + \zeta_{t+1}).$$

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

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

Proof This proof is an adaptation of Theorem 5.3 in [19]. ■

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

Lemma B.4 (Representation Gap) *Given Main Assumption (iii.), where $\sigma_{\min}, \sigma_{\max}$ are the smallest and largest singular values of the latent mapping χ respectively.*

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

This lemma controls the 'distortion' of the distance metric between the latent and control spaces by the representation mapping and performs a key role in the analysis.

Proof

We first prove the upper bound, implying to showing that the representation mapping χ is $\sigma_{max}-$Lipschitz. Given a vector v with norm equal to one, from the definition of a Jacobian the following statement holds.

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

For all ε, there exists a δ such that if $t < \delta$,

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

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

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

The above holds for all v on the unit ball and if we replace the value tv with $\theta^* - \theta$ where $\|\theta - \theta^*\| < \delta$, we conclude

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

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

With this lemma in hand, we can finally prove the consistency of SHADE.

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

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

Proof Recall Proposition A.1.

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

Note because f is strongly convex, from the definition of strong convexity we see

$$\gamma_t \langle h(x_t), x_t - x^*\rangle \geq \mu\|x_t - x^*\| = \mu E_t.$$

Thus the above inequality can be simplified as follows.

$$E_{t+1} \leq (1 - 2\mu\gamma)E_t + \gamma_t \alpha_t + \gamma_t^2 \psi_t \tag{B.20}$$

The α_t term in the inequality, given the data $\mathcal{D}_t$, has mean zero.

$$\begin{aligned}
\mathbb{E}_t[\alpha_t] &= \mathbb{E}_t[\sum_{t\leq T} \langle \mathbf{J}(\theta_t)^T \mathbf{P}(\theta_t) V_t - h(x_t), x_t - x^* \rangle | \mathcal{D}] \\
&= \sum_{t\leq T} (\mathbf{J}(\theta_t)^T \mathbf{P}(\theta_t)\, \mathbb{E}[V_t|\mathcal{D}_t] - h(x_t))^T (x_t - x^*) \\
&= \sum_{t\leq T} (\mathbf{J}(\theta_t)^T \mathbf{P}(\theta_t) \mathbf{J}(\theta_t)^T h(x_t) - h(x_t))^T (x_t - x^*) \\
&= \sum_{t\leq T} (h(x_t) - h(x_t))^T (x_t - x^*) \\
&= 0
\end{aligned}$$

Taking expectations, we claim that.

$$\begin{aligned}
\mathbb{E}_t[E_{t+1}] &\leq \mathbb{E}_t[(1 - 2\mu\gamma) E_t + \gamma_t \alpha_t + \gamma_t^2 \psi_t] \\
&= (1 - 2\gamma\mu) E_t + \gamma_t^2\, \mathbb{E}_t[\psi_t] \\
&\leq (1/2 - \gamma\mu)((1 + C\gamma_t^2 + C B_t^{-1} \gamma_t^2) \|\Delta_t\|^2 \\
&\quad + C\gamma_t^2 (1 + B_t {-} 1) + 4\gamma_t^2 \phi^{1-2/p}(B_{t-1}) v_t + 4C\gamma_t \sqrt{\phi^{1-2/p}(B_t) v_t} \\
&\quad + 2\gamma_t^2 \sqrt{\phi^{1-2/p}(B_t) v_t} \sqrt{C B_t^{-1}(1 + \|\Delta_t\|^2)} \\
&\quad + 2\gamma^3 \mathbf{J}(\theta_t) V_t\, \mathcal{O}(\|\theta_t - \theta_{t-1}\|^2) + \gamma_t^4\, \mathcal{O}(\|\theta_t - \theta_{t-1}\|^4))
\end{aligned}$$

Because $\sum_{t\geq 1} \gamma_d^2 < \infty, d > 2$ and $\sum_{t\geq 1} \phi^{1-2/p} < \infty$, the terms with these coefficients, are asymptotically zero. Thus we can apply the Robbins-Siegmund theorem to show that E_t converges to zero as $t \to \infty$. To show convergence of θ note that from Lemma B.4

$$\sigma_{\min} \|\theta_t - \theta^*\| \leq \|\Delta_t\| \leq \sigma_{\max} \|\theta_t - \theta^*\| \tag{B.21}$$

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

Appendix C. Proofs of asymptotic normality

In this section we establish the asymptotic normality of SHADE and its bootstrap counterpart. For convenience, we restate these theorems below. For ease of notation, we define $G = \nabla^2 f(x^*)$ to be the Hessian matrix at the optimal state $\chi(\theta^*)$.

