UNIVERSITY OF CALIFORNIA,
IRVINE

Studies of Residual Diffusivity and Curvature Dependent Effective Velocity in Fluid Flows by Analytical and Mechine Learning Methods

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Jiancheng Lyu

Dissertation Committee:
Professor Jack Xin, Chair
Professor Yifeng Yu
Professor Long Chen

2018

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

List of Algorithms

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my two advisors Prof. Jack Xin and Prof. Yifeng Yu for their continuous support and help since I entered the program. It was their smart ideas, invaluable guidance and endless patience that made my Ph.D. career smooth and enjoyable. I have benefited a lot from not only their immense knowledge but their perspectives as great mathematicians.

I would like to thank Prof. Long Chen for serving as my committee member. My sincere thank also goes to my collaborators Dr. Penghang Yin, Dr. Shuai Zhang, Dr. Yingyong Qi and Prof. Stanley Osher for the simulating discussions.

Last but not the least, I would like to thank my family members for supporting me throughout persuing this degree and my life in general.

This work was partially supported by NSF CAREER award DMS-1151919 and NSF grants DMS-1211179, DMS-1522383 and IIS-1632935.

CURRICULUM VITAE

Jiancheng Lyu

EDUCATION

Doctor of Philosophy in Applied Mathematics **2018**
University of California at Irvine *Irvine, California, U.S.A.*

Master of Science in Mathematics **2013**
Peking University *Beijing, China*

Bachelor of Science in Mathematics **2010**
Peking University *Beijing, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2018**
University of California at Irvine *Irvine, California, U.S.A.*

TEACHING EXPERIENCE

Teaching Assistant **2015–2017**
University of California at Irvine *Irvine, California, U.S.A.*

Teaching Assistant **2011–2013**
Peking University *Beijing, China*

ABSTRACT OF THE DISSERTATION

Studies of Residual Diffusivity and Curvature Dependent Effective Velocity in Fluid Flows
by Analytical and Mechine Learning Methods

By

Jiancheng Lyu

Doctor of Philosophy in Mathematics

University of California, Irvine, 2018

Professor Jack Xin, Chair

In chaotic advection generated by a class of time periodic cellular flows, the residual diffusion refers to the non-zero effective (homogenized) diffusion in the limit of zero molecular diffusion as a result of chaotic mixing of the streamlines. We study the residual diffusion phenomenon computationally and analytically.

We make use of the Poincaré map of the advection-diffusion equation to bypass long time simulation and gain accuracy in computing effective diffusivity and learning adaptive basis. We observe a non-monotone relationship between residual diffusivity and the amount of chaos in the advection, though the overall trend is that sufficient chaos leads to higher residual diffusivity. The adaptive orthogonal basis with built-in sharp gradient structures is constructed by taking snapshots of solutions in time, preprocessing with deep neural network (DNN) if necessary and performing singular value decomposition of the matrix consisting of those snapshots as column vectors. The trained orthogonal adaptive basis makes possible low cost computation of the effective diffusivities at smaller molecular diffusivities. The testing errors decrease as the training occurs at smaller molecular diffusivities.

We also study the enhanced diffusivity in the so called elephant random walk model with stops by including symmetric random walk steps at small probability ϵ. At any $\epsilon > 0$, the large time behavior transitions from sub-diffusive at $\epsilon = 0$ to diffusive in a wedge shaped parameter regime where the diffusivity is strictly above that in the un-perturbed model in the $\epsilon \downarrow 0$ limit. The perturbed model is shown to be solvable with the first two moments and their asymptotics calculated exactly in both one and two space dimensions. The model provides a discrete analytical setting of the residual diffusion phenomenon as molecular diffusivity tends to zero. On a related nonlinear case, we give theoretical proof that the turbulent flame speed as an effective burning velocity is decreasing with respect to the curvature diffusivity (Markstein number) for shear flows in the well-known G-equation model. Besides, we solve the selection problem of weak solutions when the Markstein number goes to zero and solutions approach those of the inviscid G-equation model. The limiting solution is given by a closed form analytical formula.

Finally for the dimensionality reduction on DNNs, we propose BinaryRelax, a simple two-phase algorithm, for training DNNs with quantized weights. We relax the hard constraint that characterizes the quantization of weights into a continuous regularizer via Moreau envelope, which turns out to be the squared Euclidean distance to the set of quantized weights. The pseudo quantized weights are obtained by linearly interpolating between the float weights and their quantizations. A continuation strategy is adopted to push the weights towards the quantized state by gradually increasing the regularization parameter. We test BinaryRelax on the benchmark CIFAR and ImageNet color image datasets to demonstrate the superiority of the relaxed quantization approach and the improved accuracy over the state-of-the-art training methods. Moreover, we prove the convergence of BinaryRelax under an approximate orthogonality condition.

Chapter 1

Introduction

1.1 Residual Diffusion and Turbulent Flame Speed

Diffusion enhancement in fluid advection has been studied for nearly a century, dating back to the pioneering work of Taylor [66] in 1921. It is a fundamental problem to characterize and quantify the large scale effective diffusion (denoted by D^E) in fluid flows containing complex and turbulent streamlines. Much progress has been made based on the passive scalar model [47]:

$$T_t + (\boldsymbol{v} \cdot D)\, T = D_0\, \Delta T, \tag{1.1}$$

where T is a scalar function (e.g. temperature or concentration), $D_0 > 0$ is a constant (the so-called molecular diffusion), $\boldsymbol{v}\,(x,t)$ is a prescribed incompressible velocity field, D and Δ are the spatial gradient and Laplacian operators.

When the flow is steady, periodic and two dimensional, precise asymptotics of D^E are known. A prototypical example is the steady cellular flow [17, 26], $\boldsymbol{v} = (-H_{x_2}, H_{x_1})$, $H = \sin x_1 \sin x_2$, see also [53, 70, 71] for its application in effective speeds of front propagation. The asymptotics of the effective diffusion along any unit direction in the cellular flow obeys the square root law in the advection dominated regime: $D^E = O(\sqrt{D_0}) \gg D_0$ as $D_0 \downarrow 0$, [26, 29]. However, if the streamlines are fully chaotic (well-mixed), the enhancement can follow a very different law. The simplest example is the time periodic cellular flow:

$$\boldsymbol{v} = (\cos(x_2), \cos(x_1)) + \theta\ \cos(t)\ (\sin(x_2), \sin(x_1)), \quad \theta \in (0, 1]. \tag{1.2}$$

The first term of (1.2) is a steady cellular flow with a $\pi/4$ rotation, and the second term is a time periodic perturbation that introduces an increasing amount of disorder in the flow trajectories as θ becomes larger. At $\theta = 1$, it is fully mixing, and empirically subdiffusive [85]. The flow (1.2) has served as a model of chaotic advection for Rayleigh-Bénard experiment [8]. Numerical simulations [5, 51] suggest that at $\theta = 1$, the effective diffusion along the x_1-axis, $D^E_{11} = \mathcal{O}(1)$ as $D_0 \downarrow 0$, the so called *residual diffusion* arises. As $D_0 \downarrow 0$, the solutions develop sharp gradients, and render accurate computation costly, especially if one is interested in D^E parametrized by θ.

Recall the formula for effective diffusivity tensor [5]:

$$D^E_{ij} = D_0 \left(\delta_{ij} + \langle Dw_i \cdot Dw_j \rangle\right), \tag{1.3}$$

where w is a mean zero space-time periodic vector solution of:

$$w_t + (\boldsymbol{v} \cdot Dw) - D_0 \Delta w = -\boldsymbol{v}, \tag{1.4}$$

and the bracket denotes space-time average over the periods. The solution of (1.4) is unique by the Fredholm alternative. The correction to D_0 is positive definite in (1.3). We will be focusing on the singular solutions of (1.4) at small D_0. In Chapter 2, we compute w by the spectral method since Fourier basis can represent cellular flow with few modes [44]. By truncating the Fourier expansion, we find an approximate system of ordinary differential equations (ODEs). The time periodic solution is constructed as the unique fixed point of the Poincaré map of the ODE's time 2π flow. Formula (1.3) is then used to calculate D^E.

The motion of a diffusing particle in the flow (1.2) also satisfies the stochastic differential equation (SDE):

$$dX_t = \boldsymbol{v}(X_t, t)\, dt + \sqrt{2\, D_0}\, dW_t, \quad X(0) = (x_0, y_0) \in \mathbf{R}^2, \tag{1.5}$$

where $D_0 > 0$ is molecular diffusivity as above, W_t is the standard 2-dimensional Wiener process. The effective diffusivity in the unit direction e can be given by mean square displacement at large times [3]:

$$D^E\,(D_0, e, \theta) = \lim_{t \uparrow +\infty} E(|(X(t) - X(0)) \cdot e|^2)/t. \tag{1.6}$$

In Chapter 3, we analyze the residual diffusion phenomenon in a random walk model which is solvable in the sense of moments and has certain statistical features of the SDE model (1.5). The baseline random walk model is the so-called elephant random walk model with stops (ERWS) [35] which is non-Markovian and exhibits sub-diffusive, diffusive and super-diffusive regimes. The ERWS plays the role of flow (1.2) in that there is a sub-diffusive statistical regime, which is absent in the earlier version of the ERW model without stops [63]. Stops in random walk models are often interpreted as occasional periods of rest during an animal's movement [68]. Recall that the chaotic system from (1.5) is sub-diffusive [85] at $D_0 = 0$ and transitions to diffusive with residual diffusion at $D_0 > 0$. To mimic this in the ERWS model, we add a small probability of symmetric random walk in the sub-diffusive regime and examine the large time behavior of the mean square displacement [45]. Interestingly, the sub-diffusive regime also transitions into diffusive regime and a wedge shaped parameter region appears where the diffusivity is strictly above that of the baseline ERWS model in the zero probability limit of the symmetric random walk (analogue of the zero molecular diffusivity limit). In the context of animal dispersal in ecology, the emergence of residual diffusion indicates that the large time statistical behavior of the movement can pick up positive normal diffusivity when the animal's rest pattern is slightly disturbed consistently in time. We also extend our analysis to a two dimensional ERWS model (see [20] for a related solvable model).

