Strange springs in many dimensions: how parametric resonance can explain divergence under covariate shift.

by

Kirby Banman

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

Abstract

Most convergence guarantees for stochastic gradient descent with momentum (SGDm) rely on independently and identically ditributed (iid) data sampling. Yet, SGDm is often used outside this regime, in settings with temporally correlated inputs such as continual learning and reinforcement learning. Existing work has shown that SGDm with decaying step-size can converge under Markovian temporal correlation. In this work, we show that SGDm under covariate shift with fixed step-size can be unstable and diverge. In particular, we show SGDm under covariate shift is a parametric oscillator, and so can suffer from a phenomenon known as resonance. We characterize the learning system as a time varying system of ordinary differential equations (ODEs), and leverage existing theory to characterize learning system divergence/convergence as resonant/nonresonant modes of the ODE system. The theoretical result is limited to the linear setting with periodic covariate shift, so we empirically supplement this result to show that resonance phenomena persist across other problem settings having non-periodic covariate shift, nonlinear dynamics with neural networks, and optimizers other than SGDm.

Preface

This thesis is based upon a project which began in Summer 2020 as a playful probe into the intersection of machine learning and signal processing. Significant experimental development occurred in collaboration with Liam Peet-Pare in Fall 2020 as a course project for CMPUT 655 under Dr. Martha White. Further experimental development and all theoretical development was completed in preparation for the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) where the work now awaits review with coathors Liam Peet-Pare, Dr. Martha White, Dr. Nidhi Hegde, and Dr. Alona Fyshe.

To my dad.

You couldn't witness this journey, but you still carried me through it.

Acknowledgements

Thank you to my supervisor, Dr. Martha White, for supporting in countless ways, but especially for so generously trusting me throughout our work together.

Thank you to Liam Peet-Pare for coauthoring the vast majority of this work, and for suggesting several direction changes which turned out to be critical to making progress.

Thank you to Dr. Lei Ma and Dr. Alona Fyshe for your critical feedback as committee members for this dissertation.

Thank you to Dr. Nidhi Hegde, Dr. Alona Fyshe, Dr. Csaba Szepesvári, Logan Gilmour, Ryan Thiessen, and Yurij Salmaniw for significant guidance in many areas of this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Stochastic gradient descent (SGD) [36] – and its variants such as Adagrad [13], ADAM [22] and RMSprop [19] – are very widely used optimization algorithms across machine learning. SGD is conceptually straightforward, easy to implement, and often performs well in practice. Among the variants of SGD, accelerated versions based on Polyak's or Nesterov's acceleration [32], [34], known generally as Stochastic Gradient Descent with Momentum (SGDm), are used widely due to the improvements in convergence rate they offer. SGDm can give up to a quadratic speedup to SGD on many functions and is in fact optimal among all methods having only information about the gradient at consecutive iterates [16], [33]. SGDm has the same computational complexity as SGD, but exhibits superior convergence rates under reasonable assumptions [41].

These convergence results for SGDm, however, rely on independent and identically distributed (iid) sampling. Little is known about the convergence properties of SGDm under non-iid sampling, yet the non-iid setting is critical. In many machine learning problems it is expensive or impossible to obtain iid samples. In online learning [35] and reinforcement learning (RL) [44], the data becomes available in a sequential order and there is a temporal dependence among the samples. There is a particularly strong temporal dependence in RL, where observed states are sampled according to the transition dynamics of the underlying Markov decision process (MDP). Federated learning [20] and time-series learning [26] provide further examples of when non-iid sampling is

essential to the learning problem.

Without momentum, SGD's convergence rate has been examined under non-iid sampling with the stochastic approximation framework in [5], [25], and more recently under specific assumptions of ergodicity [14] or Markovian sampling [12], [31], [42]. Convergence rates under Markovian sampling are also known for ADAM-type algorithms when applied to policy gradient and temporal difference learning [47]. To our knowledge, however, there has been little work on providing convergence rates or guarantees for SGDm under non-iid sampling. In [11], a progress bound is provided for SGDm under Markovian sampling based on mixing time—the time required for a distribution's convergence toward its stationary distribution—along with a convergence rate guarantee under decaying step-sizes. In this work, we assume fixed step-size, since it is a common and reasonable choice in the online setting, especially when the practitioner is unsure of mixing rate or stationarity. We also do not explicitly assume Markovianness or ergodicity in sampling. Instead, we theoretically characterize rates of convergence and divergence by assuming periodicity in the expected loss surface induced by the non-iid sampling, and empirically demonstrate similar convergence and divergence phenomena when the periodicity assumption is relaxed.

There is a broad literature using ordinary differential equations (ODEs) to analyze gradient descent methods by approximating descent updates as continuous-time flows. Early work is comprehensively discussed in [5], [25]. Despite the age and establishment of the linear setting, the gradient flow lens continues to reveal new insights (e.g. implicit rank reduction in [2]) and new perspectives on old insights (e.g. regularization of early stopping in [1].) Recent work has paid particular attention to the flow induced by momentum accelerated methods, [3], [6], [7], [10], [23], [24], [27]–[29], [37]–[41], [45], [46], [48]. In these works, it is often demonstrated that linear regression under iid sampling with SGDm can be represented as an ODE resembling a harmonic oscillator, sometimes with a time-decaying damping coefficient.

In this work, we show that non-iid sampling induces a related system: the parametric oscillator, a harmonic oscillator having coefficients which vary

over time in a manner capable of exponentially exciting the system [9], [18], [30]. We show that such exponential resonant modes highlight the role that sampling can play in convergence, as they allow full characterization of the learning system's instability (subject to the aformentioned assumption of loss surface periodicity) which we demonstrate by leveraging the well-established mathematics of Floquet theory [18].

We empirically extend beyond our theoretical guarantees by borrowing from the methods of empirical nonlinear control. Specifically, we investigate learning systems whose input sampling does not have the strict periodic time variation required by our theoretical guarantees. We treat the input sampling process as an input signal which is noisy and possibly stochastic, and we vary the signal's frequency content across a wide band so that the learning system's output can be empirically measured as a response across input frequencies. This is analogous to a mechanical engineer triggering an oscillatory vibration within a machine's engine compartment, sweeping across frequencies, and measuring the resulting vibration amplitude in the machine's housing. Such methods are made systematic in the modal analysis literature [4].

In the following sections, we start by precisely explaining the problem setting, then show that SGDm under non-iid sampling is a discretization of a particular time-varying ODE, and give conditions for the ODE's stability and instability. We then empirically demonstrate that resonance-driven divergence occurs in a learning system which aligns well with theoretical assumptions. Similar demonstrations follow on learning systems which progressively relax further and further away from our theoretical assumptions.

1.1 Problem Setting

We investigate the effect of non-iid sampling on a model optimized using SGDm. We assume labelled training data sampled from the pair of discrete-indexed, real-valued stochastic processes $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ such that $\{X_k\}_{k\in\mathbb{N}}$ has a unique stationary distribution Π. Y_k is a function of X_k with zero-mean observation noise, $Y_k = f(X_k) + \epsilon_k$. We denote pairs of data sampled from X_k

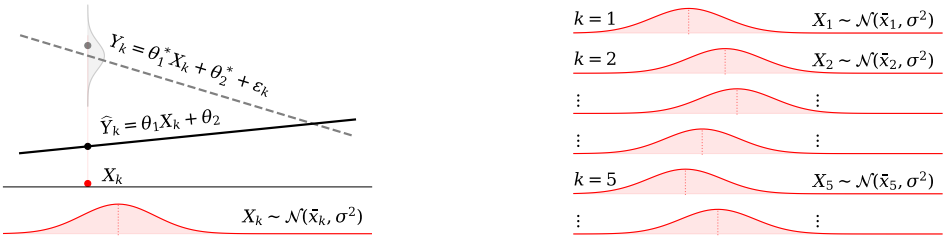


Figure 1.1: The problem setting in a single dimension with a linear model $\widehat{Y}_k$ and target Y_k, and Gaussian covariates X_k. The covariates X_k are shifting in the mean $\bar{x}_k$ over time k.

and Y_k as $z_k = (x_k, y_k)$.

The learning algorithm does not have access to iid samples from Π. Instead, the learning algorithm may only update parameters θ_k at time k using samples (x_k, y_k), where the marginal over time of X_k converges to Π. In this problem setting, the goal is to train a model with parameters θ to minimize an objective function $L(z; \theta)$ with respect to the stationary distribution Π,

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}_\Pi[L(z;\theta)] = \int_{\mathbb{R}} L(z;\theta) d\Pi(x)$$

This setting is similar to that explored in prior work on Markovian sampling [31], [42], [47], but we do not require that our stochastic processes adhere to the Markov property. Note, also, that the underlying functional relationship between the inputs, X_k, and the targets, Y_k, does not change over time. That is, the non-iid sampling is a result of covariate shift rather than a changing relationship between the inputs and targets.

We consider Polyak's Heavy Ball method [34] with the specific formulation from [43] as follows

$$\begin{aligned} v_{k+1} &= \mu v_k - \eta \nabla_{\theta_k} L(z; \theta_k) \\ \theta_{k+1} &= \theta_k + v_{k+1} \end{aligned} \tag{1.1}$$

where η is the learning rate, $\mu \in (0, 1)$ is the momentum coefficient, θ_k is the parameters of the model at time k, and $\nabla_{\theta_k} L(z; \theta_k)$ is the gradient of the objective function with respect to θ_k evaluated at sample z and weights θ_k.

1.2 Contributions

This work contains the following elements which, to our knowledge, are novel contributions to the machine learning research community.

- For supervised learning under covariate shift, we formulate the dynamics of SGDm as a system of ODEs. In existing literature, the SGD-as-ODE formulation has already been extensively used without covariate shift, and we contribute the specific role that covariate shift plays. Since existing ODE formulations of SGD and SGDm have been useful for a wide variety of analyses, we believe our formulation may be of independent interest to those studying covariate shift.

- Using the ODE formulation, we give conditions on covariate shift which cause problematic SGDm behaviour known as divergence. In existing literature, it is known that SGDm is capable of diverging for many different reasons. We contribute to this list of reasons, demonstrating that covariate shift can induce nonlinear resonance, which leads to divergence in systems whose configuration would otherwise converge if not for the presence of covariate shift.

- Finally, we provide empirical evidence for the hypothesis that parametric resonance can occur beyond the assumptions necessary for our theoretical contribtions. In existing literature, there does not yet exist any experiments which specifically control the frequency content of data generating processes. We contribute twofold to this gap, by providing several synthetic data generating processes with controllable frequency content, and demonstrating that non-negligible frequency response exists in several learning systems with momentum.

The first two contributions are primarily theoretical, and the final contribution is empirical.

Chapter 2

Theory: Covariate Shift as a Driving Force

Here we characterize SGDm as a discretization of a particular parametric oscillator in continuous time. A parametric oscillator resembles a harmonic oscillator ODE, but has time-varying system coefficients which are capable of driving the system. In the context of SGDm, the expected loss gradient manifests as a time-varying coefficient matrix induced by non-iid sampling. It is well understood that parametric oscillators can suffer from global solution instability due to coefficients[18] oscillating at particular frequencies, a condition known as *parametric resonance*. We will show the parametric resonance conditions necessary to induce exponential divergence in SGDm.

2.1 Notation

Unless explicitly noted otherwise, we use capital letters for matrices and random variables, with the difference clarified explicitly or by context. $\{X_k\}$ is a stochastic process, shorthand for $\{X_k\}_{k\in\mathbb{N}}$. Similarly, $\{\theta_k\}$ is a sequence, shorthand for $\{\theta_k\}_{k\in\mathbb{N}}$. k and t index discrete and continuous time, respectively. When a discrete sequence and a function approximate each other, they share the same symbol and are differentiated by subscript k and function argument t, e.g. the sequence $\{\theta_k\}$ and the function $\theta(t)$. At the risk of abusing notation, we will adhere to this convention and use $\{\dot{\theta}_k\}$ to be a discrete sequence approximating the function of time $\dot{\theta}(t)$, which refers to the time derivative of $\theta(t)$. This

means the symbol $\dot{\theta}_k$ is the k-th iterate of a sequence approximating $\dot{\theta}(t)$.