Theorem 3 (Asymptotic Normality of SHADE) *Given the assumptions above, the following theorem occurs*

$$\frac{T}{\sqrt{\sum_{t=1}^T B_t^{-1}}} (\hat{\theta}_T - \theta^*) \xrightarrow{d.} \mathcal{N}(0, \Sigma)$$

where $r(t) = \mathbb{E}[\nabla L(\theta_{k+t}; \Omega_{k+t}) \nabla L(\theta_k; \Omega_k)^T]$, $H = \nabla^2 f(x^*) \mathbf{J}(\theta^*)$, $V = (2r(0) + 4\sum_{k\geq 1} r(k))$, *and* $\Sigma = H^+ V [H^+]^T$.

Theorem 4 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

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

where $\mathcal{D} = \{\omega_i | i \in \Omega_t^a \cup \Omega_t^b\}$ and represents the data used in the empirical risk minimization process, and $\hat{\Sigma} = \Sigma$ is the covariance matrix in Theorem 4

This section establishes the second criterion needed to reliably use empirical risk minimization, that of asymptotic normality. Here we leverage the strongly convex geometry of the latent space to simplify analysis. We additionally, analyze the behavior of bootstrap PHGD and show that its asymptotic distribution is identical to PHGD. To this end, we use the following arguments.

1. In Lemma C.1-C.2 we equate the time-averaged parameter estimates with scaled gradients of the loss function. This is similar to the strategy undertaken by Polyak [25] and Liu et al. [22]. To achieve this, we exploit the strongly convex nature of the objective in the latent space to create sharp bounds.
2. In Lemma C.3, we show that the sum of the autocorrelation coefficients is finite and conclude via the Lindeberg condition that the sum of gradients exhibits asymptotic normality.
3. In Theorem 3, the bootstrap estimates are shown to match the distribution of PHGD. Leveraging the requirements in Assumption 3 and a theorem from Kuczmaszewka, we can apply Theorem 2 to achieve the desired result.

C.1. Establishment of asymptotic normality

Thus far our analysis has dwelt on bounding terms of the martingale loss gradient. To proceed, we establish a correspondence between these expression and the parameter estimates through supplementary lemmas found in Polyak [25]. By unrolling the iteration sequence into a product of terms dictated by the Hessian, we extend the gradient bounds to the iteration parameters.

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

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

$$D_j^j = I$$

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

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

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* $\sum_{j\geq 0} a_j = \infty$, *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 [26], (iii) can be found in Chen et al. [5] and (iv) is in Liu et al. [22]. ■

Lemma C.1 *Under the assmuptions above, 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. [22]. ■

Lemma C.2 (Lemma S.7 [22]) *Suppose* a_n *is a decreasing sequence with* $\lim_{n\to\infty} a_n = 0$, *and* b_n *is a sequence, such that* $\sum_{n=1}^{\infty}|b_n| < \infty$. *If* $\sum_{n=1}^{\infty} a_n = \infty$, *then*

$$\lim_{T\to\infty}\sum_{n=1}^{T} a_n \sum_{k=n}^{T} b_k / (\sum_{n=1}^{T} a_n) = \lim_{T\to\infty}\sum_{n=1}^{T} a_n b_n / (\sum_{n=1}^{T} a_n) = 0$$

Proof This fact is found in [22] ■

Lemma C.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.

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

Defining $\Delta_t = (x_t - x^*)$ we can see that

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

The average of this expression can be rewritten as.

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

We see from Proposition C.1 that S_1 is asymptotically $o_P(\frac{\sqrt{\sum_{j=1}^{T}B_j^{-1}}}{T})$. S_2 is bounded as follows. Define $Q_j^T = \sum_{t=j}^{T}[\prod_{k=j+1}^{t}(I-\gamma_k G)]\gamma_j$, Proposition C.1 again implies that $\|Q_j^T\| < C$. Then applying Lemma B.1, we have that

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

Taylor expanding the above yields.

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

This implies a bound for S_4.