A nonlinear case in multi-scale fluid dynamics is the propagation of flame, which can be formulated as the following G-equation model

$$G_t + V(x) \cdot DG + s_l \, |DG| = 0 \quad \text{in } \mathbf{R}^n \times (0, \infty).$$

Here the zero level set of $G(x,t)$ represents the flame front, and the burnt and unburnt regions are $\{G(x,t) < 0\}$ and $\{G(x,t) > 0\}$, respectively, see Figure 1.1. The velocity of ambient fluid $V : \mathbf{R}^n \to \mathbf{R}^n$ is assumed to be smooth, $\mathbf{Z}^n$-periodic and incompressible (i.e. $\mathrm{div} V = 0$). The G-equation follows from a simple motion law: $\vec{v}_n = s_l + V(x) \cdot n$, i.e. the normal velocity is the laminar flame speed s_l plus the projection of V along the normal direction.

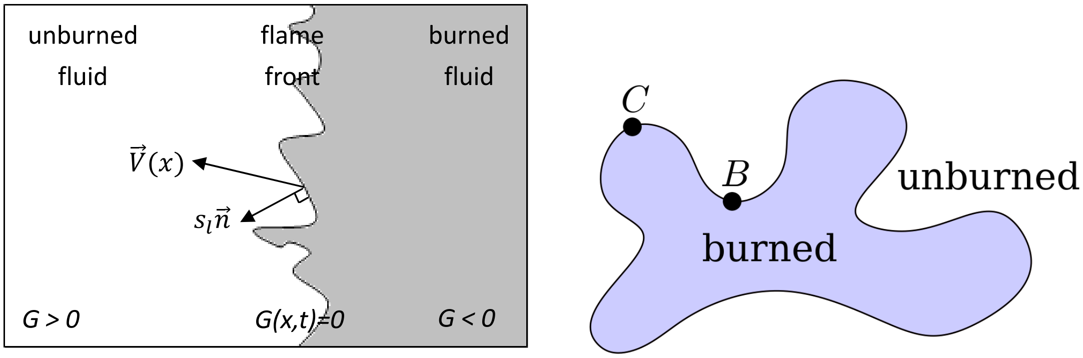


Figure 1.1: Level-set formulation of front propagation and curvature effect.

The curvature effect in turbulent combustion was first studied by Markstein [48], which says that if the flame front bends toward the cold region (unburned area, point C in Figure 1.1), the flame propagation slows down. If the flame front bends toward the hot spot (burned area, point B in Figure 1.1), it burns faster. An empirical linear relation proposed by Markstein [48] to approximate the dependence of the laminar flame speed on the curvature (see also [56], [64], etc) is

$$s_l = s_l^0(1 - \tilde{d}\,\kappa). \tag{1.7}$$

Here s_l^0, the mean value, is a positive constant. The parameter $\tilde{d} > 0$ is the so-called Markstein length proportional to the flame thickness. The mean curvature along the flame front is κ. κ changes sign along a curved flame front in general. Plugging the expression of the laminar flame speed (1.7) into the G-equation and normalizing the constant $s_l^0 = 1$, we obtain a mean curvature type equation

$$G_t + V(x) \cdot DG + |DG| - \tilde{d}\,|DG|\,\mathrm{div}\left(\frac{DG}{|DG|}\right) = 0. \tag{1.8}$$

A mathematically interesting and physically important question is: *how does the "averaged" flame propagation speed depend on the curvature term?* There is a consensus in combustion literature that the curvature effect slows down flame propagation [61]. Heuristically, this is because the curvature term smooths out the flame front and reduces the total area of chemical reaction [64]. However, this folklore has never been rigorously justified mathematically. The decrease of turbulent flame speed with respect to the Markstein number has been experimentally observed (e.g., [15]). In Chapter 4, we consider the shear flow:

$$V(x) = (v(x_2), 0) \quad \text{for } x = (x_1, x_2) \in \mathbf{R}^2,$$

where $v : \mathbf{R} \to \mathbf{R}$ is a smooth periodic function. Establishing a highly sophisticated class of inequalities, we prove the average flame speed (an effective burning velocity) is decreasing with respect to the curvature diffusivity (Markstein number) for shear flows [46].

1.2 Adaptive Basis Learning and Dimensionality Reduction on DNNs

In Chapter 5, we shall do dimensionality reduction on solving the cell problem (1.4) and construct adaptive basis functions to handle the singular solutions. The snapshots of solutions of (1.4) in the time interval $[0, 2\pi]$ are saved into columns of a matrix W. Adaptive basis can be learned from W at a few sampled D_0 or θ values. The equation (1.4) at other D_0 or $\theta \in (0, 1]$ will be solved in terms of the adaptive basis trained at the closest sample D_0 or θ value. We shall see that the number of adaptive basis functions is under a few hundred, much less than that of Fourier basis by several orders of magnitude. The relative error of the adaptive solution from a resolved spectral solution is under 6.5 % when testing at $D_0 = 10^{-5}$ and training at $D_0 = 10^{-4}$. Thus we manage to achieve accurate enough solutions at much lower costs in the regime of small D_0 where the number of Fourier basis functions grows rapidly.

A straightforward way to find adaptive basis functions is to compute left singular eigenvectors corresponding to the top singular values of W [44]. The procedure of taking snapshots and performing singular value decomposition (SVD) is standard in reduced order modeling [57] and is known as proper orthogonal decomposition (POD) in the fluid dynamics literature [43, 30]. The residual diffusivity problem we study here however offers an ideal testing ground for the evaluation of POD which lacks theoretical guarantees in general. The success of POD relies on the underlying dynamics being governed by a unique low dimensional attractor. In our case, the time periodicity of $\boldsymbol{v}$ helps to reduce the evolution problem (1.4) to a Poincaré map problem. The snapshots (training data) are directly drawn from the time periodic solution, hence more effective for learning.

It is observed that as D_0 gets smaller, thinner structures arises in the solution. A natural approach to improve the adaptive basis learning is to incoperate in the basis functions that characteristic. In other words, we would like to have adaptive basis with thinner layer structures. One way to get sharpened adaptive basis is to compute modified "sparse PCA" of solution matrix W, where the L^1 norm of the gradient of basis functions is added as a penalty. However, the modified "sparse PCA" usually involves convergence issue and is not efficient especially when the orthogonality constraint is imposed. Here we try pre-processing of W instead and map W to its counterpart with thinner structures. That procedure is quite similar to super-resolution [72, 84, 58], i.e. reconstructing a high-resolution image from its low-resolution counterpart. Thus we modify a generative adversial network based super-resolution [37] (SRGAN) and apply it to the preprocessing, where W's at larger D_0 and smaller D_0 are used for input and target data respectively during training. Hence instead of computing SVD for some particular W itself, we feed the trained deep neural network (DNN) model with W and find singular vectors of the output.

We are also interested in dimensionality reduction on DNNs themselves. There is a growing interest in deploying DNNs on low-power embedded systems with limited memory storage and computing power, such as cell phones and other battery-powered devices. However, DNNs typically require hundreds of megabytes of memory storage for the trainable full-precision floating-point parameters or weights, and need billions of FLOPs to make a single inference. Recent efforts have been devoted to the training of DNNs with coarsely quantized weights which are represented using low-precision (8 bits or less) fixed-point arithmetic [31, 19, 39, 81, 82, 76, 80, 54, 7, 79, 41]. Quantized neural networks enable substantial memory savings and computation/power efficiency, while achieving competitive performance with that of full-precision DNNs. Moreover, quantized weights can exploit hardware-friendly bit-wise operations and lead to dramatic acceleration at inference time.

The simplest way to perform quantization would be directly rounding the weights of a pre-trained full-precision network, which often leads to poor accuracy at bit-width under 8. From the perspective of optimization, the training of quantized networks can be naturally abstracted as a constrained optimization problem of minimizing some empirical risk subject to a set constraint that characterizes the quantization of weights:

$$\min_{x\in\mathbf{R}^n} f(x) := \frac{1}{N}\sum_{j=1}^{N} \ell_j(x) \quad \text{subject to} \quad x \in \mathcal{Q}. \tag{1.9}$$

Given a training sample of input I_j and label u_j, the corresponding training loss is

$$\ell_j(x) = \ell(\sigma_l(x_l * \cdots \sigma_1(x_1 * I_j)), u_j),$$

where $x = [x_{(1)}^\top, \dots, x_{(l)}^\top]^\top$ and $x_{(i)} \in \mathbf{R}^{n_i}$ contains the n_i weights in the i-th linear (fully connected or convolutional) layer with $\sum_{i=1}^{l} n_i = n$, σ_i is some element-wise nonlinear function. "$*$" denotes either matrix-vector product or convolution operation; reshaping is necessary to avoid mismatch in dimensions. For layer-wise quantization, the set $\mathcal{Q}$ takes the form of $\mathcal{Q}_1 \times \cdots \times \mathcal{Q}_l$, where $x_{(i)} \in \mathcal{Q}_i := \mathbf{R}_+ \times \{\pm q_1, \pm q_2, \dots, \pm q_m\}^{n_i}$. Here $\mathbf{R}_+$ denotes the set of nonnegative real numbers and $0 \le q_1 < q_2 < \cdots < q_m$ represent the m quantization levels and are pre-determined. The weight vector in the i-th layer enjoys the factorization $x_{(i)} = s_i \cdot Q_{(i)}$ for some $Q_{(i)} \in \{\pm q_1, \pm q_2, \dots, \pm q_m\}^{n_i}$ and some trainable layer-wise scalar $s_i \ge 0$. s_i is shared by all weights across the i-th linear layer and will be stored separately from the quantized numbers q_i for deployment efficiency. The storage for the scaling factors is *negligible* as there are so few of them. Weight quantization has two special cases as follows.