The nonnegative natural numbers, the real numbers, and the real-valued d-dimensional square matrices are denoted $\mathbb{N}, \mathbb{R}, \mathbb{R}^{d\times d}$, respectively. $I_{d\times d}, 0_{d\times d}$ are the identity and zero matrix. $\langle\cdot,\cdot\rangle$ denotes the inner product. We denote the joint distribution of X_k and Y_k as P_k. Given a scalar-valued loss function $L(z;\theta)$ for arbitrary training pair $z = (x, y)$ and weights θ, we denote the time-varying expected gradient function $g_k(\theta)$ as the expectation of the gradient of L with respect to θ. That is: $g_k(\theta) := \mathbb{E}_{P_k}[\nabla_\theta L(z;\theta)]$

Wherever we write the training input generating process $\{X_k\}$, it is implicit that the last dimension is fixed at 1 if one wishes to describe linear models with a bias term. In this way, the notation $\langle\theta, X_k\rangle$ accommodates linear models with or without a bias term.

We use ξ to denote the *phase space* coordinates of weights θ, meaning $\xi = \begin{bmatrix}\theta\\ \dot{\theta}\end{bmatrix}$, where vectors $\theta, \dot{\theta} \in \mathbb{R}^d$ are stacked so that the resulting $\xi \in \mathbb{R}^{2d}$. We denote the d-dimensional zero vector as 0_d. For an ODE system $\dot{\xi}(t) = f(\xi(t), t)$ with arbitrary solution trajectory $\xi(t)$ approximated by a sequence of iterates $\{\xi_k\}$, we denote the integration time step as h, and the *numerical flow* as $\phi_{h,k}$, which is the map such that $\xi_{k+1} = \phi_{h,k}(\xi_k)$. We denote the k-th discrete timestep as t_k, which we always use to refer to the k-th multiple of integration time step h, i.e. $t_k = hk$.

2.2 Linear Time-Varying Expected Gradient via Covariate Shift

We begin by showing that linear least squares regression with covariate shift and a fixed target induces a linear time-varying expected loss gradient. That is, the gradient function $g_k(\theta)$ is a linear function of $\theta - \theta^*$ (the weights' position w.r.t. the target) defined in terms of a matrix B_k. Intuitively, the covariate shift causes our B_k to vary in time.

Assumption 1. *The covariate generating process $\{X_k\}_{k\in\mathbb{N}}$ is non-iid, such that for some $k_1, k_2 \in \mathbb{N}$ we have inequality in covariance $Cov(X_{k_1}, X_{k_1}) \neq Cov(X_{k_2}, X_{k_2})$ or inequality in expectation $\mathbb{E}_{P_{k_1}}[X_{k_1}] \neq \mathbb{E}_{P_{k_2}}[X_{k_2}]$.*

Assumption 2. *The targets Y_k are a fixed linear function of X_k with iid zero-mean observation noise, ϵ_k such that $\mathbb{E}[\epsilon_k] = 0$. That is, $Y_k = \langle \theta^*, X_k \rangle + \epsilon_k$ where $\theta^* \in \mathbb{R}^d$ is fixed for all k.*

Proposition 1. *Under Assumptions 1, 2, a linear model with weights θ_k making predictions $\widehat{Y}_k = \langle \theta_k, X_k \rangle$ with a mean squared error (MSE) objective will induce time-varying linear expected loss gradients $g_k(\theta_k) = B_k(\theta_k - \theta^*)$ for all $k \in \mathbb{N}$, where $B_k \in \mathbb{R}^{d \times d}$ and $B_{k_1} \neq B_{k_2}$ for some k_1, k_2.*

Proof. We will show that the gradient of MSE loss $\nabla_{\theta_k} L(Z; \theta_k)$ is a random variable whose expectation will take the linear form $B_k(\theta_k - \theta^*)$. We start with gradient for arbitrary time step k

$$
\begin{aligned}
\nabla_{\theta_k} L(Z; \theta_k) &= \nabla_{\theta_k} (\widehat{Y}_k - Y_k)^2 && \text{MSE def'n} \\
&= \nabla_{\theta_k} [(\langle \theta_k, X_k \rangle - Y_k)^2] && \widehat{Y}_k \text{ def'n from Prop. 1} \\
&= \nabla_{\theta_k} [(\langle \theta_k, X_k \rangle - \langle \theta^*, X_k \rangle + \epsilon_k)^2] && Y_k \text{ def'n from Ass. 2} \\
&= \nabla_{\theta_k} [(\langle \theta_k - \theta^*, X_k \rangle + \epsilon_k)^2] && \langle \cdot, \cdot \rangle \text{ distributivity} \\
&= 2(\langle \theta_k - \theta^*, X_k \rangle + \epsilon_k) \nabla_{\theta_k} [\langle \theta_k - \theta^*, X_k \rangle + \epsilon_k] && \text{chain rule} \\
&= 2(\langle \theta_k - \theta^*, X_k \rangle + \epsilon_k) X_k \\
&= 2 \langle \theta_k - \theta^*, X_k \rangle X_k + 2 \epsilon_k X_k && (*)
\end{aligned}
$$

Taking expectation with respect to distribution P_k:

$$
\begin{aligned}
g_k(\theta_k) &= \mathbb{E}_{P_k} [\nabla_{\theta_k} L(Z; \theta_k)] && g \text{ def'n} \\
&= \mathbb{E}_{P_k} [2 \langle \theta_k - \theta^*, X_k \rangle X_k + 2 \epsilon_k X_k] && \text{by (*)} \\
&= 2 \mathbb{E}_{P_k} [\langle \theta_k - \theta^*, X_k \rangle X_k] + 2 \mathbb{E}_{P_k} [\epsilon_k] \mathbb{E}_{P_k} [X_k] && \mathbb{E} \text{ linear}, \epsilon_k \text{ indep.} \\
&= 2 \mathbb{E}_{P_k} [\langle \theta_k - \theta^*, X_k \rangle X_k] && \mathbb{E}[\epsilon_k] = 0 \\
&= 2 \mathbb{E}_{P_k} [X_k \langle \theta_k - \theta^*, X_k \rangle] && \text{scalar and vector commute} \\
&= 2 \mathbb{E}_{P_k} [X_k \langle X_k, \theta_k - \theta^* \rangle] && \langle \cdot, \cdot \rangle \text{ commutative} \\
&= 2 \mathbb{E}_{P_k} [X_k (X_k^T (\theta_k - \theta^*))] && \langle \cdot, \cdot \rangle \text{ matrix form} \\
&= 2 \mathbb{E}_{P_k} [X_k X_k^T] (\theta_k - \theta^*) && \text{matrix mult. associative} \\
&= B_k (\theta_k - \theta^*)
\end{aligned}
$$

In the last step, we have defined the matrix B_k as twice the expected outer product of X_k with itself, which has the following property.

$$\begin{aligned} B_k &= 2\mathbb{E}_{P_k}[X_k X_k^T] \\ &= 2\left(\mathrm{Cov}(X_k, X_k) + \mathbb{E}_{P_k}[X_k]\mathbb{E}_{P_k}[X_k]^T\right) \end{aligned}$$

So if there exists $k_1, k_2 \in \mathbb{N}$ such that the covariance matrices differ or the mean vectors differ

$$\mathrm{Cov}(X_{k_1}, X_{k_1}) \neq \mathrm{Cov}(X_{k_2}, X_{k_2}) \qquad \text{or} \qquad \mathbb{E}_{P_{k_1}}[X_{k_1}] \neq \mathbb{E}_{P_{k_2}}[X_{k_2}]$$

then $B_{k_1} \neq B_{k_2}$. Assumption 1 provides the necessary k_1, k_2, so the proposition is proved.

□

From the proof, we see that B_k depends only on the first two moments of X_k:

$$B_k = 2\left(\mathrm{Cov}(X_k, X_k) + \mathbb{E}_{P_k}[X_k]\mathbb{E}_{P_k}[X_k]^T\right) \tag{2.1}$$

2.3 ODE Correspondence

Next, we show that the iterates $\{\theta_k\}$ generated by SGDm are a first order numerical integration of a particular ODE. The procedure is similar to [29], but we pay specific attention to the conditions on the ODE necessary to have integration consistency, given the time-varying loss gradient. We start in terms of the matrix-valued function $B(t)$, which is assumed below to be a continuous-time extension of the gradient matrix sequence B_k.

Assumption 3. *Assume there exists matrix-valued function $B(t)$, Lipschitz continuous in t, such that $\{B_k\}$ are samples spaced $\sqrt{\eta}$ apart, i.e. $B_k = B(\sqrt{\eta}k)$.*

Assumption 3 means that B_k is sampled from a continuous $B(t)$ with step-size η acting as a scaling constant. One unit of t is $\approx \frac{1}{\sqrt{\eta}}$ discrete steps of k. (Exact when $\frac{1}{\sqrt{\eta}}$ is an integer.)

Assumption 4. *Step-size, $\eta > 0$, and momentum parameter, $\mu \in [0, 1]$, are constant $\forall k \in \mathbb{N}$.*

Proposition 2. *Under Assumptions 3 and 4 the SGDm iterates $\{\theta_k\}$ numerically integrate the ODE system (2.2) with integration step $\sqrt{\eta}$ and first order consistency.*

$$\ddot{\theta}(t) + \frac{1-\mu}{\sqrt{\eta}}\dot{\theta}(t) + B(t)(\theta(t) - \theta^*) = 0 \tag{2.2}$$

Proof. The proof proceeds two parts. First we derive a first order operator-splitting integrator for the ODE (2.2), then we show that it is equivalent to SGDm (1.1).

(Part 1: Operator-Splitting Integrator)

Let $\xi : \mathbb{R} \mapsto \mathbb{R}^{2d}$ be the vector valued function of time whose first d elements are $\theta(t)$, and last d elements are $\dot{\theta}(t)$:

$$\xi(t) = \begin{bmatrix} \theta(t) - \theta^* \\ \dot{\theta}(t) \end{bmatrix} \tag{2.3}$$

i.e. ξ is a phase space transformation allowing us to rewrite ODE (2.2) in the form of (2.9), which can then be split into a sum of separate systems $f^{[1]}, f^{[2]}$ as follows

$$\begin{aligned}
\dot{\xi}(t) &= f(\xi(t), t) \\
\dot{\xi}(t) &= \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -B(t) & -\frac{1-\mu}{\sqrt{\eta}} I_{d\times d} \end{bmatrix} \xi(t) \\
\dot{\xi}(t) &= \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ 0_{d\times d} & 0_{d\times d} \end{bmatrix} \xi(t) + \begin{bmatrix} 0_{d\times d} & 0_{d\times d} \\ -B(t) & -\frac{1-\mu}{\sqrt{\eta}} I_{d\times d} \end{bmatrix} \xi(t) \\
\dot{\xi}(t) &= f^{[1]}(\xi(t), t) + f^{[2]}(\xi(t), t)
\end{aligned} \tag{2.4}$$

Let $h > 0$ be an integration time step, $\phi^{[1]}_{h,k}(\xi_k)$ be the implicit Euler numerical flow of $f^{[1]}(\xi(t), t)$:

$$\begin{aligned}
\phi^{[1]}_{h,k}(\xi_k) &= \xi_k + h f^{[1]}(\phi^{[1]}_{h,k}(\xi_k), t_{k+1}) \\
&= \xi_k + h \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ 0_{d\times d} & 0_{d\times d} \end{bmatrix} \phi^{[1]}_{h,k}(\xi_k)
\end{aligned} \tag{2.5}$$

Let $\phi^{[2]}_{h,k}(\xi_k)$ be explicit Euler numerical flow of $f^{[2]}(\xi(t), t)$:

$$\begin{aligned}\phi^{[2]}_{h,k}(\xi_k) &= \xi_k + hf^{[2]}(\xi_k, t_k) \\ &= \xi_k + h\begin{bmatrix} 0_{d\times d} & 0_{d\times d} \\ -B(t_k) & -\frac{1-\mu}{\sqrt{\eta}} I_{d\times d}\end{bmatrix}\xi_k \end{aligned} \tag{2.6}$$

The composed flow

$$\phi_{h,k} := \phi^{[1]}_{h,k} \circ \phi^{[2]}_{h,k} \tag{2.7}$$

is a sequentially split operator, which has splitting error order 1 because (2.4) is time-varying linear system [15]. The operators being composed, implicit and explicit Euler, are both order 1 consistent with their respective systems [21]. The overall order of consistency of a split operator is the minimum of splitting error, and the orders of the composed flows [8]. So we have that (2.7) approximates (2.4) with order 1 consistency.