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

The consistency of our estimator implies $\Delta_t \xrightarrow{a.s.} 0$, thus the set of all j such that $\|\Delta_j\|^2 < \delta$ is finite almost surely, implying the second term of the sequence will be dominated by the $\frac{1}{T}$ term.

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

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

■

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

We now recall a result from Fan and Yao [10], which ensures the autocorrelation coefficients do not grow too large in a $\phi-$mixing sequence.

Lemma C.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 [10], 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 [3], we can conclude the desired claim. ■

The primary vehicle for demonstrating asymptotic normality is the celebrated Lindeberg condition.

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

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

Then

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

Proof The proof can be found in Brown [4]. ■

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

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

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

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

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

We now aim to bound these expressions via our assumptions on the $\phi-$mixing nature of these sequences. When $t > k + 1$, the difference of the indices between Ω_t^a and Ω_k^b are at least B_{t-1} apart. Thus we can use Lemma B.1 to see that

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

An analogous result holds for when $k > t + 1$

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

Combining the above, we conclude.

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

Thus using Lemma C.2 and the requirements of the $\phi-$mixing sequence in the Descent Parameters, we see that this term is $o_P(\sum_{t=1}^{T} B_t^{-1})$.

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

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

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

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

Then we see via Lemma C.2 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

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

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

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

Putting all of these pieces together, we define the quantity $V_{T,t} = (\hat{g}_t^{a*}+\hat{g}_t^{b*})/\sqrt{\sum_{k=1}^{T}B_k^{-1}}$. Above, we have just shown that

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

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

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

We get this via Markov's inequality. Then using the definition of strong convexity, and the assumptions on our loss, we arrive at the following.

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

Thus combining this all together we satisfy the Lindeberg condition.

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

The final equality is due to lemma. ■

To establish Theorem 2, we recall a useful lemma.

Lemma C.6 *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 [16]. ■

Theorem 3 (Asymptotic Normality of SHADE) *Given the assumptions above, the following theorem occurs*

$$\frac{T}{\sqrt{\sum_{t=1}^{T}B_t^{-1}}}(\hat{\theta}_T-\theta^*)\xrightarrow{d.}\mathcal{N}(0,\Sigma)$$

where $r(t)=\mathbb{E}[\nabla L(\theta_{k+t};\Omega_{k+t})\nabla L(\theta_k;\Omega_k)^T]$, $H=\nabla^2 f(x^*)\mathbf{J}(\theta^*)$, $V=(2r(0)+4\sum_{k\ge1}r(k))$, *and* $\Sigma=H^+V[H^+]^T$.

Proof Given lemma B.4, we have that

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

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

Then applying Lemma C.3 we can conclude.

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

Applying the delta method results in the desired claim. ■

Proof This proof can be shown using the Central Limit theorem and Taylor's theorem. A full proof can be found [16]. ■

C.2. Bootstrap normality

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

Theorem 4 (Bootstrap Normality) *Suppose the assumptions in part 3 hold. Then*

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

where $\mathcal{D} = \{\omega_i | i \in \Omega_t^a \cup \Omega_t^b\}$ and represents the data used in the empirical risk minimization process, and $\hat{\Sigma} = \Sigma$ is the covariance matrix in Theorem 4

Proof Define $v_t = \hat{g}_t^{*i} + \hat{g}_t^{*j}$, and $\chi(\hat{\theta}) = \bar{x}_T^*$. Then we claim that

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

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

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

Moreover, we cap the variance of $v_t v_t^T$ with Lemma A.2 and the assumptions defined above, we see that

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

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

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

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

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

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

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

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

Likewise, looking at the Lindeberg condition, we can see

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

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

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

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

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

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

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

The last convergence comes from Lemma C.2. Thus we have shown the Lindeberg condition is satisfied and can conclude that $Y_t|D_t \to \mathcal{N}(0, V)$. Thus applying Theorem 4, and the delta method the claim is proven. ■

Appendix D. 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 usefulness of our methods. The section begins with a overview on how the inference is conducted, and proceeds to detail each of the applications.

In each of the proceeding examples, we define an estimation problem, in which the loss function is latent convex. The latent parameters are interfaced via a set of control parameters through a preconfigured neural network which acts as the representation map $\chi(\theta)$. Attached to the parameter estimates are confidence regions generated by bootstrap algorithm detailed below.

The dimensions of the mapping vary with the datasets, and when applicable, we provide both point estimates and confidence regions.