- 1-bit binarization: $m = 1$ and $\mathcal{Q}_i = \mathbf{R}_+ \times \{\pm 1\}^{n_i}$. The storage of $Q_{(i)}$'s only needs 1 bit for representing the signs. Compared to the full-precision model, we have 32× memory savings.
- 2-bit ternarization: $m = 2$ and $\mathcal{Q}_i = \mathbf{R}_+ \times \{0, \pm 1\}^{n_i}$. The storage needs 2 bits for the signs and the binary numbers $\{0, 1\}$, which gives 16× model compression rate.

On the computational side, with sampled mini-batch gradient ∇f_k at the k-th iteration, the classical projected stochastic gradient descent (PSGD) [18, 62]

$$\begin{cases} y^{k+1} = x^k - \gamma_k \nabla f_k(x^k) \\ x^{k+1} = \mathrm{proj}_{\mathcal{Q}}(y^{k+1}), \end{cases} \tag{1.10}$$

performs poorly however, and gets stagnated when updated with a small learning rate γ_k. It is the quantization/projection of weights that "rounds off" small gradient updates and causes the plateau as explained by Li et al. in a recent study [40]. Instead of using the standard gradient step in (1.10), a hybrid gradient update, referred as BinaryConnect [40]

$$y^{k+1} = y^k - \gamma_k \nabla f_k(x^k)$$

was adopted by Courbariaux et al. [19] and has become the workhorse algorithm for training quantized DNN models such as Xnor-Net [59] and TWN [39]. By introducing the augmented Lagrangian of (1.9), more complicated algorithms based on alternating minimization were proposed in [11] and [38]. Despite the succinctness and effectiveness of BinaryConnect, the only convergence analysis of it appeared in [40] under convexity assumption on the loss function. Different quantizers have also been explored [19, 59, 81, 39, 76, 54, 80, 12, 36], all of which maintain a sequence of purely quantized weights during the training.

In Chapter 6, we propose a novel relaxed quantization approach called BinaryRelax, to explore more freely the non-convex landscape of the objective function of the DNNs under the discrete quantization constraint [75]. We relax the set constraint into a continuous regularizer, which leads to a relaxed quantization update. Besides, we set an increasing regularization parameter, driving x^k slowly to the quantized state. When the training error stops decaying at small γ_k, we switch to regular quantization to get genuinely quantized weights as desired. By exploiting the structure of quantization set $\mathcal{Q}$, we prove the convergence of BinaryRelax in the non-convex setting, which naturally covers that of BinaryConnect.

Chapter 2

Computing Residual Diffusivity via Spectral Method

2.1 Effective Diffusivity

2.1.1 Multi-scale analysis

Let $\boldsymbol{v}\left(x,t\right)$ be a velocity flow periodic in $x \in \mathbf{R}^2$ and t, $\nabla \cdot \boldsymbol{v} = 0$ and have mean zero. The advection-diffusion (passive scalar) equation is

$$u_t + \left(\boldsymbol{v} \cdot \nabla\right) u = D_0 \Delta u, \tag{2.1}$$

where $D_0 > 0$ is a constant.

Remark 2.1. *Since $\boldsymbol{v}\left(x,t\right)$ is incompressible and has mean zero in space, there exists a 2×2 skew-symmetric matrix $\mathbf{H} = \left(H_{ij}\left(x,t\right)\right)$ such that $\nabla \cdot \mathbf{H} = \boldsymbol{v}$. In fact, without loss of generality, suppose $\boldsymbol{v}\left(x,t\right)$ is 2π-periodic in spatial and temporal variables, and $\boldsymbol{v} = \left(v_1, v_2\right)$.*

Define

$$H_{pq}\left(x,t\right)=\frac{1}{i}\sum_{k\neq\mathbf{0}}e^{ik\cdot x}\frac{k_p v_{q,k}\left(t\right)-k_q v_{p,k}\left(t\right)}{\left|k\right|^2},$$

for $p,q=1,2,$ *where* $v_p\left(x,t\right)=\sum\limits_{k\in\mathbf{Z}^2}e^{ik\cdot x}v_{p,k}\left(t\right),\ p=1,2.$ *It follows from* $\nabla\cdot\boldsymbol{v}=0$ *that* $\nabla\cdot\mathbf{H}=\boldsymbol{v}$*. Hence equation* (2.1) *can be written in the form*

$$u_t-\left(a_{ij}\left(x,t\right)u_{x_j}\right)_{x_i}=0,$$

where

$$a_{ij}\left(x,t\right)=D_0\delta_{ij}+H_{ij}\left(x,t\right).$$

The matrix $\left(a_{ij}\left(x,t\right)\right)$ *is periodic and uniformly elliptic for* $D_0>0$*.*

In the large-distance and large-time scaling $x\to x/\epsilon$, $t\to t/\epsilon^2$, equation (2.1) becomes

$$u_t^\epsilon\left(x,t\right)+\frac{1}{\epsilon}\left(\boldsymbol{v}\left(\frac{x}{\epsilon},\frac{t}{\epsilon^2}\right)\cdot\nabla\right)u^\epsilon\left(x,t\right)=D_0\Delta u^\epsilon\left(x,t\right).$$

Initial data are independent of ϵ,

$$u^\epsilon\left(x,0\right)=U\left(x\right).$$

Solution can be sought as a multi-scale expansion of the form:

$$u^\epsilon\left(x,t\right)=u^{(0)}\left(x,t;y,\tau\right)+\epsilon u^{(1)}\left(x,t;y,\tau\right)+\epsilon^2u^{(2)}\left(x,t;y,\tau\right)+\cdots,$$

where $y=x/\epsilon$ and $\tau=t/\epsilon^2$.

Let $\boldsymbol{\partial}$ and ∇ denote gradient operator with respect to fast and slow space variables respectively, $\boldsymbol{w}\left(x,t\right)$ be the periodic solution with vanishing average over periodicities to the cell problem

$$\boldsymbol{w}_t\left(x,t\right)+\left(\boldsymbol{v}\left(x,t\right)\cdot\nabla\right)\boldsymbol{w}\left(x,t\right)-D_0\nabla^2\boldsymbol{w}\left(x,t\right)=-\boldsymbol{v}\left(x,t\right). \tag{2.2}$$

It can be calculated directly that $u^{(0)}$ and $u^{(1)}$ are in the form

$$\begin{aligned}
&u^{(0)}\left(x,t;y,\tau\right)=u^{(0)}\left(x,t\right),\\
&u^{(1)}\left(x,t;y,\tau\right)=u^{(1)}\left(x,t\right)+\boldsymbol{w}\left(y,\tau\right)\cdot\nabla u^{(0)}\left(x,t\right).
\end{aligned}$$

Solvability of the equation

$$u^{(2)}_{\tau}+\left(\boldsymbol{v}\cdot\boldsymbol{\partial}\right)u^{(2)}-D_0\boldsymbol{\partial}^2u^{(2)}=-u^{(0)}_t-\left(\boldsymbol{v}\cdot\nabla\right)u^{(1)}+D_0\nabla^2u^{(0)}+2D_0\boldsymbol{\partial}\cdot\nabla u^{(1)}$$

implies the zero average of the right-hand side, so $u^{(0)}$ satisfies the effective equation

$$\begin{aligned}
&u^{(0)}_t\left(x,t\right)=D^E_{ij}\nabla^2u^{(0)}\left(x,t\right),\\
&u^{(0)}\left(x,0\right)=U\left(x\right),
\end{aligned}$$

where the effective diffusivity tensor

$$D^E_{ij}=D_0\left(\delta_{ij}+\langle\boldsymbol{\partial}\boldsymbol{w}_i\cdot\boldsymbol{\partial}\boldsymbol{w}_j\rangle\right),$$

and $\langle\cdot\rangle$ denotes space time average. Given $\left(a_{ij}\left(x,t\right)\right)$ defined in Remark 2.1 being periodic and uniformly elliptic, Theorem 2.1 in Chapter 2 of [3] says that u^{ϵ} converges to $u^{(0)}$ weakly in the L^2 sense as $\epsilon\downarrow 0$.

Explicit upper and lower bounds of D^E are known [29] when $\boldsymbol{v}\left(x,t\right)=\nabla^{\perp}H\left(x\right)$ is time independent. Under appropriate assumptions on $H(x)$, in particular for steady cellular flows (Eqn. (1.2) with $\theta = 0$),

$$D_{ii}^E = \mathcal{O}\left(\sqrt{D_0}\right), \qquad i = 1, 2, \qquad D_0 \downarrow 0.$$

For n-dimensional steady flow, $n \geq 2$, see [83] for the asymptotic limit of $D_0 D^E$ as D_0 tends to zero. Shear layer structure is the typical case when the limit is not zero. Numerical results [5, 51] suggest that if the streamlines of the flow are chaotic,

$$D_{11}^E = \mathcal{O}(1), \qquad D_0 \downarrow 0.$$

We shall recover this result and compute also D_{12}^E with our method.