(Part 2: SGDm Equivalency)

We will now show that (2.7) is equivalent to SGDm when integration timestep $h = \sqrt{\eta}$, where η is the SGDm step-size. We start with the definition of ξ_{k+1} as the numerical flow of ξ_k:

$$\begin{aligned}
\xi_{k+1} &:= \phi_{h,k}(\xi_k) \\
&= \phi^{[1]}_{h,k}(\phi^{[2]}_{h,k}(\xi_k)) && \text{by (2.7)} \\
&= \phi^{[2]}_{h,k}(\xi_k) + h\begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ 0_{d\times d} & 0_{d\times d}\end{bmatrix}\phi^{[1]}_{h,k}(\phi^{[2]}_{h,k}(\xi_k)) && \text{subst'n from (2.5)} \\
&= \phi^{[2]}_{h,k}(\xi_k) + h\begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ 0_{d\times d} & 0_{d\times d}\end{bmatrix}\xi_{k+1} && \text{def'n of } \xi_{k+1} \\
&= \phi^{[2]}_{h,k}(\xi_k) + h\begin{bmatrix} \dot\theta_{k+1} \\ 0_d \end{bmatrix} && \text{matrix mult.} \\
&= \left(\xi_k + h\begin{bmatrix} 0_{d\times d} & 0_{d\times d} \\ -B(t_k) & -\frac{1-\mu}{\sqrt{\eta}} I_{d\times d}\end{bmatrix}\xi_k\right) + h\begin{bmatrix} \dot\theta_{k+1} \\ 0_d \end{bmatrix} && \text{subst'n from (2.6)} \\
&= \left(\xi_k + h\begin{bmatrix} 0_d \\ -B(t_k)(\theta_k - \theta^*) - \frac{1-\mu}{\sqrt{\eta}}\dot\theta_k\end{bmatrix}\right) + h\begin{bmatrix} \dot\theta_{k+1} \\ 0_d \end{bmatrix} && \text{matrix mult.} \\
\xi_{k+1} &= \xi_k + h\begin{bmatrix} \dot\theta_{k+1} \\ -B(t_k)(\theta_k - \theta^*) - \frac{1-\mu}{\sqrt{\eta}}\dot\theta_k\end{bmatrix} && \text{simplification}
\end{aligned}$$

Recalling that $\xi_k = \begin{bmatrix}\theta_k - \theta^* \\ \dot\theta_k\end{bmatrix}$, the above immediately provides the following

two recurrence relations:

$$\theta_{k+1} = \theta_k + h\dot{\theta}_{k+1} \qquad \dot{\theta}_{k+1} = \dot{\theta}_k + h\left(-B(t_k)\theta_k - \frac{1-\mu}{\sqrt{\eta}}\dot{\theta}_k\right)$$

Now let integration time step h be the square root of SGDm step-size, $h = \sqrt{\eta}$, and define $v_k = \sqrt{\eta}\dot{\theta}_k$. Substituting h, v_k, we proceed via elementary algebra:

$$\theta_{k+1} = \theta_k + \sqrt{\eta}\left(\frac{v_{k+1}}{\sqrt{\eta}}\right) \quad \left(\frac{v_{k+1}}{\sqrt{\eta}}\right) = \left(\frac{v_k}{\sqrt{\eta}}\right) + \sqrt{\eta}\left(-B(t_k)\theta_k - \frac{1-\mu}{\sqrt{\eta}}\left(\frac{v_k}{\sqrt{\eta}}\right)\right)$$

$$\theta_{k+1} = \theta_k + v_{k+1} \qquad v_{k+1} = v_k + \eta\left(-B(t_k)\theta_k - \frac{(1-\mu)v_k}{\eta}\right)$$

$$v_{k+1} = v_k - \eta B(t_k)\theta_k - (1-\mu)v_k$$

$$v_{k+1} = \mu v_k - \eta B(t_k)\theta_k$$

Note that $B(t_k) = B_k$ (Assumption 3), and that B_k defines the time-varying gradients induced by covariate shift and linear regression to a linear target (Proposition 1). Hence, under those conditions, we have arrived at SGDm (1.1). The proof is complete.

□

The reader may observe that, given a solution $\xi(t)$ to the system $\dot{\xi}(t) = A(t)\xi(t)$, the proof above shows that the SGDm iterates $\{\theta_k\}$ are precisely a first order numerical approximation of the first d dimensions of $\xi(t) = \begin{bmatrix}\theta(t) - \theta^* \\ \dot{\theta}(t)\end{bmatrix}$, but the remaining d dimensions are approximated only up to a scale factor $\sqrt{\eta}$ by iterates $\{v_k\}$. However, this is not an issue. Solutions $\theta(t)$ to the system (2.2) are embedded within the first d dimensions of $\xi(t)$, so the remaining d dimensions and iterates $\{v_k\}$ do not affect the result.

The first order consistency guarantee in Proposition 2 means that when step-size η is small and initial conditions agree between continuous and discrete time, i.e. $\theta_0 = \theta(0)$, the continuous and discrete time trajectories approximate each other as $\theta_k \approx \theta(k\sqrt{\eta})$, where the difference between them depends on the integration time step $h = \sqrt{\eta}$. The difference accrued in one step of k (local

error) and over arbitrarily many steps (global error) behave as follows [17].

$$\begin{array}{ccccc} & \text{Local Error} & & \text{Global Error} & \\ \theta_0 = \theta(0) \implies \theta_1 = \theta(h) + O(h^2) & & \text{and} & \theta_k = \theta(hk) + O(h^2 k) & (2.8) \\ = \theta(\sqrt{\eta}) + O(\eta) & & & = \theta(\sqrt{\eta}k) + O(\eta k) & \end{array}$$

2.4 Parametric Resonance for ODE Convergence and Divergence

Next we connect Propositions 1 and 2, so that conditions sufficient for convergence and divergence in (2.2) may be shown using established dynamical systems theory. The conditions sufficient for divergence are precisely the conditions for parametric resonance. We can theoretically guarantee this when the matrix $B(t)$ is periodic according to Theorem 1.

As per elementary results in linear ODE theory, system (2.2) can be transformed into a linear time-varying first order form

$$\dot{\xi}(t) = A(t)\xi(t) \qquad A(t) = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ B(t) & \frac{1-\mu}{\sqrt{\eta}} I_{d\times d} \end{bmatrix} \qquad (2.9)$$

such that solution trajectories $\theta(t)$ of (2.2) are embedded in solution trajectories $\xi(t)$ of (2.9). Moreover, (2.9) admits a fundamental solution matrix[1] $\psi(t)$ such that the spectral radius ρ of $\psi(T)$ characterizes ODE instability, which implies divergence for SGDm, as stated in Theorem 1.

Theorem 1. *When $B(t)$ is periodic such that $B(t) = B(t+T)$, the spectral radius ρ of $\psi(T)$ characterizes the stability of solution trajectories of (2.2) as follows:*

- $\rho > 1 \implies$ *trivial solution $\theta(t) = \theta^*$ is unstable. All other solutions diverge as $\theta(t) \to \infty$ exponentially with rate ρ.*
- $\rho < 1 \implies$ *trivial solution is asymptotically stable, all other solutions converge as $\theta(t) \to \theta^*$ exponentially with rate ρ.*

[1] A fundamental solution matrix for system (2.9) is any matrix-valued function of time $\psi(t)$ whose columns are linearly independent solutions to (2.9). We choose $\psi(t)$ such that $\psi(0) = I_{2d\times 2d}$.

Proof. The proof of this theorem relies heavily on the well-established mathematics of Floquet theory and the stability result is contained in Theorem 1.9 and Theorem 1.10 of [18]. Theorem 1.10 states that for a system of differential equations of the form

$$\dot{\xi} = A(t)\xi, \qquad A(t+T) = A(t) \tag{2.10}$$

where $A(t)$ is piecewise continuous, the stability of the trivial solution $\xi(t) \equiv 0$ is determined by the spectral radius of the system's *monodromy matrix* M defined below

$$M = \psi^{-1}(0)\psi(T) \qquad \text{where } \dot{\psi}(t) = A(t)\psi(t)$$

(Monodromy Matrix)

Matrix-valued functions $\psi(t)$ are the system's *fundamental solution matrix*, and elementary existence results for linear systems allow one to choose a $\psi(t)$ such that $\psi(0) = I$ so that the monodromy matrix simplifies as

$$\begin{aligned} M &= \psi^{-1}(0)\psi(T) \\ &= I^{-1}\psi(T) \\ &= I\psi(T) \\ M &= \psi(T) \end{aligned}$$

Below we denote the spectral radius of $\psi(T)$ (and hence of M) as ρ.

(First Order Linear Form, Stability via Floquet)

Below we show that (2.2) can be transformed to the form (2.10) with $A(t)$ continuous such that trivial solution stability of (2.10) is equivalent to stability of the solution $\theta(t) = \theta^*$.

Let $\xi(t)$ be the phase space transformation of $\theta(t)$, similarly to the proof of Proposition 2

$$\xi(t) := \begin{bmatrix} \theta(t) - \theta^* \\ \dot{\theta}(t) \end{bmatrix}$$

so that for each t, $\theta(t) \in \mathbb{R}^d$ and $\xi(t) \in \mathbb{R}^{2d}$. This means (2.2) (restated below)

$$\ddot{\theta}(t) + \frac{1-\mu}{\sqrt{\eta}}\dot{\theta}(t) + B(t)(\theta(t) - \theta^*) = 0$$

is equivalent to the following first order linear form

$$\dot{\xi}(t) = A(t)\xi(t) \qquad \text{where } A(t) = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -B(t) & -\frac{1-\mu}{\sqrt{\eta}} I_{d\times d} \end{bmatrix}$$

Since $B(t)$ is periodic and continuous, which is stronger than the requisite piecewise continuity. Since all other submatrices of $A(t)$ are constant, we have that $A(t)$ is also periodic and (piecewise) continuous. Now Theorem 1.10 in [18] immediately provides the following

- $\rho > 1 \implies$ trivial solution $\xi(t) = 0$ is unstable. All other solutions diverge as $\xi(t) \to \infty$ exponentially with rate ρ.
- $\rho < 1 \implies$ trivial solution is asymptotically stable, all other solutions converge as $\xi(t) \to 0$ exponentially with rate ρ.

Since $\xi(t) = \begin{bmatrix} \theta(t) - \theta^* \\ \dot{\theta}(t) \end{bmatrix}$:

- The trivial solution $\xi(t) = 0$ is equivalent to $\theta(t) = \theta^*$
- $\xi(t) \to 0$ is equivalent to $\theta(t) \to \theta^*$
- $\xi(t) \to \infty$ is equivalent to $\theta(t) \to \infty$

The theorem is proved.

