THE UNIVERSITY OF CHICAGO

Stochastic Optimization in High-Dimensional Dynamical Systems

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

ELEANOR J. HARROW

CHICAGO, ILLINOIS

JUNE 2025

Unofficial template — not affiliated with or endorsed by the university.

Sample content only.

To my parents, Clara and Martin Harrow,

for teaching me that every question

deserves the patience of a thorough answer.

Abstract

High-dimensional dynamical systems pervade modern scientific inquiry, from climate modeling and neural population dynamics to financial market microstructure and genomic regulatory networks. This dissertation develops a unified framework for stochastic optimization in such systems by bridging three previously disconnected threads: spectral regularization of non-convex objectives, kernel-based sensitivity analysis under temporal dependence, and adaptive step-size schemes that exploit the geometry of the loss landscape. The central contribution is the **Spectral Annealed Gradient (SAG)** algorithm, which couples a preconditioned Langevin dynamics sampler with a spectrally adaptive temperature schedule, yielding provable convergence guarantees in the overparameterized regime. We validate the framework through extensive simulation studies on synthetic systems of up to 10^5 dimensions, and we demonstrate practical efficacy on two real-world problems: predicting short-horizon volatility surfaces from order-book data and inferring cortical connectivity graphs from multi-electrode array recordings. Our results establish that SAG outperforms Adam, RMSProp, and standard Langevin Monte Carlo across a suite of metrics including wall-clock convergence, test-set log-likelihood, and posterior coverage of credible intervals.

Keywords: stochastic optimization, Langevin dynamics, high-dimensional inference, spectral methods, dynamical systems

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has witnessed an explosion in the dimensionality of data collected across the natural and social sciences. High-throughput genomic sequencers, dense sensor arrays, and tick-by-tick financial data feeds routinely produce observations in spaces of dimension p that may exceed the sample size n by orders of magnitude. Classical inferential machinery — maximum likelihood, method of moments, and even modern Bayesian computation via Hamiltonian Monte Carlo — buckles under this regime, either because the requisite linear algebra scales superlinearly in p or because the geometry of the posterior distribution becomes pathologically concentrated.

This dissertation is concerned with the following question:

> *Given a high-dimensional dynamical system whose state evolves according to a stochastic differential equation (SDE) with unknown parameters $\theta \in \mathbb{R}^p$, how can we design an optimization procedure that recovers θ with statistical efficiency guarantees while remaining computationally feasible when $p \gtrsim 10^4$?*

We tackle this question from three complementary angles. First, we develop a spectral regularization framework that leverages the eigendecomposition of the empirical Fisher information matrix to identify and shrink directions of high curvature without requiring full Hessian access. Second, we introduce a kernel-based gradient estimator that exploits temporal structure in the underlying SDE to reduce the variance of stochastic gradients by an order of magnitude relative to naive subsampling. Third, we propose an adaptive

temperature schedule that anneals the injected noise in proportion to a local flatness measure, ensuring rapid mixing near saddle points without sacrificing convergence to the global minimizer.

1.1 Motivation and Scope

Consider the prototypical setting of a d-dimensional Itô process:

$$dX_t = f(X_t; \theta)\, dt + \sigma(X_t)\, dW_t, \tag{1.1}$$

where $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ is a smooth drift function, σ is a state-dependent volatility matrix, and W_t is a standard d-dimensional Brownian motion. Observations $y_1, \ldots, y_n$ arrive at discrete times $t_1, \ldots, t_n$, possibly corrupted by additive Gaussian noise. The log-likelihood $\ell_n(\theta) = \sum_{i=1}^{n} \log p(y_i \mid \theta)$ is almost always non-convex and may contain exponentially many suboptimal local minima when p is large [2].

Why should we expect any algorithm to succeed here? A growing body of work in the theory of overparameterized models [3, 7] suggests that the loss landscape of high-dimensional problems is generically benign: saddle points dominate local minima, and all global minima are connected by flat, low-loss valleys. Our contributions show how to exploit this structure algorithmically.

1.2 Summary of Contributions

1. **Chapter 2** introduces the Spectral Annealed Gradient (SAG) algorithm, proves its convergence in Wasserstein-2 distance, and characterizes its bias-variance trade-off through a sharp non-asymptotic bound.

2. **Chapter 3** develops the temporal kernel gradient estimator and establishes a central limit theorem for the resulting stochastic gradients under β-mixing.

3. **Chapter 4** presents the adaptive temperature schedule, which interpolates between exploration and exploitation phases without requiring manual tuning of a cooling rate parameter.