2.1.2 ODEs from Fourier basis

Let us write $\boldsymbol{v} = (v, \tilde{v})$ and $\boldsymbol{w} = (w, \tilde{w})$ in component form. Consider the first equation in the cell problem (2.2):

$$w_t + (\boldsymbol{v} \cdot \boldsymbol{\partial})\, w - D_0 \partial^2 w = -v. \tag{2.3}$$

Equation (2.3) can be rewritten as an infinite system of ODEs of Fourier modes

$$\frac{dw_k}{dt} + D_0 |k|^2 w_k + i \sum_{j \in \mathbf{Z}^2} \left[(k_1 - j_1)\, v_{\boldsymbol{j}}(t) + (k_2 - j_2)\, \tilde{v}_j(t)\right] w_{k-j} = -v_k(t),$$

where $w = \sum_{k \in \mathbf{Z}^2} w_k(t)\, e^{ik \cdot x}$, $v = \sum_{k \in \mathbf{Z}^2} v_k(t)\, e^{ik \cdot x}$, and $\tilde{v} = \sum_{k \in \mathbf{Z}^2} \tilde{v}_k(t)\, e^{ik \cdot x}$.

Set $\|k\| = \max\{|k_1|, |k_2|\}$. A truncated solution with $(2N+1)^2$ modes

$$w^N(x,t) = \sum_{\|k\| \leq N} w_k^N(t)\, e^{ik \cdot x} \tag{2.4}$$

solves

$$\frac{dw_k^N}{dt} + D_0 |k|^2 w_k^N + i \sum_{\|k-j\| \leq N} \left[(k_1 - j_1)\, v_j(t) + (k_2 - j_2)\, \tilde{v}_j(t)\right] w_{k-j}^N = -v_k(t)\,. \tag{2.5}$$

Thus D_{11}^E is approximated by

$$D_{11,N}^E = D_0 \left(1 + \sum_{\|k\| \leq N} |k|^2 \left\langle w_k^N \overline{w}_k^N \right\rangle \right).$$

2.1.3 Poincaré map

Vectorize $\left\{w_k^N(t)\right\}_{\|k\| \leq N}$ column-wise and denote the vector by $\mathbf{w}(t)$, then

$$\frac{d\mathbf{w}}{dt} = A(t)\,\mathbf{w} + \mathbf{v}(t)\,, \tag{2.6}$$

where A is a $(2N+1)^2 \times (2N+1)^2$ matrix and $\mathbf{v}$ is a $(2N+1)^2 \times 1$ vector determined by (2.5).

Define the Poincaré map $P : \mathbf{R}^{(2N+1)^2} \to \mathbf{R}^{(2N+1)^2}$ as:

$$P(x) = \mathbf{x}(2\pi)\,, \quad x \in \mathbf{R}^{(2N+1)^2},$$

where $\mathbf{x}(t)$ solves

$$\begin{cases} \dfrac{d\mathbf{x}}{dt} = A(t)\,\mathbf{x} + \mathbf{v}(t)\,, \\ \mathbf{x}(0) = x. \end{cases} \tag{2.7}$$

Also define $P_0 : \mathbf{R}^{(2N+1)^2} \to \mathbf{R}^{(2N+1)^2}$,

$$P_0(x) = \mathbf{x}(2\pi)\,, \quad x \in \mathbf{R}^{(2N+1)^2},$$

where $\mathbf{x}(t)$ solves

$$\begin{cases} \dfrac{d\mathbf{x}}{dt} = A(t)\,\mathbf{x}, \\ \mathbf{x}(0) = x. \end{cases} \tag{2.8}$$

Let $e_1, e_2, \ldots, e_{(2N+1)^2}$ be the standard basis of $\mathbf{R}^{(2N+1)^2}$,

$$M = \left[P_0(e_1) \quad P_0(e_2) \quad \ldots \quad P_0\left(e_{(2N+1)^2}\right) \right], \qquad b = P(\mathbf{0})\,,$$

then

$$P(x) = Mx + b, \quad x \in \mathbf{R}^{(2N+1)^2}.$$

Let us impose $\displaystyle\int_{[0,2\pi]^2} w^N(x,t)\,dx = 0$, then $w^N_{\mathbf{0}}(t) = 0$. Hence the initial value of $\mathbf{w}(t)$ is the solution to

$$x = Mx + b$$

with $x_{2N^2+2N+1} = 0$.

2.2 Numerical Method

We shall use $N_t + 1$ equally spaced grid points in the time interval $[0, 2\pi]$.

2.2.1 Assemble matrices in the Poincaré map

Apply the classical Runge-Kutta method (e.g. RK4) to ODE (2.7) with zero initial value for N_t steps,

$$\begin{aligned}\hat{\mathbf{x}}_0 &= \mathbf{0},\\ \hat{\mathbf{x}}_{n+1} &= \mathcal{L}\left(A, \mathbf{v}; \hat{\mathbf{x}}_n, t_n\right),\end{aligned}$$

where $\mathcal{L}$ is the induction operator in RK4. Approximate b by $\hat{b} = \hat{\mathbf{x}}_{N_t}$. Similarly, apply RK4 to ODE (2.8) with initial value e_j,

$$\begin{aligned}\hat{\mathbf{x}}_{j,0} &= e_j,\\ \hat{\mathbf{x}}_{j,n+1} &= \mathcal{L}\left(A, \mathbf{0}; \hat{\mathbf{x}}_{j,n}, t_n\right),\end{aligned}$$

for $j = 1, \ldots, (2N+1)^2$. Approximate $P_0\left(e_j\right)$ by $\hat{m}_j = \hat{\mathbf{x}}_{j,N_t}$ and M by

$$\hat{M} = \begin{bmatrix}\hat{m}_1 & \hat{m}_2 & \ldots & \hat{m}_{(2N+1)^2}\end{bmatrix}.$$

In practice, only half of $\hat{m}_j$'s are computed since $w_{-k}^N = \bar{w}_k^N$. It follows from (2.5) that $\hat{m}_{2N^2+2N+1} = e_{2N^2+2N+1}$. Let $j_1, j_2, \ldots, j_l$ be the vector indices corresponding to Fourier modes indices

$$\left\{k = (k_1, k_2) \,|\, k \neq \mathbf{0}, k_1 \leq k_2\right\},$$

where $l = 2N^2 + 2N$, and

$$\hat{X}_0 = \left[e_{j_1} \quad e_{j_2} \quad \dots \quad e_{j_l}\right],$$

then the following iteration for matrix

$$\hat{X}_{n+1} = \mathcal{L}\left(A, \mathbf{0}; \hat{X}_n, t_n\right)$$

gives

$$\left[\hat{m}_{j_1} \quad \hat{m}_{j_2} \quad \dots \quad \hat{m}_{j_l}\right] = \hat{X}_{N_t}.$$

For $j \notin \{j_1, j_2, \dots, j_l\}$,

$$\hat{m}_j = F\left(\overline{\hat{m}}_{2l+1-j}\right), \tag{2.9}$$

where F is to flip components of a vector. Thus the matrix M in the Poincaré map is assembled. We note that the assembling of matrix M can be implemented in parallel.

2.2.2 Solve ODE and estimate effective diffusivity

The initial data $\hat{x}$ for discretized form of ODE (2.6) is solved from the linear system

$$\hat{x} = \hat{M}\hat{x} + \hat{b}.$$

Again by RK4, the numerical periodic solution to ODE (2.6) is computed as

$$\begin{aligned} \hat{\mathbf{w}}_0 &= \hat{x}, \\ \hat{\mathbf{w}}_{n+1} &= \mathcal{L}\left(A, \mathbf{v}; \hat{\mathbf{w}}_n, t_n\right). \end{aligned}$$

The entire algorithm can be summarized as below.

Algorithm 1 Solving cell problem with Fourier basis.

```
1. Set $l = 2N^2 + 2N$, $\hat{X}_0 = [e_{j_1} \quad e_{j_2} \quad \dots \quad e_{j_l}]$ to be rearrangement of $e_k$ with $\{k = (k_1, k_2)\,|k \neq \mathbf{0}, k_1 \leq k_2\}$.
for n = 0, 1,..., $N_t - 1$ do
    $\hat{X}_{n+1} = \mathcal{L}\left(A, \mathbf{0}; \hat{X}_n, t_n\right)$
end for
Assemble $\hat{M}$ using $[\hat{m}_{j_1} \quad \hat{m}_{j_2} \quad \dots \quad \hat{m}_{j_l}] = \hat{X}_{N_t}$ and (2.9).
2. Set $\hat{\mathbf{x}}_0 = \mathbf{0}$.
for n = 0, 1,..., $N_t - 1$ do
    $\hat{\mathbf{x}}_{n+1} = \mathcal{L}\left(A, \mathbf{v}; \hat{\mathbf{x}}_n, t_n\right)$
end for
$\hat{b} = \hat{\mathbf{x}}_{N_t}$.
3. Solve $\hat{x} = \hat{M}\hat{x} + \hat{b}$.
Set $\hat{\mathbf{w}}_0 = \hat{x}$.
for n = 0, 1,..., $N_t - 1$ do
    $\hat{\mathbf{w}}_{n+1} = \mathcal{L}\left(A, \mathbf{v}; \hat{\mathbf{w}}_n, t_n\right)$
end for
```

For $n = 0, 1, \dots, N_t$, reorder $\hat{\mathbf{w}}_n$ as Fourier modes $\{\hat{w}_{k,n}\}_{\|k\| \leq N}$, then $D^E_{11,N}$ is estimated by

$$\hat{D}^E_{11,N} = D_0 \left(1 + \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{\|\boldsymbol{k}\| \leq N} |k|^2 \left|\hat{w}_{k,n}\right|^2\right).$$

2.3 Numerical Results

In this section, we first present computational results of D^E_{11} by spectral method and Poincaré map for small D_0, recovering the early finding in [5] on time periodic cellular flows. We then perform a parameter dependence study of D_{11}^E on a family of such flows, and discover a non-monotone relationship between D_{11}^E and the amount of chaos in the flows. Similar results hold for D_{12}^E.