□

The spectral radius conditions $\rho > 1$ and $\rho < 1$ characterize when ODE solution trajectories $\theta(t)$ will converge or diverge, and Proposition 2 tells us that these solution trajectories are approximations of discrete time SGDm trajectories $\{\theta_k\}$ under identical initialization. But does convergence or divergence of $\theta(t)$ imply the same for SGDm? Experiment 3.1 empirically suggests that the approximation is sufficient, since the boundary at $\rho = 1$ in Figure 3.1a agrees with both continuous and discrete time.

However, given Theorem 1 and Proposition 2, we have a theoretical guarantee of SGDm's divergence, but not its convergence. This is because the divergence rate of $\theta(t)$ is exponential, and Proposition 2 provides a bound on

long tail behaviour as $\theta_k = \theta(\sqrt{\eta}k) + O(\eta k)$ from (2.8). Since the approximation error is linear in time k, and the divergence rate is exponential, the divergence rate dominates, and we are guaranteed that a diverging $\theta(t)$ corresponds to a diverging $\{\theta_k\}$. However, the same argument does not imply that a convergent $\theta(t)$ corresponds to a convergent $\{\theta_k\}$, because the linear error bound technically permits the discrete trajectory to escape θ^* at a linear rate. Despite the lack of theoretical guarantee, we see empirical agreement for *both* convergent and divergent cases, but we defer theoretical proof to future work.

2.5 Example

A learning system which satisfies all of the above assumptions is online linear least squares regression using SGDm with a fixed η, μ under a periodic covariate shift. To be concrete, let data be sampled from the sequence of random variables $\{X_k, Y_k\}_{k \in \mathbb{N}}$ defined as follows

$$\begin{aligned} X_k &\sim \mathcal{N}(\bar{x}_k, 1) \\ \bar{x}_k &= a\cos(2\pi f k) \\ Y_k &= \theta_1^* X_k + \theta_2^* + \epsilon \\ \epsilon &\sim \mathcal{N}(0, 1) \end{aligned}$$

where $\bar{x}_k$ is the mean of covariates X_k at time k, $\theta^* = [\theta_1^*, \theta_2^*]^T$ are the target weights, and ϵ is the iid observation noise. Note, this is precisely the setup for Experiment 3.1.

We can use Theorem 1 to determine which values of η, μ, and covariate shift frequency, f, will result in divergence or convergence of the learning system. We numerically compute the fundamental solution matrix of the ODE corresponding to this learning system and then calculate the spectral radius of this matrix evaluated at time T. By convergence we mean that the learned parameters will converge to the true underlying θ_1^* and θ_2^*, and divergence means they will move arbitrarily far from the true θ_1^* and θ_2^*.

We consider a squared error loss function,

$$\begin{aligned} L(z;\theta) &= (\hat{y} - y)^2 \\ &= [\langle \theta, \mathbf{x} \rangle - (\langle \theta^*, \mathbf{x} \rangle + \epsilon)]^2 \\ &= [(\theta_1 x + \theta_2) - (\theta_1^* x + \theta_2^* + \epsilon)]^2 \end{aligned}$$

Hence, taking gradients at time k w.r.t. θ_k we have

$$\nabla_{\theta_k} L(z_k;\theta_k) = \begin{bmatrix} 2x_k[(\theta_{k_1} x_k + \theta_{k_2}) - (\theta_{k_1}^* x_k + \theta_{k_2}^* + \epsilon)] \\ 2[(\theta_{k_1} x_k + \theta_{k_2}) - (\theta_{k_1}^* x_k + \theta_{k_2}^* + \epsilon)] \end{bmatrix}$$

Taking expected gradients,

$$\begin{aligned} \mathbb{E}[\nabla_{\theta_k} L(z_k;\theta_k)] &= \begin{bmatrix} \mathbb{E}[2[(\theta_{k_1} x_k^2 + \theta_{k_2} x_k) - (\theta_{k_1}^* x_k^2 + \theta_{k_2}^* x_k + \epsilon x_k)]] \\ \mathbb{E}[2[(\theta_{k_1} x_k + \theta_{k_2}) - (\theta_{k_1}^* x_k + \theta_{k_2}^* + \epsilon)]] \end{bmatrix} \\ &= \begin{bmatrix} 2[(\theta_{k_1}(1+\bar{x}_k^2) + \theta_{k_2}\bar{x}_k) - (\theta_{k_1}^*(1+\bar{x}_k^2) + \theta_{k_2}^*\bar{x}_k + \mathbb{E}[\epsilon]\bar{x}_k)] \\ 2[(\theta_{k_1}\bar{x}_k + \theta_{k_2}) - (\theta_{k_1}^*\bar{x}_k + \theta_{k_2}^* + \mathbb{E}[\epsilon])] \end{bmatrix} \\ &= 2\begin{bmatrix} (\theta_{k_1}(1+\bar{x}_k^2) + \theta_{k_2}\bar{x}_k) - (\theta_{k_1}^*(1+\bar{x}_k^2) + \theta_{k_2}^*\bar{x}_k) \\ (\theta_{k_1}\bar{x}_k + \theta_{k_2}) - (\theta_{k_1}^*\bar{x}_k + \theta_{k_2}^*) \end{bmatrix} \\ &= 2\begin{bmatrix} \theta_{k_1}(1+\bar{x}_k^2) - \theta_{k_1}^*(1+\bar{x}_k^2) + \theta_{k_2}\bar{x}_k - \theta_{k_2}^*\bar{x}_k \\ \theta_{k_1}\bar{x}_k - \theta_{k_1}^*\bar{x}_k + \theta_{k_2} - \theta_{k_2}^* \end{bmatrix} \\ &= 2\begin{bmatrix} (1+\bar{x}_k^2) & \bar{x}_k \\ \bar{x}_k & 1 \end{bmatrix}\begin{bmatrix} \theta_{k_1} - \theta_{k_1}^* \\ \theta_{k_2} - \theta_{k_2}^* \end{bmatrix} \end{aligned}$$

In the notation of Proposition 1, we can write this as

$$g_k(\theta_k) = B_k[\theta_k - \theta^*]^T \qquad \text{with} \quad B_k = 2\begin{bmatrix} (1+\bar{x}_k^2) & \bar{x}_k \\ \bar{x}_k & 1 \end{bmatrix}$$

If we let $\bar{x}(t) = a\cos(2\pi f \eta^{-\frac{1}{2}} t)$, then we can satisfy Assumption 3 (i.e. $B_k = B(\sqrt{\eta}k)$ with the matrix-valued function $B(t)$

$$B(t) = 2\begin{bmatrix} (1+\bar{x}(t)^2) & \bar{x}(t) \\ \bar{x}(t) & 1 \end{bmatrix}$$

From Proposition 2, we have that our learning system numerically integrates an LTV of the form

$$\ddot{\theta}(t) + \frac{1-\mu}{\sqrt{\eta}}\dot{\theta}(t) + B(t)(\theta(t) - \theta^*) = 0 \quad \text{with} \quad (\theta(t) - \theta^*) = \begin{bmatrix} \theta_1(t) - \theta_1^*(t) \\ \theta_2(t) - {\theta_2}^*(t) \end{bmatrix} \tag{2.11}$$

Note that $B(t)$ is piecewise continuous and periodic with period $T = \frac{1}{f}$.

This second order ODE can be represented as a system of first order ODEs with the transformation

$$\begin{aligned}\xi_1 &= \theta_1 - \theta_1^* \\ \xi_2 &= \theta_2 - \theta_2^* \\ \xi_3 &= \dot{\theta}_1 \\ \xi_4 &= \dot{\theta}_2\end{aligned}$$

so that the equation (2.11) can be written in standard first order linear form $\dot{\xi} = A(t)\xi$:

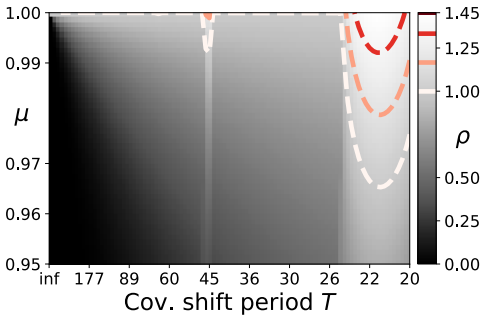
$$\begin{bmatrix}\dot{\xi}_1 \\ \dot{\xi}_2 \\ \dot{\xi}_3 \\ \dot{\xi}_4\end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2(1+\bar{x}^2(t)) & -2\bar{x}(t) & \frac{\mu-1}{\sqrt{\eta}} & 0 \\ -2\bar{x}(t) & -2 & 0 & \frac{\mu-1}{\sqrt{\eta}} \end{bmatrix} \begin{bmatrix}\xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4\end{bmatrix} \tag{2.12}$$

The matrix $B(t)$ is piecewise continuous and periodic, which implies the matrix $A(t)$ also shares these properties. The ODE (2.11) satisfies all the assumptions needed to make use of Theorem 1 to determine the stability of solution trajectories. We can obtain a fundamental solution matrix, $\psi(t)$, of (2.12) satisfying initial conditions $\psi(0) = I_{4\times 4}$ by numerical computation through the use of an ODE solver.

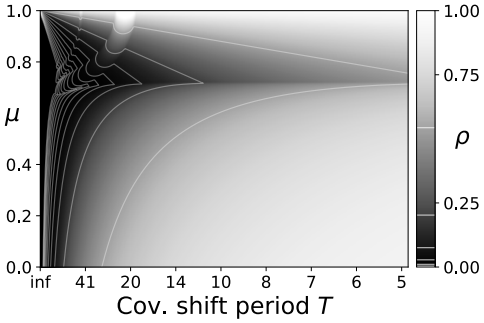
Once we have this matrix $\psi(t)$, we evaluate it at time $T = \frac{1}{f}$ to obtain the system's monodromy matrix. Now we only need to determine the spectral radius, ρ, of $\psi(T)$ to determine the stability of solution trajectories for the system. Details regarding why this is true are given in the proof of Theorem 1.

We repeat this process, sweeping over values of μ, η, and f to determine the stability of solution trajectories for systems to triples of fixed values for these hyperparameters. The results of this process are given in Figure 2.1 below. Interestingly, the right column of 2.1 shows a band of minimum spectral radius, which suggests that there exists an optimal μ_{best} which does not resonate like larger values of μ, and also has the fastest convergence rate. Further, the optimal band is nearly horizontal, with only minor deviations toward the far left where data approaches iid, which suggests that the optimality of μ_{best} applies across a very wide band of frequencies However, the location of μ_{best}

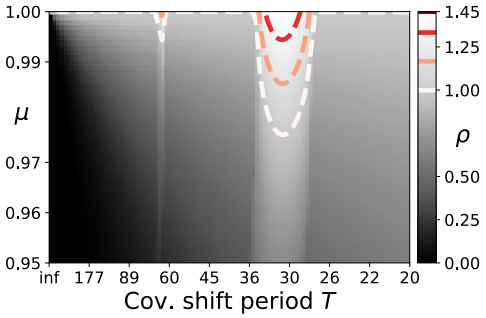
clearly changes with step-size η, and also likely depends upon other aspects of the specific learning problem.