4. **Chapter 5** reports simulation benchmarks on synthetic dynamical systems of varying dimension, comparing SAG against six baseline optimizers.

5. **Chapter 6** applies the full framework to two empirical problems: volatility surface reconstruction from limit-order-book data, and functional connectivity estimation from neural spike trains.

Chapter 2

Spectral Annealed Gradient

This chapter develops the core algorithmic contribution of the dissertation. We begin by establishing notation and reviewing the relevant background on Langevin dynamics and spectral regularization, then present the SAG update rule and prove its main convergence guarantee.

2.1 Preliminaries

Let $\mathcal{L}(\theta) = -\ell_n(\theta)$ denote the empirical risk (negative log-likelihood). The unadjusted Langevin algorithm (ULA) [9] generates a Markov chain $\{\theta_k\}_{k\geq 0}$ via the recursion

$$\theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}(\theta_k) + \sqrt{2\eta\tau}\,\xi_k, \tag{2.1}$$

where $\eta > 0$ is the step size, $\tau > 0$ is the temperature parameter, and $\xi_k \sim \mathcal{N}(0, I_p)$ are independent Gaussian vectors. Under mild regularity conditions on $\mathcal{L}$, the stationary distribution of (2.1) is proportional to $\exp(-\mathcal{L}(\theta)/\tau)$, making ULA a natural tool for both sampling from and optimizing over the posterior.

However, when p is large and $\mathcal{L}$ is poorly conditioned, the naive ULA update (2.1) mixes prohibitively slowly. The condition number $\kappa = \lambda_{\max}(\nabla^2\mathcal{L})/\lambda_{\min}(\nabla^2\mathcal{L})$ can be on the order of 10^3 to 10^5 in the problems we study, forcing $\eta = O(1/\lambda_{\max})$ and yielding a mixing time that scales as $O(\kappa^2)$.

2.2 The SAG Algorithm

The SAG algorithm replaces the isotropic noise injection in (2.1) with a spectrally adapted counterpart. For each iteration k, we compute a low-rank approximation of the empirical Fisher information:

$$\hat{F}_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla \log p(y_i \mid \theta_k) \, \nabla \log p(y_i \mid \theta_k)^\top, \tag{2.2}$$

where $\mathcal{B}_k$ is a mini-batch of size B. Let $\hat{F}_k = U_k \Lambda_k U_k^\top$ be the eigendecomposition with $\Lambda_k = \operatorname{diag}(\lambda_{k,1}, \dots, \lambda_{k,p})$ and $\lambda_{k,1} \geq \cdots \geq \lambda_{k,p} \geq 0$. Define the truncated precision matrix

$$\Gamma_k = U_k \operatorname{diag}\big(\max(\lambda_{k,1}, \varepsilon), \dots, \max(\lambda_{k,p}, \varepsilon)\big)^{-1/2} U_k^\top, \tag{2.3}$$

with $\varepsilon = 10^{-8}$ for numerical stability. The SAG update is then

$$\boxed{\theta_{k+1} = \theta_k - \eta_k \, \Gamma_k \nabla \mathcal{L}(\theta_k) + \sqrt{2\eta_k \tau_k} \, \Gamma_k^{1/2} \, \xi_k}, \tag{2.4}$$

where η_k and τ_k follow adaptive schedules defined in Chapter 4.

Convergence in W_2 Distance

Assume $\mathcal{L}$ is M-smooth and μ-strongly convex outside a compact set. Let $\{\theta_k\}$ be generated by (2.4) with $\eta_k = \eta_0 k^{-\gamma}$ and $\tau_k = \tau_0 k^{-\delta}$ for $0 < \delta < \gamma < 1$. Then there exists a constant $C > 0$ such that for all $k \geq 1$,

$$W_2\big(\mathcal{L}(\theta_k), \, \pi_{\tau_k}\big) \leq C \, k^{-(\gamma - \delta)/2}, \tag{2.5}$$

where π_τ denotes the Gibbs distribution $\propto \exp(-\mathcal{L}(\theta)/\tau)$ and W_2 is the Wasserstein-2 distance.

The proof, deferred to Appendix A, hinges on a coupling argument that synchronizes two copies of the chain — one at equilibrium and one evolving from an arbitrary initialization

— and bounds the contraction induced by the preconditioner Γ_k.

2.3 Comparison with Related Methods

Table 2.1 summarizes the key differences between SAG and several widely used optimizers.

Table 2.1: Comparison of optimization methods for high-dimensional stochastic problems.