2.3.1 Two-dimensional time-dependent flow

As in [5], we consider the time periodic cellular flow with chaotic Lagrangian trajectories:

$$\begin{aligned} v\left(x,t\right) &= \cos\left(x_2\right) + \sin\left(x_2\right)\cos\left(t\right), \\ \tilde{v}\left(x,t\right) &= \cos\left(x_1\right) + \sin\left(x_1\right)\cos\left(t\right). \end{aligned} \tag{2.10}$$

Rewrite

$$\begin{aligned} v\left(x,t\right) &= \frac{1}{2}\left(1 - i\cos t\right)e^{ix_2} + \frac{1}{2}\left(1 + i\cos t\right)e^{-ix_2}, \\ \tilde{v}\left(x,t\right) &= \frac{1}{2}\left(1 - i\cos t\right)e^{ix_1} + \frac{1}{2}\left(1 + i\cos t\right)e^{-ix_1}. \end{aligned}$$

Set $e_1 = (1,0)$, $e_2 = (0,1)$, then

$$\begin{aligned} & v_{\pm e_2}\left(t\right) = \tilde{v}_{\pm e_1}\left(t\right) = \frac{1}{2}\left(1 \mp i\cos t\right), \\ & v_k\left(t\right) = 0, \quad k \neq \pm e_2, \\ & \tilde{v}_k\left(t\right) = 0, \quad k \neq \pm e_1. \end{aligned}$$

Hence (2.5) is reduced to

$$\begin{aligned} \frac{dw_k^N}{dt} + D_0\left|k\right|^2 w_k^N + \frac{1}{2}\big[& k_1\left(i+\cos t\right)w_{k-e_2}^N + k_1\left(i-\cos t\right)w_{k+e_2}^N \\ & + k_2\left(i+\cos t\right)w_{k-e_1}^N + k_2\left(i-\cos t\right)w_{k+e_1}^N\big] + v_k = 0. \end{aligned}$$

Both A and $\mathbf{v}$ are sparse. Estimates of $D_{11,N}^E$ for some varied $D_0/N/N_t$'s are shown in Table

2.1-2.5. $D^E_{11,N}$'s vs. D_0 are plotted in Figure 2.1 which resembles Figure 4 of [5] in the regime $D_0 \leq 0.1$.

$\hat{D}^E_{11,N}$ N N_t	30	35	40	45	50
1000	1.3412	1.3412	1.3412	1.3412	1.3412
1500	1.3412	1.3412	1.3412	1.3412	1.3412

Table 2.1: $\hat{D}^E_{11,N}$ for flow (2.10) with $D_0 = 10^{-2}$.

$\hat{D}^E_{11,N}$ N N_t	40	45	50	55	60
1000	1.3847	1.3787	1.3790	1.3795	1.3778
1500	1.3790	1.3795	1.3778	1.3773	1.3772

Table 2.2: $\hat{D}^E_{11,N}$ for flow (2.10) with $D_0 = 10^{-3}$.

$\hat{D}^E_{11,N}$ N N_t	40	45	50	55	60
1000	1.5448	1.5961	1.5087	1.5035	1.4936
1500	1.5459	1.5971	1.5099	1.5050	1.4949
2000	1.5460	1.5972	1.5101	1.5051	1.4951

Table 2.3: $\hat{D}^E_{11,N}$ for flow (2.10) with $D_0 = 10^{-4}$.

2.3.2 Two-dimensional time-dependent flow with $\theta \in (0,1]$

Let us consider now the one parameter family of time periodic cellular flows

$$\begin{aligned} v\left(x,t\right) &= \cos\left(x_2\right) + \theta \sin\left(x_2\right) \cos\left(t\right), \\ \tilde{v}\left(x,t\right) &= \cos\left(x_1\right) + \theta \sin\left(x_1\right) \cos\left(t\right). \end{aligned} \tag{2.11}$$

$\hat{D}^{E}_{11,N}$ N N_t	55	60	65	70	75	80
2000	1.6774	1.6268	1.7604	1.7528	1.8265	1.6984
2500	1.6793	1.6301	1.7651	1.7558	1.8336	1.7056

Table 2.4: $\hat{D}^{E}_{11,N}$ for flow (2.10) with $D_0 = 10^{-5}$.

$\hat{D}^{E}_{11,N}$ N N_t	55	60	65	70	75	80
2000	1.5676	1.6114	1.7351	1.7074	2.0494	1.5528
2500	1.6270	1.7410	1.8016	1.7882	2.1849	1.6831

Table 2.5: $\hat{D}^{E}_{11,N}$ for flow (2.10) with $D_0 = 10^{-6}$.

As θ increases, the flow trajectories are more and more mixing and chaotic [85]. The Fourier modes for the flow are:

$$\begin{aligned}
&v_{\pm e_2}(t) = \tilde{v}_{\pm e_1}(t) = \frac{1}{2}\left(1 \mp i\theta \cos t\right), \\
&v_k(t) = 0, \quad k \neq \pm e_2, \\
&\tilde{v}_k(t) = 0, \quad k \neq \pm e_1.
\end{aligned}$$

Similarly, (2.5) is reduced to

$$\begin{aligned}
\frac{dw_k^N}{dt} + D_0 |k|^2 w_k^N + \frac{1}{2} \big[& k_1 (i + \theta \cos t)\, w^N_{k-\boldsymbol{e_2}} + k_1 (i - \theta \cos t)\, w^N_{k+e_2} \\
& + k_2 (i + \theta \cos t)\, w^N_{k-e_1} + k_2 (i - \theta \cos t)\, w^N_{k+e_1} \big] + v_k = 0.
\end{aligned}$$

A and $\mathbf{v}$ are still sparse. Estimates $\hat{D}^{E}_{11,N}$ are shown in Table 2.6 and plotted in Figure 2.2. These results are computed according to numerical parameters in Table 2.7. Larger values of N/N_t did not alter the results significantly. We observed a non-monotone dependence of D^E_{11} vs. θ in the small D_0 regime, though the overall trend is that D^E_{11} increases with the amount of chaos in the flows.

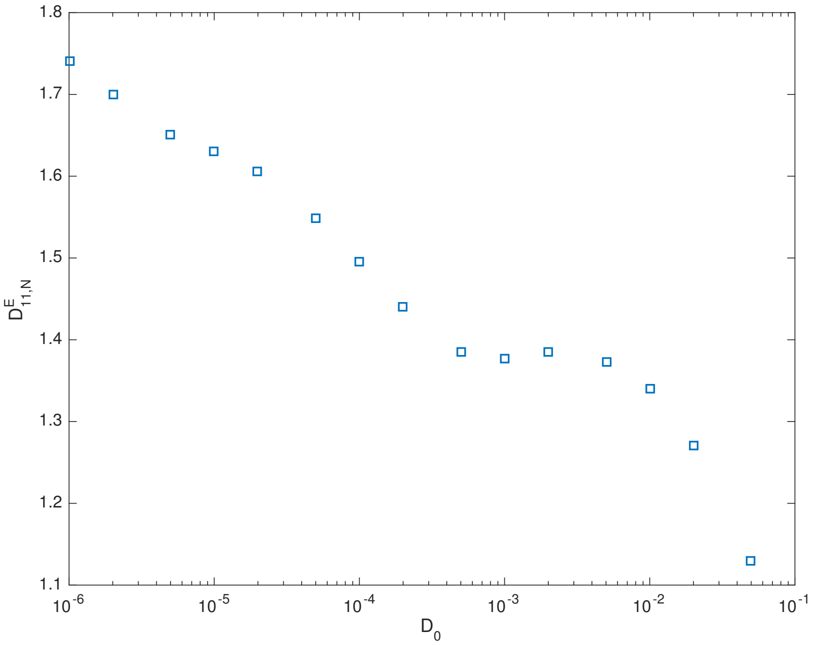


Figure 2.1: Computed $D^E_{11,N}$ vs. D_0 for flow (2.10), resembling Fig. 4 of [5] in the regime $D_0 \leq 0.1$.

2.3.3 Estimates of D^E_{12}

The second component of $\boldsymbol{w} = (w, \tilde{w})$ can be approximated by Fourier modes in a similar way to (2.4),

$$\tilde{w}^N(x,t) = \sum_{\|k\| \leq N} \tilde{w}^N_k(t)\, e^{ik\cdot x}.$$

Hence an estimate of D^E_{12} is

$$D^E_{12,N} = D_0 \sum_{\|k\| \leq N} |k|^2 \left\langle w^N_k \overline{\tilde{w}}^N_k \right\rangle.$$

Computed $D^E_{12,N}$'s vs. D_0 for flow (2.10) are plotted in Figure 2.3. $D^E_{12,N}$'s vs. θ for flow (2.11) with the same numerical parameters in Table 2.7 are plotted in Figure 2.4.

$\hat{D}^E_{11,N}$ D_0 θ	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
0.1	0.1625	0.0733	0.0466	0.0560	0.0954
0.2	0.2079	0.1507	0.1270	0.1054	0.1573
0.3	0.3020	0.3754	0.5615	0.7544	1.0406
0.4	0.3967	0.3921	0.3887	0.3820	0.4040
0.5	0.4315	0.3348	0.3432	0.3063	0.2563
0.6	0.4129	0.2823	0.2425	0.2120	0.2934
0.7	0.3954	0.2177	0.1708	0.1612	0.2156
0.8	0.5740	0.4902	0.5625	0.5708	0.5497
0.9	0.9543	1.1608	1.3140	1.2939	1.1494
1.0	1.3412	1.3778	1.4951	1.6301	1.7410

Table 2.6: Computed $D^E_{11,N}$ vs. θ for the time periodic cellular flow (2.11).