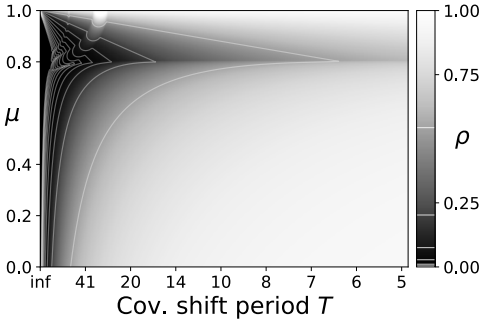
(a) Spectral radius heatmap, $\eta = 0.01$



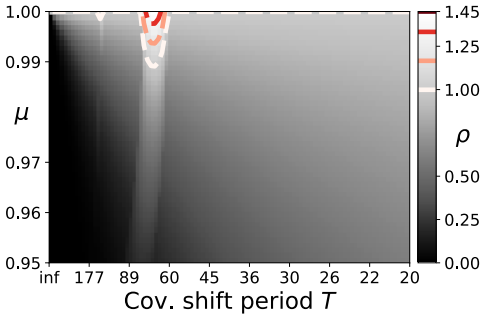
(b) Spectral radius heatmap, wide range, $\eta = 0.01$



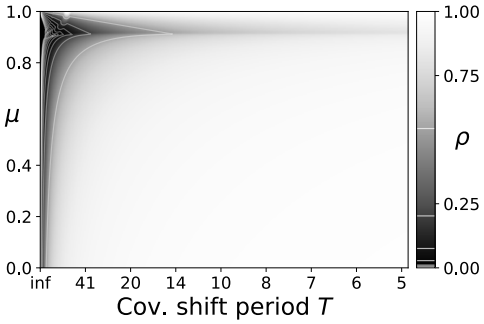
(c) Spectral radius heatmap, $\eta = 0.005$



(d) Spectral radius heatmap, wide range, $\eta = 0.005$



(e) Spectral radius heatmap, $\eta = 0.001$



(f) Spectral radius heatmap, wide range, $\eta = 0.001$

Figure 2.1: Spectral radii of the monodromy matrices induced by particular momentum μ and period T values (x and y pixel coordinates, respectively). Step-size η decreases with each row. Each column shares the same range of μ, T. The right column shows the full range $\mu \in [0, 1]$ and a wide range of periods T. Each left figure 'zooms in' to the upper left corner of the figure to its right. (a) corresponds to the μ, η, T as Experiments 3.1 and 3.2, with contour lines identical to 3.1a 3.1b such that the white line separates the convergent region below from the divergent regions above.

Chapter 3

Validating Theory, then Ablating Towards Conditions in the Wild

We now empirically evaluate the effect of resonance in learning problems. First, Experiment 3.1 validates the theoretical predictions by investigating the dynamics of a learning problem which closely matches the theoretical assumptions in Chapter 2. Specifically, we assess whether or not the spectral radius ρ predicts optimizer convergence or divergence. Then, in order to investigate when similar phenomena can be seen in more realistic learning problems, the remaining Experiments 3.2 - 3.6 sequentially move away from theoretical assumptions. Each experiment makes a single step away from a theoretical assumption, and sweeps a variable which controls the frequency content of covariate shift. Hence, we are able to empirically asses the frequency response in the resulting losses for increasingly realistic learning problems.

Across all experiments, input samples at each training step k are drawn from Gaussian distributions with diagonal covariance matrices, i.e. $X_k \sim \mathcal{N}(\bar{x}_k, cI_{d\times d})$, and we induce covariate shift by constructing a time-varying mean sequence $\{\bar{x}_k\}$. Each mean sequence is designed such that its frequency content (and hence the frequency content of X_k) can be swept with a single parameter f or T. All sequences $\{\bar{x}_k\}$ are designed such that iid sampling is induced by setting $f = 0$ or $T = 0$, so that each experiment has an iid baseline for comparison. The specification of $\bar{x}_k$ in terms of its frequency sweep parameter f or T is provided in Table 3.1 for each experiment.

Since divergence conditions are our focus, f, T are swept such that the

resonating regions are highlighted, with μ sufficiently high that resonance will occur. To demonstrate how SGDm can still be used effectively, all experiments are repeated in Section 4 with μ reduced such that resonance is mitigated.

Table 3.1: Covariate shift details for each experiment.

Experiment	Sweep Param.	Mean Sequence $\{\bar{x}_k\}_{k\in\mathbb{N}}$
3.1 Validating Theory	$f \in [0, 0.05]$	$\bar{x}_k = 0.5\sin(2\pi f k)$
3.2 Ablating Periodicity	$f \in [0, 0.05]$	$\bar{x}_k = \phi_1 \bar{x}_{k-1} + \phi_2 \bar{x}_{k-2} + \xi_k$ $\phi_1 = 4\phi_2(\phi_2 - 1)^{-1}\cos(2\pi f)$ $\phi_2, \xi_k, \bar{x}_1, \bar{x}_2$ see Table 3.3
3.3 Ablating Expected Gradient	$T \in [0, 120]$	$\bar{x}_k = \begin{cases} \frac{\xi}{2\|\|\xi\|\|} \text{ if } \lfloor \frac{2k}{T} \rfloor \equiv 0 \mod 1 \\ \frac{-\xi}{2\|\|\xi\|\|} \text{ if } \lfloor \frac{2k}{T} \rfloor \equiv 1 \mod 1 \end{cases}$ $\xi \sim \mathcal{N}(0, I_{d\times d})$ iid
3.4 Ablating Periodicity Further	$T \in [0, 50]$	$\bar{x}_k = \xi_i$ where $i = \left\lfloor \frac{k}{T} \right\rfloor$ $\xi_i \sim \mathcal{N}(0, vI_{d\times d})$ iid
3.5 Ablating Optimizer Linearity	$T \in [0, 100]$	$\bar{x}_k$ same as above.
3.6 Ablating Model Linearity	$T \in [0, 100]$	$\bar{x}_k$ same as above.

3.1 Experiment 1: Validating Theory

We start in a setting as close as possible to the theoretical predictions, with linear regression for a quadratic loss, and covariate shift such that the mean $\mathbb{E}[X_k]$ varies as a strict sinusoid. We perform regression in two weights (i.e. inputs X_k and labels Y_k are scalar-valued.) While fully stochastic SGD suggests a single sample should be drawn from each X_k, we draw 20 samples from each X_k, so that the loss gradient for each time step k is nearer to its expected value. To be clear, this is not batching over time, as would be done for conventional minibatches. Instead, we are drawing more samples from the distribution of X_k at each instant in time k. Refer to Table 3.2 for details.

Since we have only two weights, the learning system's underlying ODE can easily be completely specified, such that the fundamental solution matrix can be numerically computed. As per Theorem 1, we can use the spectral radius of the

fundamental solution matrix evaluated at time T to make theoretical predictions of where the system should converge or diverge. As depicted in Figure 3.1a, the theory agrees very well with empirical results. This suggests that parametric resonance, as predicted by Floquet theory, is indeed the dominant mechanism behind SGDm divergence under covariate shift. Refer to example 2.5 for the procedure used to compute the theoretical predictions (i.e. spectral radii ρ), and Figure 2.1a for the full surface from which the contour lines in Figure 3.1a are rendered.

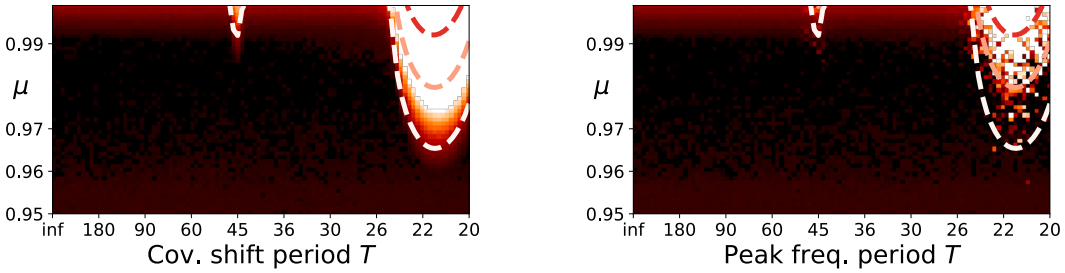
Table 3.2: Details for linear regression sinusoidal covariate shift problem.

Sinusoid Mean Frequency	$f \in [0, 0.05]$
Covariate Shift Mean	$\bar{x}_k = 0.5 \sin(2\pi f k)$
Input Sampling	$X_k \sim \mathcal{N}(\bar{x}_k, 1)$
Target Function (Fixed $\forall k$)	$Y_k = \theta_1^* X_k + \theta_2^*$ $\theta_1^*, \theta_2^* \sim \text{Uniform}[-1, 1]$
Model and Optimizer	$\widehat{Y}_k = \theta_{k,1} X_k + \theta_{k,2}$ $\theta_{0,1}, \theta_{0,2} \sim \text{Uniform}[-1, 1]$ $(\theta_k)_{k \in \mathbb{N}} \leftarrow \text{SGDm}(\eta, \mu)$ $\eta = 0.01, \mu \in [0.95, 0.999]$ $k \in [0, 10^4]$

3.2 Experiment 2: Ablating Periodicity

Here we repeat the same experiment, but with the mean of X_k varying stochastically instead of deterministically.

The precise characterization of solution trajectory stability via Theorem 1 depends on the expected gradient to vary periodically over time given fixed weights θ. In most realistic scenarios, however, we should not expect strictly periodic time variation, which implies Theorem 1 is no longer strictly applicable to characterize the stability of solution trajectories. But even with aperiodic and/or stochastic time variation, our LTV system (2.9) might be similarly susceptible to instability.



(a) Resonance regions for sinusoidal $\bar{x}_k$

(b) AR(2) Stochastic Mean

Figure 3.1: Empirical heatmap of SGDm for linear regression, overlaid by contours of theoretical prediction. Each pixel is the distance $||\theta_k - \theta^*||$ averaged over the final 500 steps k and 10 runs. Coordinates are momentum μ (y-axis) and covariate shift mean signal period T (x-axis, a) or period T corresponding to the dominant frequency in the X_t signal (x-axis, b). Dark pixels converge quickly and stably, bright pixels diverge exponentially. Contour show divergence predictions from Theorem 1: white contour has $\rho = 1$, with ρ increasing with redness.

Here we demonstrate instability by replacing our periodic covariate shift mean with a mean that moves according to an AR(2) process. We run the same experiment as before but instead of having the mean of X_k drifting according to a periodic function, it moves according to an AR(2) process tuned to have a frequency peak exactly at the frequency of the sinusoidal covariate shift of Experiment 3.1. Refer to table 3.1 for details on the covariate shift. We can see from the heatmap in Figure 3.1b that under this setting we observe very similar resonance behaviour in the learning system, which aligns well with the identical predicted stability regions.

Note that both theoretical and empirical results in Figures 3.1a and 3.1b suggest that sufficiently low momentum values μ mitigate the resonance phenomenon, which agrees with the role it plays in the ODE system: decreasing μ increases damping. We also observe that resonance regions shrink as step size η is decreased, see Section 4 for plots demonstrating this trend. Analytically characterizing the bounds of stable μ, η in terms of system properties is an interesting future direction.

Table 3.3: Details for linear regression sinusoidal covariate shift problem. Rather than choosing a frequency f and using it directly as in Table 3.2, we use f together with the stationary distribution variance 0.1 to compute AR(2) coefficients ϕ_1, ϕ_2.

Expected Dominant Freq. in $\bar{x}_k$	$f \in [0, 0.05]$
$\bar{x}_k$ Stationary Dist. (Fixed $\forall f$)	$P = \mathcal{N}(0, 0.1)$
Covariate Shift Mean	$\bar{x}_k = \phi_1 \bar{x}_{k-1} + \phi_2 \bar{x}_{k-2} + \xi_k$ $\xi_k \sim \mathcal{N}(0, 10^{-5})$ iid $\bar{x}_1, \bar{x}_2 \sim P$ $\phi_1 = \frac{4\phi_2}{\phi_2 - 1} \cos(2\pi f)$ ϕ_2 s.t. $[\bar{x}_k \mid \bar{x}_{k-1} \sim P] \sim P$
Remaining Parameters	$X_k, Y_k, \widehat{Y}_k, \theta^*, \theta_0, (\theta_k)_{k \in \mathbb{N}}, \eta, \mu, k$ same as Table 3.2

3.3 Experiment 3: Ablating Expected Gradient

Here we perform linear regression with periodic mean covariate shift, and ablate the number of samples drawn from each X_k to show the effect of increasing noise in the gradient signal. Linear regression is performed mapping $\mathbb{R}^5 \to \mathbb{R}$, with the weights learned via SGDm. The mean sequence $\bar{x}_k$ oscillates with strict periodicity between $\pm\bar{x}$, where $\bar{x}$ is a unit norm 5-vector randomly chosen for each run. i.e. the mean signal is a square wave in $\mathbb{R}^5$ with period T. Refer to Table 3.1 for details of the covariate shift.