Latent Linear Model: The first model discussed is the hidden linear model found in Example 2.1, which minimizes the following non-convex penalty.

$$L(\theta; \Omega) = \|y - X\chi(\theta)\|^2, \omega = (y, \boldsymbol{x}^T)^T$$

Algorithm 1 Bootstrap SHADE

```
Input: Data $\{\omega_s\}_{s=1}^{\infty}$, Learning rate $\gamma_t$, Block size $B_t$
Initial parameter values $\theta_0, \theta_0^{(k)\bullet}$, for $k \in \{1, ..., n\}$
Significance level $\alpha \in [0, 1]$.
for $t = 1$ to $T$ do
    Construct two minibatches $\Omega^a$ and $\Omega^b$
    Update $\theta_t^i = \theta_{t-1} - \gamma_{t-1}\mathbf{P}(\theta_{t-1}^i)V_{t-1}$ for $i \in \{a, b\}$
    Update $\bar{\theta}_t = \frac{t-1}{t}\bar{\theta}_{t-1} + \frac{1}{2t}(\theta_t^a + \theta_t^b)$
    for $k = 1$ to $n$ do
        Generate the random weight $U_t^{(k)}$
        Update $\theta_t^{(k)\bullet} = \theta_{t-1}^{(ki)\bullet} - \gamma_{t-1}U_t\mathbf{P}(\theta_{t-1}^{(ki)\bullet})V_{t-1}$ for $i \in \{a, b\}$
        Update $\bar{\theta}_t^{(k)\bullet} = \frac{t-1}{t}\bar{\theta}_{t-1}^{(k)\bullet} + \frac{1}{2t}(\theta_t^{(ak)\bullet} + \theta_t^{(bk)\bullet})$
    end for
end for
Let $\hat{\Sigma} = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=t}^{T}(\theta_T^{(j)\bullet} - \bar{\theta}_T^{\bullet})(\theta_T^{(i)\bullet} - \bar{\theta}_T^{\bullet})^T$
Define $l_\alpha, u_\alpha$ be the $\alpha/2$, and $1 - \alpha/2$ quantiles of $\theta_T^{(1)\bullet}, ..., \theta_T^{(n)\bullet}$.
return point estimator $\theta_T$, empirical confidence interval $(l_\alpha, u_\alpha)$, and empirical covariance matrix $\hat{\Sigma}$.
```

The control variables $\theta \in \mathbb{R}^2$ are fed into the player's MLP given by the representation maps

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

where $\alpha^{(1)}, \alpha^{(2)}$ are randomly initialized matrices with values within $[-1, 1]$, and bias zero. The activation function CeLU, is given below for reference.

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

The data points are $\phi-$mixing and generated via a autoregressive process. To this end, we define an orthogonal matrix M and uniformly random diagonal matrix Λ, to construct our transform $P = M\Lambda M^T$. The covariates x_t are then created from the following vector autoregressive process, where ε_t is a standard normal random variable.

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

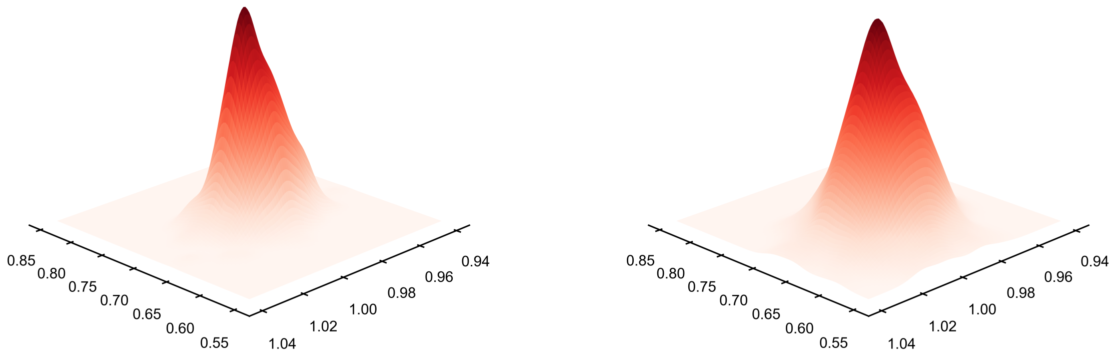
In each experiment, there are 200 bootstrap samples, 10000 training steps, and 500 trials. The learning rate $\gamma_t = (t + 10)^{-0.66}$, and the batch size $B_t = t^{0.3}$, and confidence intervals are made at the 95% significance level, the joint parameter estimate capturing the true model 90% of the time. The estimates can be seen below.