D_0	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
N	50	60	60	60	60
N_t	1500	1500	2000	2500	2500

Table 2.7: Numerical parameters for computing $\hat{D}^E_{11,N}$.

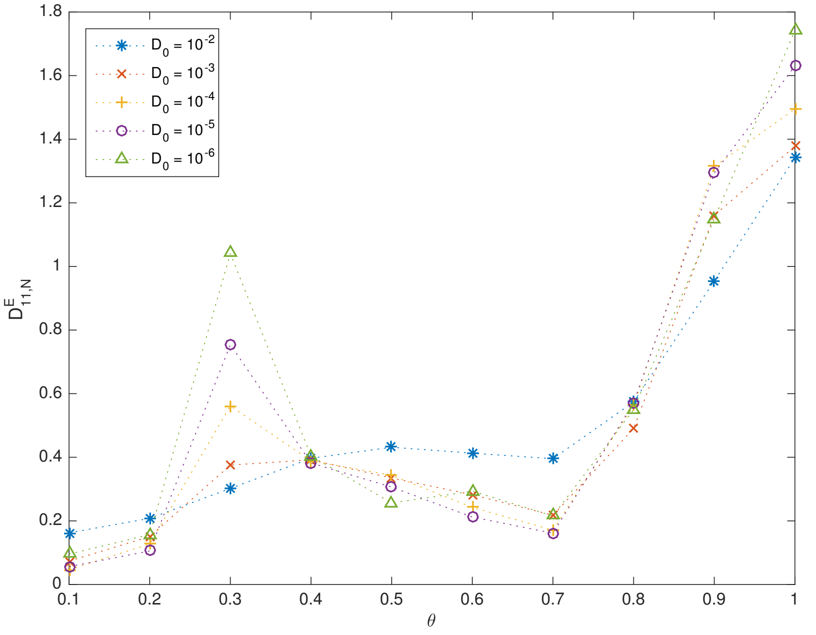


Figure 2.2: $D^E_{11,N}$ vs. θ for flow (2.11) with numerical parameters in Table 2.7.

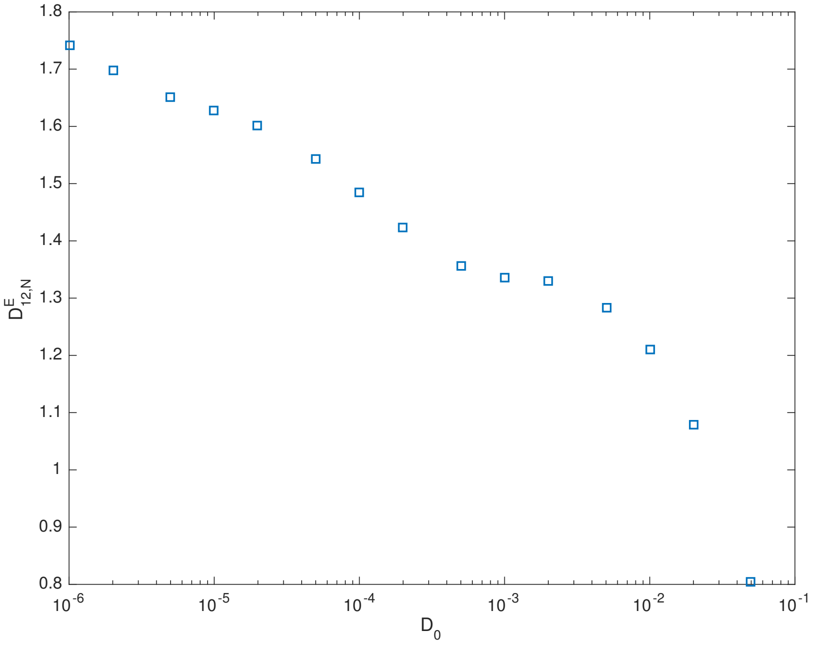


Figure 2.3: Computed $D^E_{12,N}$ vs. D_0 for flow (2.10).

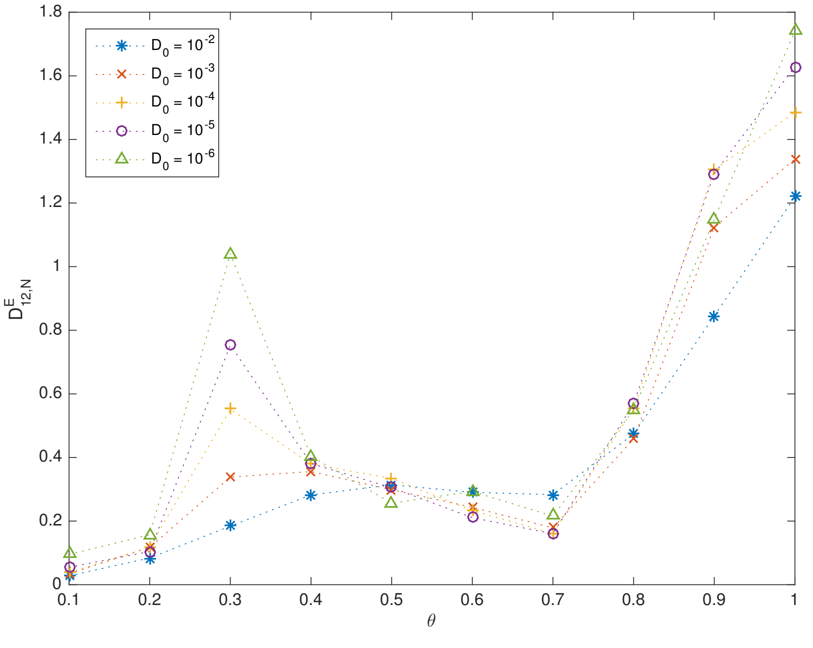


Figure 2.4: $D^E_{12,N}$ vs. θ for the time periodic cellular flow (2.11) with numerical parameters in Table 2.7.

Chapter 3

Residual Diffusivity in Elephant Random Walk Models with Stops

3.1 Perturbed ERWS Model and Moment Analysis

Consider a random walker on a one-dimensional lattice with unit distance between adjacent lattice sites. Denote the position of the walker at time t by X_t. Time is discrete ($t = 0, 1, 2, \dots$) and the walker starts at the origin, $X_0 = 0$. At each time step, $t \to t+1$,

$$X_{t+1} = X_t + \sigma_{t+1},$$

where $\sigma_{t+1} \in \{-1, 0, 1\}$ is a random number depending on $\{\sigma_t\} = (\sigma_1, \dots, \sigma_t)$ as follows. Let $p, q, r, \epsilon \in (0,1)$ and $p+q+r = 1$. The process is started at time $t = 0$ by allowing the walker to move to the right with probability s and to the left with probability $1-s$, $s \in (0,1)$. For $t \geq 1$, a random previous time $k \in \{1, \dots, t\}$ is chosen with uniform probability.

(i) If $\sigma_k = \pm 1$,

$$P\left(\sigma_{t+1} = \sigma_k\right) = p,\; P\left(\sigma_{t+1} = -\sigma_k\right) = q,$$
$$P\left(\sigma_{t+1} = 0\right) = r.$$

(ii) If $\sigma_k = 0$,

$$P\left(\sigma_{t+1} = 1\right) = P\left(\sigma_{t+1} = -1\right) = \epsilon/2,$$
$$P\left(\sigma_{t+1} = 0\right) = 1 - \epsilon.$$

When $\epsilon = 0$, the model is reduced to the elephant random walk model with stops (ERWS) [35].

3.1.1 First moment $\langle X_t \rangle$

At $t = 0$, it follows from the initial condition of the model for $\sigma = \pm 1$ that

$$P\left(\sigma_1 = \sigma\right) = \frac{1}{2}\left[1 + \left(2s - 1\right)\sigma\right].$$

Let $\gamma = p - q$, for $t \geq 1$, it follows from the probabilistic structure of the model and $\sigma_k \in \{1, -1, 0\}$ that

$$\begin{aligned}
P\left(\sigma_{t+1} = 1 \middle| \left\{\sigma_t\right\}\right) &= \frac{1}{t}\sum_{k=1}^{t}\left[\sigma_k^2\left(1+\sigma_k\right)\frac{p}{2} + \sigma_k^2\left(1-\sigma_k\right)\frac{q}{2} + \left(1-\sigma_k^2\right)\frac{\epsilon}{2}\right] \\
&= \frac{1}{t}\sum_{k=1}^{t}\left[\sigma_k^2\frac{1-r}{2} + \sigma_k\frac{\gamma}{2} + \left(1-\sigma_k^2\right)\frac{\epsilon}{2}\right] \\
&= \frac{1}{2t}\sum_{k=1}^{t}\left[\sigma_k^2\left(1-\epsilon-r\right) + \sigma_k\gamma\right] + \frac{\epsilon}{2},
\end{aligned}$$

$$
\begin{aligned}
P\left(\sigma_{t+1}=-1|\left\{\sigma_t\right\}\right) &= \frac{1}{t}\sum_{k=1}^{t}\left[\sigma_k^2\left(1-\sigma_k\right)\frac{p}{2}+\sigma_k^2\left(1+\sigma_k\right)\frac{q}{2}+\left(1-\sigma_k^2\right)\frac{\epsilon}{2}\right] \\
&= \frac{1}{t}\sum_{k=1}^{t}\left[\sigma_k^2\frac{1-r}{2}-\sigma_k\frac{\gamma}{2}+\left(1-\sigma_k^2\right)\frac{\epsilon}{2}\right] \\
&= \frac{1}{2t}\sum_{k=1}^{t}\left[\sigma_k^2\left(1-\epsilon-r\right)-\sigma_k\gamma\right]+\frac{\epsilon}{2},
\end{aligned}
$$

$$
\begin{aligned}
P\left(\sigma_{t+1}=0|\left\{\sigma_t\right\}\right) &= \frac{1}{t}\sum_{k=1}^{t}\left[\sigma_k^2 r+\left(1-\sigma_k^2\right)\left(1-\epsilon\right)\right] \\
&= \frac{1}{t}\sum_{k=1}^{t}\left[-\sigma_k^2\left(1-\epsilon-r\right)\right]+1-\epsilon \\
&= \frac{1}{2t}\sum_{k=1}^{t}\left[-2\sigma_k^2\left(1-\epsilon-r\right)\right]+1-\epsilon.
\end{aligned}
$$