Linear regression implies a quadratic loss, and periodic covariate shift implies a loss with periodic time variation in expectation, so we are very near to the setting in which Theorem 1 is directly applicable. However, as we ablate from 5 samples to 1 sample drawn from each X_k, we move further away from the expected gradient, towards the fully stochastic gradient setting. As we can see in Figure 3.2 resonance is dampened by stochasticity in the gradient signal. Losses are normalized with respect to the number of samples drawn from X_k, so that loss and gradient magnitude are independent of the number of samples.

In all following experiments, 3.4 - 3.6, only a single sample is drawn from each X_k, so that our results better reflect the modern setting where optimizers

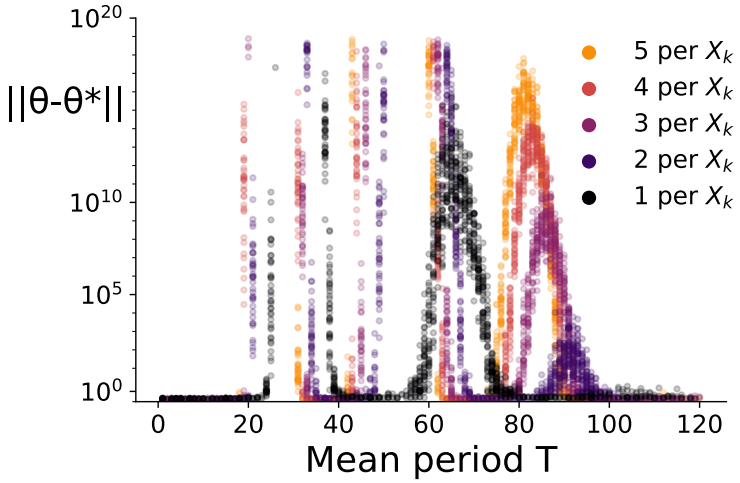


Figure 3.2: Regression w/ periodic $\bar{x}_k$. Warmer colors draw more samples from each X_k. Each dot is avg. distance $||\theta_k - \theta^*||$ over final 500 steps. Decreasing samples per X_k scales down frequency response and shifts peaks to the right, akin to mechanical damping. Each color has three peaks: the left peak is too large to appear on the y scale for all curves, the center peak is small enough to appear only for the black curve, and the right peak appears for all curves (vanishingly small for the black curve.) Resonant responses are dampened by increasing stochasticity in SGDm updates.

have access to noisy gradient estimates.

3.4 Experiment 4: Ablating Periodicity Further

We further depart from strictly periodic covariate shift by randomly sampling a new mean from a normal distribution $\mathcal{N}(0, v^2)$ after every T update steps. Since we no longer have periodic time variation Floquet theory no longer provides direct predictions of system behaviour. Nonetheless, we observe divergence characteristic of the parametric resonance we have seen thus far, with a specific band of mean switching intervals T having a highly divergent response. As seen in Figure 3.3, we observe a strong dependence on the variance v^2 of the distribution from which our means $\bar{x}_k$ are sampled, which is akin to the strong dependence of parametric resonance on driving signal amplitude.

Similar to the dependence on mean switching variance v^2 in Figure 3.3, we also observe a strong dependence on the number of input space dimensions, as seen in Figure 3.4. Also, the resonance occurs within the same band of mean

Table 3.4: Details for linear regression square wave covariate shift problem.

Square Wave Mean Period	$T \in [0, 120]$
Input Dimensionality	$d = 5$
Covariate Shift Mean	$\bar{x}_k = \begin{cases} \frac{\xi}{2\|\|\xi\|\|} \text{ if } \lfloor \frac{2k}{T} \rfloor \equiv 0 \mod 1 \\ - \quad \frac{\xi}{2\|\|\xi\|\|} \text{ if } \lfloor \frac{2k}{T} \rfloor \equiv 1 \mod 1 \end{cases}$ $\xi \sim \mathcal{N}(0, I_{d \times d})$ iid
Input Sampling	$X_k \sim \mathcal{N}(\bar{x}_k, 0.25 I_{d \times d})$
Target Function (fixed for all k)	$Y_k = \langle \theta^*_{[1:d]}, X_k \rangle + \theta^*_{d+1} + \epsilon_k$ $\theta^* \sim \mathcal{N}(0, c I_{d+1 \times d+1})$ where $c = 0.25$ $\epsilon_k \sim \mathcal{N}(0, 0.1)$
Model and Optimizer	$\widehat{Y}_k = \langle \theta_{[1:d]}, X_k \rangle + \theta_{d+1}$ $\theta_0 \sim \mathcal{N}(0, c I_{d+1 \times d+1})$ where $c = 0.25$ $(\theta_k)_{k \in \mathbb{N}} \leftarrow \text{SGDm}(\eta, \mu)$ $\eta = 0.01, \mu = 0.95$ $k \in [0, 10^4]$

switching intervals T. Given the sensitivity to driving signal amplitude in Figure 3.3, this aligns with the fact that the expected norm of samples drawn from a multivariate Gaussian increases with dimensionality, even with a fixed covariance scale. Though we are as yet unsure if this accounts for the entirety of resonance's dependence on input dimensionality.

3.5 Experiment 5: Ablating Optimizer Linearity

As explained in Section 2, a learning algorithm with linear time-varying gradients using SGDm as an optimizer corresponds to a discretization of an LTV ODE. If we change any one of those conditions, the learning system no longer corresponds to our LTV ODE. It is interesting to ask, however, whether resonance can be observed in systems which do not correspond to a discretization of our ODE.

To investigate this hypothesis we replace SGDm with ADAM as an update rule for our learning algorithm. We no longer have the ODE representation of

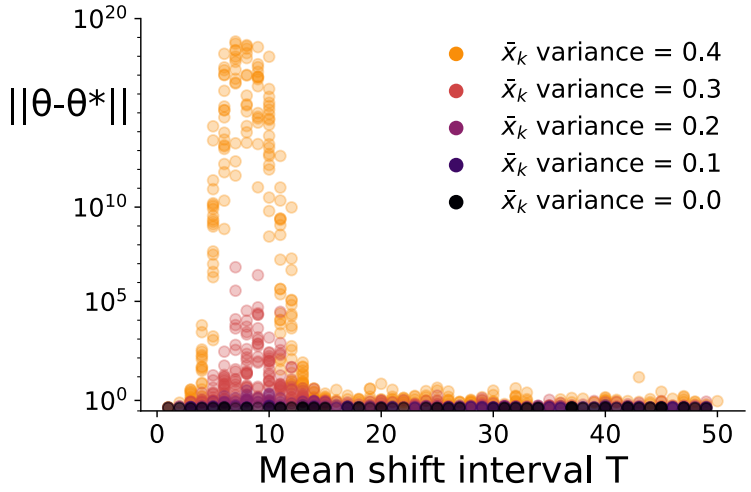


Figure 3.3: SGDm w/ stochastic $\bar{x}_k$, variance sensitivity. Regression w/ stochastic $\bar{x}_k$. Each marker is avg. distance $||\theta_k - \theta^*||$ over final 500 steps. Clearly, resonance occurs even without periodicity from Theorem 1, and resonance is very sensitive to the $\bar{x}_k$ signal variance (i.e. amplitude).

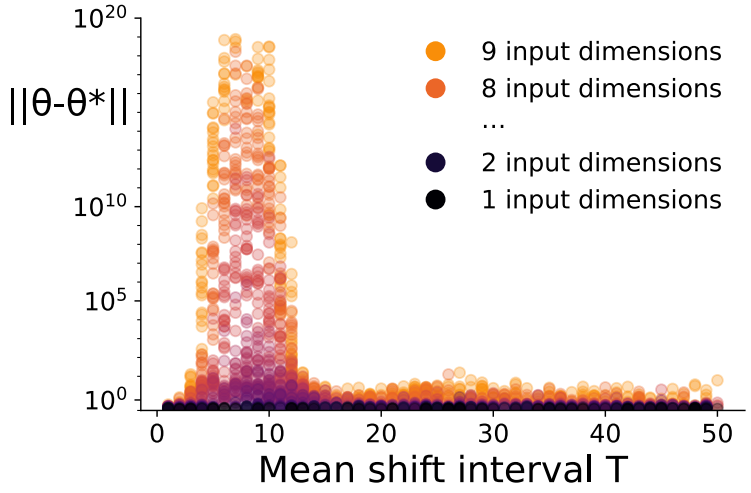


Figure 3.4: SGD w/ stochastic $\bar{x}_k$, sensitivity to d. Same configuration as Figure 3.3, but input dimensions d are varied instead of covariate shift variance. Resonance is very sensitive to the number of input dimensions d.

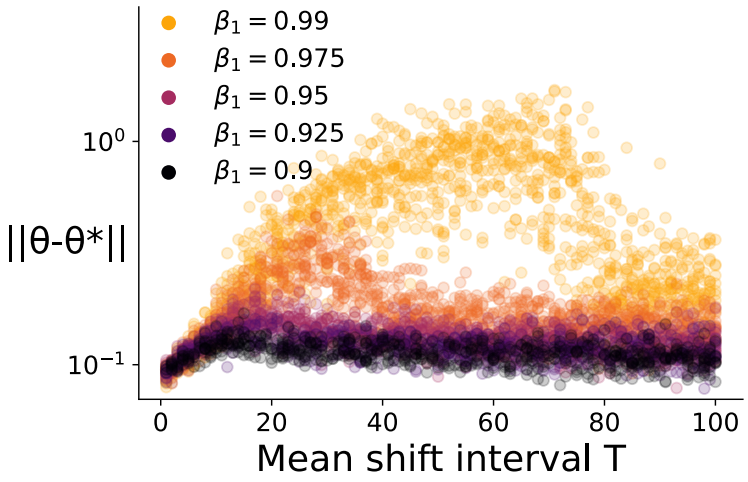


Figure 3.5: ADAM w/ stochastic $\bar{x}_k$, β_1 sensitivity. Regression w/ stochastic $\bar{x}_k$. Each marker is avg. distance $||\theta_k - \theta^*||$ over final 2000 steps. No exponential divergence as seen in Figures 3.3 and 3.4 suggests that frequency response is significantly damped for ADAM.

this learning system, but we can empirically investigate the behaviour of the system under covariate shift.

Here we use the same stochastic mean switching problem as the previous experiment, where we regress from 5 dimensions to 1 with input samples subject to a stochastic mean covariate shift. Again, we measure the distance between learned weights θ and target weights θ^*, but we replace the SGDm optimizer with ADAM, and we vary the ADAM parameter β_1. Similar to the SGDm optimizer, we can see in Figure 3.5 that there is a band of mean switching intervals T for which convergence is worse. But unlike SGDm, there is no divergence. Proper parametric resonance induces exponential divergence, which is why the previous results were presented with Euclidean distances between weights on the order of 10^{19}. Note that in Figure 3.5, the frequency response in weight space distance is instead measured on the order of 10^0.

3.6 Experiment 6: Ablating Model Linearity

Another relaxation of our conditions for the ODE representation is to use a function approximator without a quadratic loss surface. To this end, we replace our linear regression with a neural network. Like the previous experiment, we

Table 3.5: Details for linear regression problem with stochastically switching covariate shift mean.