Method	**Preconditioning**	**Noise Model**	**Temp. Schedule**
Adam [6]	Diagonal EMA of squared grads	None	N/A
RMSProp [10]	Diagonal EMA of squared grads	None	N/A
ULA [9]	None	Isotropic Gaussian	Constant
SGLD [11]	None	Isotropic Gaussian	Decreasing
pSGLD [1]	RMSProp-style diagonal	Isotropic Gaussian	Decreasing
SAG (this work)	**Low-rank Fisher**	**Spectral**	**Adaptive**

Chapter 3

Temporal Kernel Gradient Estimation

When observations arise from an SDE as in (1.1), the temporal ordering carries information that is discarded by standard i.i.d. subsampling. This chapter introduces a kernel-weighted gradient estimator that exploits the β-mixing property of discretely observed diffusion processes.

3.1 Kernel Construction

For observations at times $0 = t_0 < t_1 < \cdots < t_n = T$, define the temporal kernel

$$K_h(i, j) = \frac{1}{h}\,\phi\left(\frac{|t_i - t_j|}{h}\right), \tag{3.1}$$

where $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a symmetric, compactly supported kernel (we use the Epanechnikov kernel throughout) and $h > 0$ is the bandwidth. The kernel-smoothed gradient at iteration k is

$$\tilde{g}_k(\theta) = \frac{1}{B} \sum_{i \in \mathcal{B}_k} \sum_{j=1}^{n} K_h(i, j)\, \nabla \log p(y_j \mid \theta). \tag{3.2}$$

The inner sum over j down-weights observations that are temporally distant from i, reducing the contribution of high-variance gradient estimates from independent time blocks while preserving the low-bias signal from adjacent observations.

3.2 Theoretical Properties

Asymptotic Normality of $\tilde{g}_k$

Let the process $\{X_t\}$ in (1.1) be geometrically β-mixing with mixing coefficient $\beta(t) \leq c_0 e^{-c_1 t}$. Fix θ and let $h = h_n \to 0$ such that $nh_n \to \infty$. Then, as $n \to \infty$,

$$\sqrt{B}\left(\tilde{g}_k(\theta) - \mathbb{E}[\nabla \log p(y \mid \theta)]\right) \xrightarrow{d} \mathcal{N}\big(0,\ \Sigma_\phi(\theta)\big), \tag{3.3}$$

where $\Sigma_\phi(\theta)$ is a positive semi-definite matrix with $\operatorname{tr}(\Sigma_\phi) \leq \operatorname{tr}(\Sigma_{\text{iid}}) \,/\, (1 + c_2 h^{-1})$ for some $c_2 > 0$. In words, the kernel estimator achieves a variance reduction factor of approximately $1 + c_2/h$ relative to the standard i.i.d. estimator.

Key insight

Temporal dependence is not a nuisance to be corrected — it is a structural feature that, when properly harnessed through kernel smoothing, can *reduce* gradient variance below the i.i.d. baseline.

The bandwidth h controls the bias-variance trade-off: small h recovers the raw i.i.d. estimator, while large h averages over long temporal windows, potentially introducing bias if the latent state drifts substantially. In practice, we select h via a plug-in estimator derived from the empirical autocorrelation function of the gradient sequence.

Chapter 4

Empirical Evaluation

This chapter presents the experimental validation of the SAG framework. All experiments were conducted on a compute cluster equipped with 64-core AMD EPYC processors and Lumina A100 GPUs, using a custom implementation in JAX [5]. Code reproducing all figures and tables is publicly available at `https://github.com/example/sag-code`.

4.1 Synthetic Benchmarks

We consider three families of synthetic test problems, summarized in Table 4.1.

Table 4.1: Synthetic benchmark problems. p = parameter dimension; n = sample size; κ = condition number of Hessian at optimum.

Problem	p	n	κ	Description
Logistic bandit	5×10^3	10^5	820	Binary outcomes with sparse signal
Lorenz-96 SDE	1×10^4	2×10^5	4.3×10^3	Chaotic atmospheric toy model
Cox–Ingersoll–Ross	2×10^4	5×10^4	1.1×10^4	Multi-factor interest rate dynamics

For each problem, we compare SAG against six baseline optimizers: Adam, RMSProp, Adagrad, standard ULA, SGLD [11], and preconditioned SGLD (pSGLD) [1]. All methods are given a fixed computational budget of 10^5 gradient evaluations.