Therefore, for $\sigma=\pm1,0$,

$$
\begin{aligned}
P\left(\sigma_{t+1}=\sigma|\left\{\sigma_t\right\}\right) =& \frac{1}{2t}\sum_{k=1}^{t}\left[\sigma_k^2\left(3\sigma^2-2\right)\left(1-\epsilon-r\right)+\sigma\sigma_k\gamma\right] \\
&+\frac{\sigma^2}{2}\epsilon+\left(1-\sigma^2\right)\left(1-\epsilon\right).
\end{aligned}
$$

The conditional mean value of σ_{t+1} for $t \geq 1$ is

$$\begin{aligned}\langle \sigma_{t+1} | \{\sigma_t\} \rangle &= \sum_{\sigma=\pm 1,0} \sigma P\left(\sigma_{t+1} = \sigma | \{\sigma_t\}\right) \\ &= \sum_{\sigma=\pm 1} \sigma \left\{ \frac{1}{2t} \sum_{k=1}^{t} \left[\sigma_k^2 \left(3\sigma^2 - 2\right)(1 - \epsilon - r) + \sigma\sigma_k\gamma\right] \right. \\ &\quad \left. + \frac{\sigma^2}{2}\epsilon + \left(1 - \sigma^2\right)(1 - \epsilon) \right\} \\ &= \sum_{\sigma=\pm 1} \sigma \left\{ \frac{1}{2t} \sum_{k=1}^{t} \left[\sigma_k^2 (1 - \epsilon - r) + \sigma\sigma_k\gamma\right] + \frac{\epsilon}{2} \right\} \\ &= \sum_{\sigma=\pm 1} \frac{1}{2t} \sum_{k=1}^{t} \sigma^2 \sigma_k \gamma,\end{aligned}$$

hence,

$$\langle \sigma_{t+1} | \{\sigma_t\} \rangle = \frac{\gamma}{t} X_t. \tag{3.1}$$

It follows that

$$\langle \sigma_{t+1} \rangle = \frac{\gamma}{t} \langle X_t \rangle,$$

therefore

$$\langle X_{t+1} \rangle = \left(1 + \frac{\gamma}{t}\right) \langle X_t \rangle.$$

By the initial condition $\langle X_1 \rangle = 2s - 1$,

$$\langle X_t \rangle = (2s - 1) \frac{\Gamma(t + \gamma)}{\Gamma(1 + \gamma)\Gamma(t)}.$$

Since $\lim\limits_{t\to\infty} \dfrac{\Gamma(t+\alpha)}{\Gamma(t)\, t^{\alpha}} = 1$, $\forall\alpha$,

$$\langle X_t \rangle \sim \frac{2s-1}{\Gamma(1+\gamma)} t^{\gamma}, \quad t \to \infty.$$

From this point on, *we shall take $s = 1/2$, and so $\langle X_t \rangle = 0$, the mean square displacement agrees with the second moment.*

3.1.2 Second moment $\langle X_t^2 \rangle$

The conditional mean value of σ_{t+1}^2 for $t \geq 1$ is

$$\begin{aligned}
\langle \sigma_{t+1}^2 \big| \{\sigma_t\} \rangle &= \sum_{\sigma=\pm1,0} \sigma^2 P\left(\sigma_{t+1} = \sigma \big| \{\sigma_t\}\right) \\
&= \sum_{\sigma=\pm1} \sigma^2 \left\{ \frac{1}{2t} \sum_{k=1}^{t} \left[\sigma_k^2 \left(3\sigma^2 - 2\right)(1-\epsilon-r) + \sigma\sigma_k\gamma \right] \right. \\
&\quad \left. + \frac{\sigma^2}{2}\epsilon + \left(1-\sigma^2\right)(1-\epsilon) \right\} \\
&= \sum_{\sigma=\pm1} \left\{ \frac{1}{2t} \sum_{k=1}^{t} \left[\sigma_k^2 (1-\epsilon-r) + \sigma\sigma_k\gamma \right] + \frac{\epsilon}{2} \right\} \\
&= \sum_{\sigma=\pm1} \left\{ \frac{1}{2t} \sum_{k=1}^{t} \sigma_k^2 (1-\epsilon-r) + \frac{\epsilon}{2} \right\} \\
&= \frac{1-\epsilon-r}{t} \sum_{k=1}^{t} \sigma_k^2 + \epsilon.
\end{aligned}$$

It follows that

$$\begin{aligned}
\langle \sigma_{t+1}^2 \big| \{\sigma_t\} \rangle &= \frac{1-\epsilon-r}{t} \sum_{k=1}^{t-1} \sigma_k^2 + \frac{1-\epsilon-r}{t} \sigma_t^2 + \epsilon \\
&= \frac{t-1}{t} \left(\frac{1-\epsilon-r}{t-1} \sum_{k=1}^{t-1} \sigma_k^2 + \epsilon \right) - \frac{t-1}{t}\epsilon + \frac{1-\epsilon-r}{t} \sigma_t^2 + \epsilon \\
&= \frac{t-1}{t} \langle \sigma_t^2 \big| \{\sigma_{t-1}\} \rangle + \frac{1-\epsilon-r}{t} \sigma_t^2 + \frac{\epsilon}{t},
\end{aligned}$$

so

$$\begin{aligned} \langle \sigma_1^2 \rangle &= 1, \\ \langle \sigma_{t+1}^2 \rangle &= \left(1 - \frac{\epsilon + r}{t}\right) \langle \sigma_t^2 \rangle + \frac{\epsilon}{t}. \end{aligned} \tag{3.2}$$

Since

$$\langle X_{t+1}^2 \big| \{\sigma_t\} \rangle = X_t^2 + 2X_t \langle \sigma_{t+1} \big| \{\sigma_t\} \rangle + \langle \sigma_{t+1}^2 \big| \{\sigma_t\} \rangle,$$

by (3.1),

$$\langle X_{t+1}^2 \rangle = \left(1 + \frac{2\gamma}{t}\right) \langle X_t^2 \rangle + \langle \sigma_{t+1}^2 \rangle. \tag{3.3}$$

To motivate the solution we shall present, let us consider the ODE analogue of the difference equations (3.2) and (3.3).

$$\begin{cases} x' = -\dfrac{\epsilon + r}{t} x + \dfrac{\epsilon}{t}, \\ y' = x + \dfrac{2\gamma}{t} y. \end{cases} \tag{3.4}$$

The solution to (3.4) is

$$\begin{cases} x\left(t\right) = \dfrac{C}{t^{\epsilon + r}} + \dfrac{\epsilon}{\epsilon + r}, \\ y\left(t\right) = \dfrac{\epsilon}{\left(1 - 2\gamma\right)\left(\epsilon + r\right)} t + \dfrac{C}{1 - \epsilon - r - 2\gamma} t^{1 - \epsilon - r} + D t^{2\gamma}, \end{cases}$$

if $\gamma \neq \frac{1}{2}$, and

$$\begin{cases} x\left(t\right) = \dfrac{C}{t^{\epsilon+r}} + \dfrac{\epsilon}{\epsilon+r}, \\ y\left(t\right) = \dfrac{\epsilon}{\epsilon+r} t\ln t - \dfrac{C}{\epsilon+r} t^{1-\epsilon-r} + Dt, \end{cases}$$

if $\gamma = \frac{1}{2}$, where C and D are constants.

Proposition 3.1. *The solution to* (3.2) *is*

$$\langle \sigma_t^2 \rangle = C\frac{\Gamma\left(t-\epsilon-r\right)}{\Gamma\left(t\right)} + \frac{\epsilon}{\epsilon+r}, \tag{3.5}$$

where

$$C = \frac{r}{\left(\epsilon+r\right)\Gamma\left(1-\epsilon-r\right)}.$$

Proof. Clearly,

$$\begin{aligned} \frac{\epsilon}{\epsilon+r} &= \left(1 - \frac{\epsilon+r}{t}\right)\frac{\epsilon}{\epsilon+r} + \frac{\epsilon}{t}, \\ \frac{\Gamma\left(t+1-\epsilon-r\right)}{\Gamma\left(t+1\right)} &= \left(1-\frac{\epsilon+r}{t}\right)\frac{\Gamma\left(t-\epsilon-r\right)}{\Gamma\left(t\right)}, \end{aligned}$$

so a general solution to the recurrence equation in (3.2) is given by (3.5). The initial condition $\langle \sigma_1^2 \rangle = 1$ implies $C = \dfrac{r}{\left(\epsilon+r\right)\Gamma\left(1-\epsilon-r\right)}$. □

It follows from Proposition 3.1 and (3.3) that

$$\begin{aligned} &\langle X_1^2 \rangle = 1, \\ &\langle X_{t+1}^2 \rangle = \left(1+\frac{2\gamma}{t}\right)\langle X_t^2 \rangle + C\frac{\Gamma\left(t+1-\epsilon-r\right)}{\Gamma\left(t+1\right)} + \frac{\epsilon}{\epsilon+r}. \end{aligned} \tag{3.6}$$

Theorem 3.1. *Let* $C = \dfrac{r}{\left(\epsilon+r\right)\Gamma\left(1-\epsilon-r\right)}$.