Mean Switching Interval	$T \in [0, 50]$
Mean Switching Variance	$v = 0.25$ for Figure 3.4, $v \in [0, 0.4]$ for Figure 3.3
Input Dimensionality	$d \in [1, 9]$ for Figure 3.4, $d = 5$ for Figure 3.3
Covariate Shift Mean	$\bar{x}_k = \xi_i$ where $i = \left\lfloor \frac{k}{T} \right\rfloor$ $\xi_i \sim \mathcal{N}(0, vI_{d\times d})$ iid
Remaining Parameters	$X_k, Y_k, \widehat{Y}_k, \theta^*, \theta_0, (\theta_k)_{k\in\mathbb{N}}, \eta, \mu, k$ same as Table 3.4

no longer have correspondence with the ODE, but can empirically investigate the system's dynamics in terms of frequency response to covariate shift.

We reuse the mean switching covariate shift, but fit a nonlinear target function using a fully connected ReLU network. See Table 3.7 for details. In all previous experiments with linear regression, we measured frequency response as distance from target weights $||\theta_k - \theta^*||$. Here, we do not have access to the target weights, and so report frequency response in terms of loss L against the stationary distribution of $\{X_t\}$. In Figure 3.6, runs having loss < 0.05 converged to well-performing models, and those having loss > 0.3 perform badly.

In order to better highlight the frequency response, 10 samples were drawn from each X_k, so that the gradient signal is closer to the expected gradient. As depicted in Figure 3.6, there is a band of mean switching intervals for which convergence is damaged more than elsewhere, which is characteristic of resonance. However, the effect is bounded, which is a marked difference from previous experiments with SGDm and linear regression, where divergence was unbounded exponential.

Table 3.6: Details for linear regression problem with stochastically switching covariate shift, optimized with ADAM instead of SGDm.

Mean Switching Interval	$T \in [0, 100]$
Mean Switching Variance	$v = 1.0$
Input Dimensionality	$d = 5$
Covariate Shift Mean	$\bar{x}_k$ identical to Table 3.5
Input Sampling	$X_k \sim \mathcal{N}(\bar{x}_k, 0.1 I_{d\times d})$
Target Function (fixed for all k)	$Y_k, \theta^*, \epsilon_k$ identical to Table 3.5
Model	$\widehat{Y}_k, \theta_0$ identical to Table 3.5
Optimizer	$(\theta_k)_{k\in\mathbb{N}} \leftarrow \text{ADAM}(\eta, \beta_1, \beta_2)$ $\eta = 0.01, \beta_1 \in [0.9, 0.99], \beta_2 = 0.999$ $k \in [0, 10^4]$

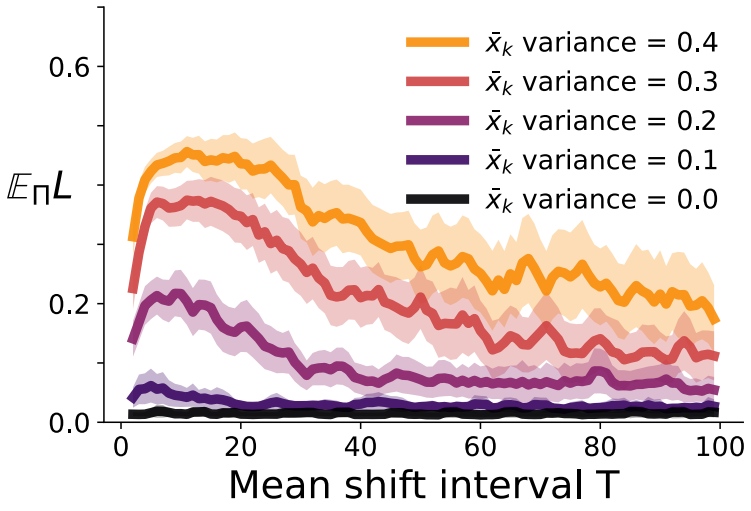


Figure 3.6: Training a neural network with SGDm shows a peak response in the loss around the band $T \in [5, 40]$. The y-axis is average test loss over the final 2000 training steps over 20 runs, with test set obtained via the stationary distribution of $\{X_k\}$. Shaded regions are 95% confidence intervals.

Table 3.7: Details for neural network regression problem with stochastically switching covariate shift.

Mean Switching Interval	$T \in [0, 100]$
Mean Switching Variance	$v \in [0, 0.4]$
Input Dimensionality	$d = 2$
Covariate Shift Mean	$\bar{x}_k$ identical to Table 3.5
Input Sampling	$X_k \sim \mathcal{N}(\bar{x}_k, 0.1 I_{d \times d})$
Target Function (fixed for all k)	$Y_k = \cos(\pi \lvert\lvert X_k \rvert\rvert) + \epsilon_k$ $\epsilon_k \sim \mathcal{N}(0, 0.1)$
Model	$\widehat{Y_k} = f(X_k; \theta_k)$ two hidden layers of 20 activations θ_0 initialized as He et. al.
Optimizer	$(\theta_k)_{k \in \mathbb{N}} \leftarrow \text{SGDm}(\eta, \mu)$ $\eta = 0.01, \mu = 0.95$ $k \in [0, 2 \times 10^4]$

Chapter 4

Reducing Resonant Responses

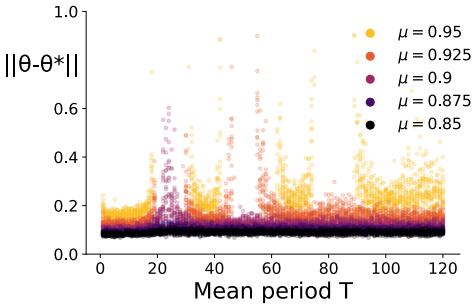
In experiments 3.1 and 3.2, we see that as the momentum parameter is reduced, the tendency to resonate is mitigated. This aligns with theoretical predictions, which (in the context of those particular experiments) suggest that a sufficiently low μ makes SGDm convergent across all frequencies in the band we evaluate. Intuitively, this aligns with the role μ plays in the ODE, as it appears in the following coefficient on the first order derivative of system (2.2) (i.e. the system's 'damping' coefficient), repeated below with the damping coefficient α explicitly labelled:

$$\ddot{\theta}(t) + \alpha\dot{\theta}(t) + B(t)(\theta(t) - \theta^*) = 0 \qquad\qquad \alpha := \frac{1-\mu}{\sqrt{\eta}} \tag{4.1}$$

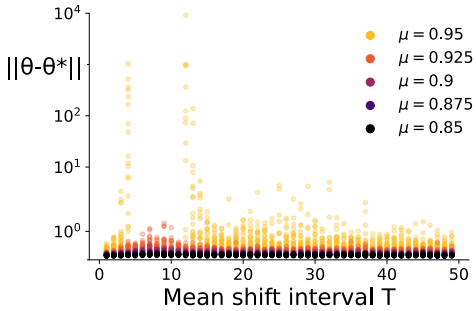
When $\alpha = 0$, the system has no damping (i.e. friction is zero, in the physical analogue) so resonant responses are maximized. As α increases, damping increases, and resonant responses are reduced. There are two ways to increase α: reduce μ or reduce η.

4.1 Reducing the momentum parameter

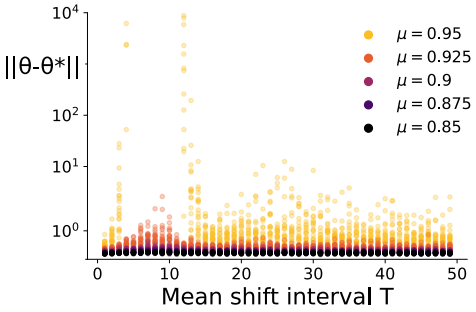
Figure 2.1b shows the theoretical heatmap across all possible momentum values, and across a much wider band of frequencies, which suggests a trend: setting μ to a sufficiently low value will completely mitigate resonance, though setting μ too low will worsen convergence rate. For now, we will set aside the observation that μ too low worsens convergence rate, and we will now show that reducing μ will reliably dampen resonance across all other experiments.



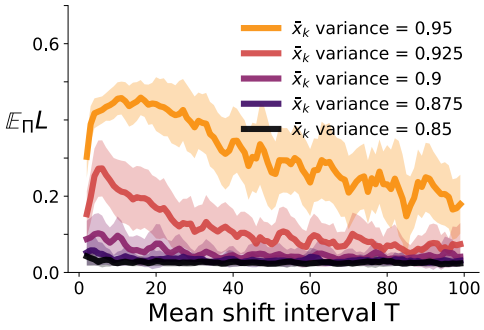
(a) Regression w/ periodic $\bar{x}_k$, 5 samples per X_k



(b) Regression w/ stochastic $\bar{x}_k$, 0.4 $\bar{x}_k$ variance



(c) Regression w/ stochastic $\bar{x}_k$, $d = 9$



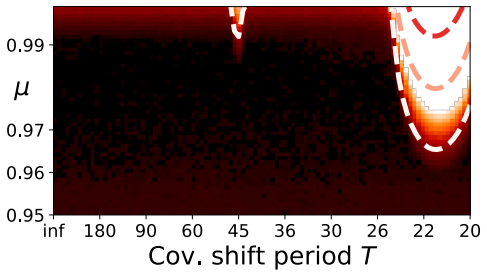
(d) Neural net w/ stochastic $\bar{x}_k$, $\bar{x}_k$ variance $= 0.4$

Figure 4.1: Re-running with reduced momentum values for highest resonance configurations chosen from Experiments 3.3 (a), 3.4 (b, c), and 3.6 (d). In all cases, reducing momentum significantly dampens resonant response, with $\mu = 0.85$ completely mitigating resonant response.

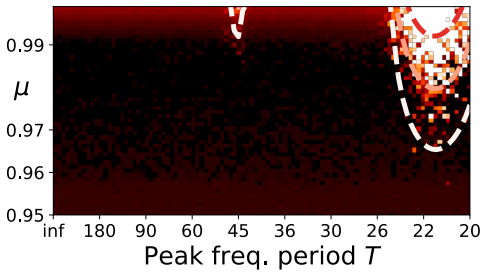
In 3.5, ADAM's parameter β_1 was varied, and we see that reducing β_1 decreases the tendency to resonate. Since β_1 is the nearest parameter to μ in SGDm, the desired trend has already been demonstrated. For the remainder of this section, we show the resonance-damping behaviour of μ in the remaining experiments: 3.3, 3.4, and 3.6. In particular, from each experiment we choose the configuration which had the highest tendency to resonate, and we modify the experiment by running them with several decreasing values of momentum μ. See Figure 4.1 for results. In Section 3, these experiments used momentum $\mu = 0.95$, and here we run with $\mu \in [0.85, 0.95]$, with all other experimental parameters identical.

4.2 Reducing the step-size

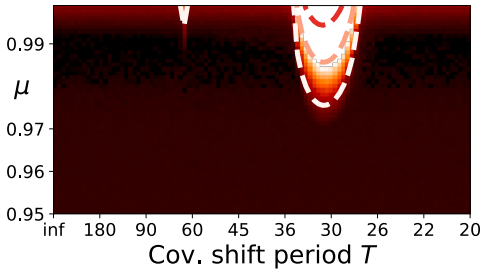
Another way to reduce tendency to resonate is suggested by the damping coefficient (4.1): reducing step-size η. Here we repeat Experiments 3.1 and 3.2, including two smaller step-sizes η to empirically demonstrate that trend. Specifically, Figure 4.2 shows the resonance heatmap and spectral radius contour lines for regression in two weights with covariate shift, identically to Experiments 3.1 and 3.2. The left column shows sinusoidal covariate shift, and the right column the AR(2) covariate shift. Each row corresponds to a fixed step-size $\eta \in \{0.01, 0.005, 0.001\}$, and it is clear that resonant, diverging regions are significantly reduced in size as step-size is decreased, with a narrower band of frequencies diverging, and the minimum momentum μ required for resonance increasing towards 1. This trend is reflected in both the empirical heatmap results, as well as the theoretically predicted spectral radius contour lines.