4.1.1 Convergence Speed

Figure **??** (appearing in the List of Figures) plots the negative log-likelihood against wall-clock time for the Lorenz-96 problem. SAG reaches a test log-likelihood of -2.87 within 1.2×10^3 seconds, compared to -3.41 for Adam at the same time point and -4.02 for standard ULA. The spectral preconditioner yields a roughly 2.5× speedup relative to the next-best method (pSGLD) on this problem class.

4.1.2 Posterior Coverage

Table 4.2: Empirical coverage of 95% credible intervals across methods on the logistic bandit problem. Nominal coverage target is 0.95.

Method	Mean Coverage	Interval Width
Adam (bootstrap)	0.82	1.24
RMSProp (bootstrap)	0.79	1.31
ULA	0.88	1.08
SGLD	0.91	0.96
pSGLD	0.93	0.91
SAG	**0.94**	**0.87**

SAG achieves 94% mean coverage with the narrowest intervals among all methods, indicating that the spectral noise injection successfully calibrates uncertainty without sacrificing precision. The gap between SAG and pSGLD narrows as dimension grows, consistent with our theoretical analysis showing that the advantage of low-rank preconditioning is most pronounced in the $p \gg n$ regime.

4.2 Real-Data Applications

4.2.1 Volatility Surface Reconstruction

We apply SAG to reconstruct implied volatility surfaces from Level-2 limit-order-book data for S&P 500 options during the volatile trading months of January–March 2024. The

dataset, sourced from the Market Information Data Analytics System (MIDAS) through a collaborative agreement with Vertex Quantitative Research, contains approximately 8×10^7 order-book snapshots at millisecond resolution.

The parameter vector θ encodes the coefficients of a stochastic volatility inspired (SVI) surface parameterization with $p = 3{,}186$ free parameters covering strikes ranging from $0.5\times$ to $2.0\times$ spot and maturities from one week to two years. Table 4.3 reports the out-of-sample root mean squared error (RMSE) in implied volatility units.

Table 4.3: Out-of-sample RMSE for volatility surface reconstruction. Lower is better.

Method	RMSE (weekly)	RMSE (monthly)
Adam	0.0241	0.0187
RMSProp	0.0258	0.0193
SGLD	0.0194	0.0151
pSGLD	0.0182	0.0144
SAG	**0.0156**	**0.0119**

4.2.2 Cortical Connectivity Inference

The second application concerns inferring functional connectivity graphs from multi-electrode array recordings of macaque primary visual cortex, collected in collaboration with the Nimbus Neuroscience Institute. The data consist of $n = 1.2 \times 10^6$ spike-time events across $d = 96$ electrodes during presentation of drifting grating stimuli. We fit a multivariate Hawkes process with a sparse graph-structured excitation matrix $\theta \in \mathbb{R}^{96 \times 96}$.

The inferred connectivity graph produced by SAG recovers known anatomical pathways with 82% precision and 74% recall, outperforming the lasso-regularized maximum-likelihood baseline (64% precision, 58% recall) reported in [8]. Notably, SAG identifies a previously unreported reciprocal connection between layers 2/3 and layer 5 of area V1, a finding that aligns with recent anatomical tracing studies [4].

Acknowledgments

I am deeply grateful to my advisor, Professor Jonathan K. Redwood, whose intellectual generosity and exacting standards shaped every page of this dissertation. I thank the members of my committee — Professors Sanjay Mehta, Diana L. Fairchild, and Thomas Wren — for their careful reading and incisive feedback. The Vertex Quantitative Research group provided computational resources and stimulating discussions; the Nimbus Neuroscience Institute generously shared electrophysiological data. Any errors that remain are mine alone.

Financial support was provided by a Synapse Foundation Graduate Fellowship and by NSF grant DMS-2145678 (PI: J.K. Redwood).

This dissertation was typeset using the LaTeX typesetting system with the New TX font family and custom UChicago-inspired styling. Unofficial template — not affiliated with the university.

References

[1] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *ICML*, 2012.

[2] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[3] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *JMLR*, 20:1–46, 2019.

[4] K. D. Harris, R. Q. Quiroga, and G. Buzsáki. Cortical connectivity and the structure of population codes. *Neuron*, 101(2):260–275, 2019.

[5] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paske, et al. JAX: composable transformations of Python+NumPy programs, 2018. https://github.com/jax-ml/jax.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[7] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. The role of over-parameterization in generalization. In *NeurIPS*, 2018.

[8] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[9] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[10] T. Tieleman and G. Hinton. Lecture 6.5: RMSProp. Coursera: Neural Networks for Machine Learning, 2012.

[11] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.