We plot the empirical parameter distributions of the models in Figure 3. The SHADE-based estimator has a tighter empirical variance, hinting at increased asymptotic efficiency.

Medical Experiments We extend the applicability of our methods to testing in medical contexts. To this end, we use the diabetes dataset from the University of California, Irvine (UCI) machine

Table 2: Latent Linear Estimates

	β_1		β_2			
	Est.	CI	Est.	CI	CP	MSE
SGD	0.6961	(0.6709, 0.7119)	0.9875	(0.9831, 0.9962)	0.872	2.701
SHADE	0.6971	(0.6732, 0.7216)	0.9878	(0.9775, 0.9996)	0.896	2.700



(*a*) Empirical Parameter Distribution of SHADE

(*b*) Empirical Parameter Distribution of Polyak-Ruppert

Figure 2: Parameter Distribution of Estimators in Latent Linear Model

learning repository [9] to predict a binary test result. We achieve this via a logistic regression augmented with a latent mapping χ.

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

The dataset covers over 250,000 patients and contains data on age, income, and other key health indicators. These scalar features are normalized into the normal interval via min-max scaling. The latent map χ in our case is a differentiable multi-layer perceptron with two hidden layers, that maps the control space $\mathbb{R}^d$ into the feature space $\mathcal{X} \subseteq \mathbb{R}^{21}$.

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

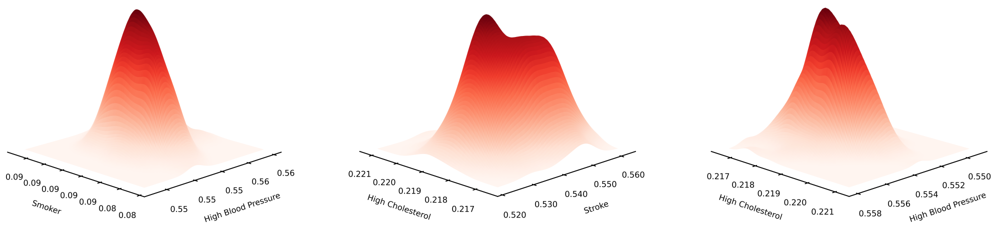
The linear mappings $\alpha^{(1)} \in [-1, 1]^{d\times 21}, \alpha^{(2)} \in [-1, 1]^{21\times 21}$ are uniformly randomly initialized. The control space is smaller than the latent space in this scenario and represents a *low-rank* encoding of the regression vector $\chi(\beta)$. The experiments are run 300 times with 10000 training steps. In accordance with the literature, the learning rate is $\gamma_t = (t + 10)^{-0.5}$ and a batch size of $B_t = t^{0.3}$.

As can be seen, both SHADE and SGD learn similarly well in low dimensional settings, as the optimal model $\theta \in \mathbb{R}^{21}$ is difficult to approximate within such a sparse representation. Yet within a more rich representation space, SHADE, powered by the information found in the Jacobian, does a better job of searching the space for the optimal parameters, bypassing the bottleneck experienced by the former.

Table 3: Accuracy of the $d-$dimensional latent linear model on diabetes detection.

d	SGD Acc.	SHADE Acc.
2	0.640	0.639
5	0.763	0.765
10	0.764	0.779

As showing a graph of the asymptotic distribution of all 21 inputs would be cumbersome, we showcase the empirical distribution of a few key parameters. To this end, we present the joint distribution of particular latent regression parameters $\chi(\theta)_i$ of interest, characterizing the interaction between the variable and the model output. This yields a multi-variate Gaussian distribution for all three pairs; the last two in particular have an anisotropic distribution, hinting at a relationship between the underlying data. We observe an anisotropic Gaussian distribution for the latter two entries, implying the two variables have a relationship.



(a) (b) (c)

Figure 3: Distribution of Latent Model Parameters Along Key Values

Detection of Fake LLM Texts: With the rise of large language models, e.g. BeRT, GPT etc., comes the issue of detecting whether text is human or machine-generated. As a show of the ubiquity of our methods, we use the SHADE parameter estimator to build a model to discriminate between academic texts and those generated by prompting ChatGPT and Galatica, under the presence of a latent convex loss, similar to the other experiments given earlier in the section. This data, given it is from the same article is highly autocorrelated and thus exemplifies the streaming demands tackled in this paper. To this end, we use the Identifying Machine Generated Scientific Papers (IDMGSP) dataset proposed by Abdalla et al. [1], which contains the abstracts, introductions, and conclusions of thousands of authentic and artificially generated papers.