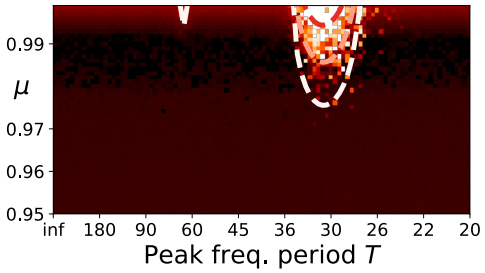
(a) Resonance regions, sinusoidal $\bar{x}_k$, $\eta = 0.01$



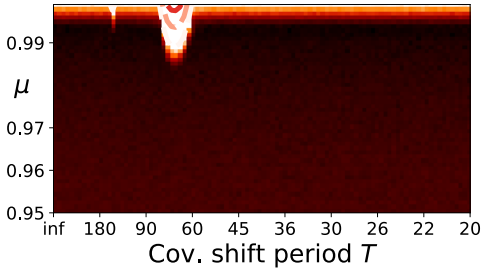
(b) Resonance regions, AR(2) $\bar{x}_k$, $\eta = 0.01$



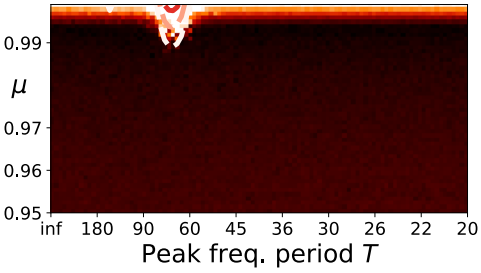
(c) Resonance regions, sinusoidal $\bar{x}_k$, $\eta = 0.005$



(d) Resonance regions, AR(2) $\bar{x}_k$, $\eta = 0.005$



(e) Resonance regions, sinusoidal $\bar{x}_k$, $\eta = 0.001$



(f) Resonance regions, AR(2) $\bar{x}_k$, $\eta = 0.001$

Figure 4.2: Re-running with reduced step-size for Experiments 3.1 (a, c, e) and 3.2 (b, d, f) and . In both cases, reducing step-size significantly dampens resonant response.

Chapter 5

Discussion and Limitations

In this work, we claim that the frequency response of SGDm is a useful property in determining when it will diverge due to the frequency content the training data process $\{X_k\}$ under non-iid sampling. While we highlight failure modes for SGDm, we do not intend to imply that SGDm is ineffective. Indeed, the heatmaps in Figures 3.1a and 3.1b indicate that SGDm converges much more often than it diverges, and that we can practically avoid resonance simply by reducing the momentum parameter μ. As per Section 4, we find that sufficiently low μ mitigates resonance.

In our theoretical analysis, we take expectation over observation noise and over the distribution of inputs X_k, so the gradient coefficient matrix $B(t)$ in the ODE (2.2) is deterministic. This allows us to rigorously characterize when resonance will occur, provided $B(t)$ is periodic. Future work might exploit the tools of stochastic differential equations in order to rigorously account for the effects of stochasticity. The case when $B(t)$ is aperiodic and stochastic is of particular interest, since Experiments 3.2 and 3.4 demonstrate that resonance still drives divergence with such $B(t)$. In particular, Experiment 3.2 aligns exactly with the theoretical predictions, suggesting that the resonance is in response to the frequency content of X_t, not its periodicity.

For the sake of consistency across experiments, our non-iid sampling is always induced by Gaussian X_k with a diagonal covariance matrix fixed over time. But Assumption 1 is far more relaxed. Even if one wishes to adhere to the strict periodicity requirements of Theorem 1, it is possible to induce time

variation by varying any aspect of X_k's distribution, including covariance structure. Comparing the effects of these different time variations is an interesting future direction.

To our knowledge, this work is the first to investigate SGDm under non-iid sampling from the ODE perspective. This perspective on non-iid sampling reveals new insights, in particular the revelation that SGDm under non-iid sampling numerically integrates a parametric oscillator, rather than the simple undriven harmonic oscillator induced by iid sampling. In terms of physical analogy, this is like the canonical spring-mass-damper oscillating system, but rather than being driven by a periodic external driving force (e.g. the wind), the system is driven by periodic changes in the stiffness of the springs. The former is a simple harmonic oscillator, and the latter a parametric oscillator. We leverage dynamical systems theory to analyze the stability of a learning system using SGDm under non-iid sampling, and are able to provide conditions for exponential divergence due entirely to the frequency content induced by the non-iid sampling.

This work connects the literature analyzing SGDm from an ODE perspective to the literature on stochastic gradient based algorithms under non-iid sampling. Despite its age and establishment, understanding SGDm is an ongoing process. We contribute to this process in a way that provides physical intuition, rather than a merely algebraic proof. Finally, we hope that the connection we have drawn to parametric oscillation in dynamical systems theory will provide a stepping stone towards more advanced work in understanding non-iid sampling and SGDm.

When assessing a new phenomenon like resonance, assessment on simple problems is a critical first step, and we have endeavored to be comprehensive in that assessment. But we only scratch the surface of more complex problems, including more advanced optimizers and nonlinear models, so it is not yet clear how big a role resonance plays in those settings. Similarly, for the sake of experimental control and interpretability of results, we assess resonance on synthetic data instead of real world data. While it is obvious that many sources of real data have nontrivial frequency content (e.g. audio samples, machine

control inputs, etc.) it is not yet clear when realistic frequency content will resonate with the system trying to learn from it. The two primary future directions for this work are to address these two gaps: to extend into the frequency *response* of more complex optimizers and models, and into the frequency *content* of real data.

Perhaps the most exciting future direction is the potential for engineered, system-specific hyperparameter choice. In order to avoid resonant response in a mechanical or electrical system under expected operations, engineers choose the parameters of dampers, linkages, capacitors, motors, and other components. These mechanical and electrical design choices are made via theoretical and empirical analysis in the frequency domain. That is, by examining system response to input signal frequencies. In the same way, machine learning engineers might choose hyperparameters to avoid resonant responses in online learning systems under covariate shift data conditions. Given sufficient understanding of the learning algorithm and the data, analytical guarantees may be possible using techniques which build upon those employed in Section 2. But even without such analytical guarantees, system design can still be aided by empirical observations of its frequency response. Such methods are used in established engineering fields [4], and they were the primary inspiration for the frequency sweep experiments in Section 3. Whether the methods are analytical or empirical, we are excited at the prospect of frequency domain methods for online learning system design.

References

[1] A. Ali, J. Z. Kolter, and R. J. Tibshirani, "A continuous-time view of early stopping for least squares regression," *International Conference on Artificial Intelligence and Statistics*, 2019.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *Neural Information Processing Systems*, 2019.

[3] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, "Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity," *Mathematical Programming*, 2018.

[4] P. Avitabile, *Modal Testing: A Practitioner's Guide*. John Wiley & Sons, 2017.

[5] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. Springer Science & Business Media, 2012.

[6] R. Berthier, F. Bach, N. Flammarion, P. Gaillard, and A. Taylor, "A continuized view on nesterov acceleration," *arXiv preprint arXiv:2102.06035*, 2021.

[7] M. Betancourt, M. I. Jordan, and A. C. Wilson, "On symplectic optimization," *arXiv preprint arXiv:1802.03653*, 2018.

[8] P. Csomós and I. Faragó, "Error analysis of the numerical solution of split differential equations," *Mathematical and Computer Modelling*, 2008.

[9] S. Csörgő and L. Hatvani, "Stability properties of solutions of linear second order differential equations with random coefficients," *Journal of Differential Equations*, 2010.

[10] J. Diakonikolas and L. Orecchia, "The approximate duality gap technique: A unified theory of first-order methods," *SIAM Journal on Optimization*, 2019.

[11] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Convergence rates of accelerated markov gradient descent with applications in reinforcement learning," *arXiv preprint arXiv:2002.02873*, 2020.

[12] ——, "Finite-time analysis of stochastic gradient descent under markov randomness," *arXiv preprint arXiv:2003.10973*, 2020.

[13] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, 2011.

[14] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic mirror descent," *SIAM Journal on Optimization*, 2012.

[15] I. Faragó, Á. Havasi, and R. Horvath, "On the order of operator splitting methods for time-dependent linear systems of differential equations," *International Journal of Numerical Analysis and Modeling*, Jan. 2011.

[16] G. Goh, "Why momentum really works," *Distill*, 2017. DOI: 10.23915/distill.00006. [Online]. Available: http://distill.pub/2017/momentum.

[17] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Science & Business Media, 2006.

[18] A. Halanay, "Stability theory," in *Differential Equations: Stability, Oscillations, Time Lags*. Elsevier, 1966, ch. 1, pp. 13–130.

[19] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a: Overview of mini-batch gradient descent," *Neural Networks for Machine Learning*, 2012.

[20] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," *International Conference on Machine Learning*, 2020.

[21] A. Iserles, "Euler's method and beyond," in *A First Course in the Numerical Analysis of Differential Equations*, 2nd ed. Cambridge University Press, 2008, ch. 1, pp. 3–18.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[23] N. B. Kovachki and A. M. Stuart, "Analysis of momentum methods," *arXiv preprint arXiv:1906.04285*, 2019.

[24] ——, "Continuous time analysis of momentum methods," *Journal of Machine Learning Research*, 2021.

[25] H. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Science & Business Media, 2003.

[26] V. Kuznetsov and M. Mohri, "Generalization bounds for time series prediction with non-stationary processes," *International Conference on Algorithmic Learning Theory*, 2014.

[27] Q. Li, C. Tai, and E. Weinan, "Stochastic modified equations and adaptive stochastic gradient algorithms," *International Conference on Machine Learning*, 2017.

[28] M. Muehlebach and M. I. Jordan, "A dynamical systems perspective on nesterov acceleration," *International Conference on Machine Learning*, 2019.

[29] ——, "Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives," *Journal of Machine Learning Research*, 2021.

[30] W. W. Mumford, "Some notes on the history of parametric transducers," *Proceedings of the Institute of Radio Engineers*, 1960.

[31] D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli, "Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms," *Neural Information Processing Systems*, 2020.

[32] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $o(1/k^2)$," *Proceedings of the USSR Academy of Sciences*, 1983.

[33] ——, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer Science & Business Media, 2014, ISBN: 1461346916.

[34] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, 1964.

[35] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Random averages, combinatorial parameters, and learnability," *Neural Information Processing Systems*, 2010.

[36] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 1951.

[37] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont, "Integration methods and optimization algorithms," *Neural Information Processing Systems*, 2017.

[38] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan, "Acceleration via symplectic discretization of high-resolution differential equations," *arXiv preprint arXiv:1902.03694*, 2019.

[39] J. W. Siegel, "Accelerated first-order methods: Differential equations and lyapunov functions," *arXiv preprint arXiv:1903.05671*, 2019.

[40] U. Simsekli, L. Zhu, Y. W. Teh, and M. Gurbuzbalaban, "Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise," *International Conference on Machine Learning*, 2020.

[41] W. Su, S. Boyd, and E. Candes, "A differential equation for modeling nesterov's accelerated gradient method: Theory and insights," *Neural Information Processing Systems*, 2014.

[42] T. Sun, Y. Sun, and W. Yin, "On markov chain gradient descent," *Neural Information Processing Systems*, 2018.

[43] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," *International Conference on Machine Learning*, 2013.

[44] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[45] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, 2016.

[46] A. C. Wilson, B. Recht, and M. I. Jordan, "A lyapunov analysis of momentum methods in optimization," *arXiv preprint arXiv:1611.02635*, 2016.

[47] H. Xiong, T. Xu, Y. Liang, and W. Zhang, "Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling," *arXiv preprint arXiv:2002.06286*, 2020.

[48] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie, "Direct runge-kutta discretization achieves acceleration," *Neural Information Processing Systems*, 2018.