To simplify our model and analysis, we transform the data from its textual representation into a vector embedding in $\mathbb{R}^{384}$, through the Sentence Transformer autoencoder architecture [29], based on the BeRT transformer autoencoder model. By reducing the dimensionality of the data from hundreds of thousands in one-hot encoded form, to mere hundreds, we are able to perform a latent logistic regression, similar to the above examples. Formally, if ψ represents the autoencoder representation mapping, this is detailed as follows.

$$L(\theta;\Omega) = \log(1 + \exp(-y \cdot \psi(x) \cdot \chi(\theta)))$$

In this case, χ is a two-layer multi-layer perceptron of that maps the control space $\mathbb{R}^d$ into the feature space $\mathcal{X} \subseteq \mathbb{R}^{384}$.

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

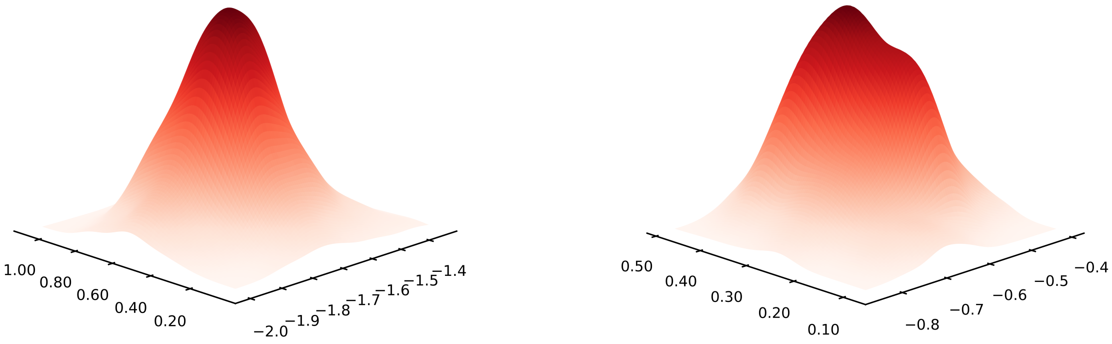
The control variables represent a *low-rank* encoding of the latent model parameters $x = \chi(\theta)$, and can be also seen as hedges against overfitting, given the model simplicity.

To the right are the accuracies of both Polyak-Ruppert averaging based-methods and SHADE, tested using a variety of control variable sizes. It is evident, that while the estimator performances are similar in smaller dimensions, where the optimal parameters are easier to search through, the complications of searching a solution space orders of magnitude larger, leave the SGD based method unable to properly comb through the parameter space, again demonstrating the power of the second-order information within the Jacobian, yielding a model that reduces the number of errors made by 15%.

Table 4: Accuracy of the $d-$dimensional latent linear model on artificial text detection.

d	SGD Acc.	SHADE Acc.
25	0.550	0.535
50	0.560	0.580
100	0.588	0.622
200	0.568	0.626

We conclude with another avenue of examination of the differences within the empirical parameter distributions created by SHADE and the Polyak-Ruppert averaging scheme. Listing point estimates and confidence intervals for each of the 384 unique parameters would prove infeasible and cumbersome, we instead, graph the distributions of the fiftieth and hundredth latent parameter $\chi(\theta)$ over the 200 runs, to elucidate their joint relationship, with the former on the left-hand axis and the latter on the right. The SHADE estimator's distribution is uni-modal multivariate Gaussian, while the Polyak-Ruppert estimator's distribution has skew, albeit with a tighter distribution overall



(*a*) SHADE Distribution

(*b*) PR Distribution

Figure 4: Distribution of Parameters Along Varying Parameters

Remarks: These experiments exemplify the advantages of the SHADE estimator over the existing framework within the realm of stream data and latent convex losses in both real world and synthetic contexts. By conditioning the gradient, the *low-rank* parameter estimates construct asymptotically consistent estimators with superior performance. The bootstrap SHADE algorithm in particular allows for the formation of accurate, balanced confidence intervals in light of the challenges presented above.