(1) If $\gamma \neq \frac{1}{2}$, *the solution to* (3.6) *is*

$$\langle X_t^2 \rangle = \frac{\epsilon}{(1-2\gamma)(\epsilon+r)} t + \frac{C}{1-\epsilon-r-2\gamma} \frac{\Gamma(t+1-\epsilon-r)}{\Gamma(t)} + D\frac{\Gamma(t+2\gamma)}{\Gamma(t)}, \tag{3.7}$$

where

$$D = -\frac{1}{\Gamma(2\gamma)} \left[\frac{\epsilon}{(\epsilon+r)(1-2\gamma)} + \frac{r}{(\epsilon+r)(1-\epsilon-r-2\gamma)} \right].$$

(2) If $\gamma = \frac{1}{2}$, *the solution to* (3.6) *is*

$$\langle X_t^2 \rangle = \frac{\epsilon}{\epsilon+r} t \sum_{k=1}^{t} \frac{1}{k} - \frac{C}{\epsilon+r} \frac{\Gamma(t+1-\epsilon-r)}{\Gamma(t)} + Dt, \tag{3.8}$$

where

$$D = \frac{\epsilon}{(\epsilon+r)^2} - 1.$$

Proof. Motivated by the ODE solution, we check the formula of the solution to (3.6).

If $\gamma \neq \frac{1}{2}$, by the identity $\Gamma(x+1) = x\Gamma(x)$,

$$\frac{\epsilon}{(1-2\gamma)(\epsilon+r)}(t+1) = \left(1+\frac{2\gamma}{t}\right) \frac{\epsilon}{(1-2\gamma)(\epsilon+r)} t + \frac{\epsilon}{\epsilon+r}, \tag{3.9}$$

$$\frac{C}{1-\epsilon-r-2\gamma} \frac{\Gamma(t+2-\epsilon-r)}{\Gamma(t+1)} = \left(1+\frac{2\gamma}{t}\right) \frac{C}{1-\epsilon-r-2\gamma} \frac{\Gamma(t+1-\epsilon-r)}{\Gamma(t)} + C\frac{\Gamma(t+1-\epsilon-r)}{\Gamma(t+1)}, \tag{3.10}$$

$$\frac{\Gamma(t+1+2\gamma)}{\Gamma(t+1)} = \left(1+\frac{2\gamma}{t}\right) \frac{\Gamma(t+2\gamma)}{\Gamma(t)}. \tag{3.11}$$

Hence a general solution to the recurrence equation in (3.6) is given by (3.7) for some constant

D. Then $\langle X_1^2 \rangle = 1$ and $C = \dfrac{r}{(\epsilon{+}r)\,\Gamma\,(1{-}\epsilon{-}r)}$ imply

$$\frac{\epsilon}{(1{-}2\gamma)\,(\epsilon{+}r)} + \frac{r\Gamma\,(2{-}\epsilon{-}r)}{(\epsilon{+}r)\,(1{-}\epsilon{-}r{-}2\gamma)\,\Gamma\,(1{-}\epsilon{-}r)} + D\Gamma\,(1{+}2\gamma) = 1,$$

so

$$D = -\frac{1}{\Gamma\,(2\gamma)}\left[\frac{\epsilon}{(\epsilon+r)\,(1-2\gamma)} + \frac{r}{(\epsilon+r)\,(1-\epsilon-r-2\gamma)}\right].$$

If $\gamma = \dfrac{1}{2}$, (3.10) and (3.11) still hold,

$$\begin{aligned} -\frac{C}{\epsilon+r}\frac{\Gamma\,(t+2-\epsilon-r)}{\Gamma\,(t+1)} &= \left(1+\frac{1}{t}\right)\left(-\frac{C}{\epsilon+r}\frac{\Gamma\,(t+1-\epsilon-r)}{\Gamma\,(t)}\right) \\ &\quad + C\frac{\Gamma\,(t+1-\epsilon-r)}{\Gamma\,(t+1)}, \\ t+1 &= \left(1+\frac{1}{t}\right)t. \end{aligned}$$

For the recurrence relation

$$a_{t+1} = \left(1+\frac{1}{t}\right)a_t + \frac{\epsilon}{\epsilon+r},$$

suppose $a_t = tb_t$, then

$$b_{t+1} = b_t + \frac{\epsilon}{\epsilon+r}\frac{1}{t+1},$$

so for $t \geq 1$,

$$b_t = b_0 + \frac{\epsilon}{\epsilon+r}\sum_{k=1}^{t}\frac{1}{k}.$$

Set $b_0 = 0$, then

$$a_t = \frac{\epsilon}{\epsilon + r} t \sum_{k=1}^{t} \frac{1}{k}.$$

Hence a general solution to the recurrence equation in (3.6) in this case is (3.8). Similarly, the initial condition gives

$$D = \frac{\epsilon}{(\epsilon + r)^2} - 1.$$

□

The corollary below follows from (3.7) and (3.8).

Corollary 3.1. *Let C and D be the same as in Theorem 3.1.*

(1) If $\gamma \neq \frac{1}{2}$,

$$\langle X_t^2 \rangle \sim \frac{\epsilon}{(1 - 2\gamma)(\epsilon + r)} t + \frac{C}{1 - \epsilon - r - 2\gamma} t^{1-\epsilon-r} + D t^{2\gamma}, \quad t \to \infty.$$

(2) If $\gamma = \frac{1}{2}$,

$$\langle X_t^2 \rangle \sim \frac{\epsilon}{\epsilon + r} t \ln t - \frac{C}{\epsilon + r} t^{1-\epsilon-r} + D t, \quad t \to \infty.$$

3.2 Residual Diffusivity

The occurrence of residual diffusivity relies on the choice of γ as a function of ϵ. To this end, we show three cases: 1) case 1 only recovers the un-perturbed diffusivity, 2) case 2 reveals the residual diffusivity exceeding the un-perturbed diffusivity in the limit of $\epsilon \downarrow 0$, 3) case 3 results in residual super-diffusivity. The cases 2 and 3 are illustrated in Figure 3.1. As $\epsilon \to 0$, the parameter region of the residual diffusion shrinks towards $\gamma = \frac{1}{2}$ while the enhanced diffusivity remains strictly above the un-perturbed diffusivity.

3.2.1 Regular diffusivity

Let $\gamma = \frac{1-\epsilon}{2}$, then $D = 0$ and

$$\langle X_t^2 \rangle = \frac{1}{(\epsilon + r)} t - \frac{1}{(\epsilon + r)\,\Gamma(1-\epsilon-r)} \frac{\Gamma(t+1-\epsilon-r)}{\Gamma(t)},$$

so

$$\langle X_t^2 \rangle \sim \frac{1}{(\epsilon + r)} t - \frac{1}{(\epsilon + r)\,\Gamma(1-\epsilon-r)} t^{1-\epsilon-r}, \quad t \to \infty,$$

and diffusivity equals $\frac{1}{\epsilon + r}$.

For fixed $r \in \left(0, \frac{1}{2}\right)$, let $\epsilon \in (0,1)$, then

$$p = \frac{3-\epsilon-2r}{4}, \quad q = \frac{1+\epsilon-2r}{4}.$$

Recall the second moment formula of [35] (equation (18)),

$$\begin{aligned} \langle X_t^2 \rangle &= \frac{1}{(2\gamma+r-1)\,\Gamma(t)} \left(\frac{\Gamma(t+2\gamma)}{\Gamma(2\gamma)} - \frac{\Gamma(1+t-r)}{\Gamma(1-r)} \right) \\ &\sim \frac{1}{(2\gamma+r-1)} \left(\frac{t^{2\gamma}}{\Gamma(2\gamma)} - \frac{t^{1-r}}{\Gamma(1-r)} \right), \end{aligned} \tag{3.12}$$

which is diffusive at $\gamma = 1/2$ with diffusivity $1/r$.

We see that for $\gamma = (1-\epsilon)/2$, $\epsilon \in (0,1)$ and the above (p,q), the diffusivity of the perturbed ERW problem $1/(\epsilon + r)$ approaches $1/r$, the diffusivity of the un-perturbed model as $\epsilon \downarrow 0$. Hence no residual diffusivity exists.

3.2.2 Residual diffusivity

Let $\gamma = \dfrac{1-\epsilon r}{2}$, then

$$\begin{aligned}\langle X_t^2 \rangle =& \frac{1}{r(\epsilon+r)} t - \frac{r\Gamma(t+1-\epsilon-r)}{(\epsilon+r)(\epsilon+r-\epsilon r)\Gamma(1-\epsilon-r)\Gamma(t)} \\ &- \frac{1}{\Gamma(1-\epsilon r)}\left[\frac{1}{r(\epsilon+r)} - \frac{r}{(\epsilon+r)(\epsilon+r-\epsilon r)}\right]\frac{\Gamma(t+1-\epsilon r)}{\Gamma(t)},\end{aligned}$$

and

$$\begin{aligned}\langle X_t^2 \rangle \sim& \frac{1}{r(\epsilon+r)} t - \frac{r}{(\epsilon+r)(\epsilon+r-\epsilon r)\Gamma(1-\epsilon-r)} t^{1-\epsilon-r} \\ &- \frac{1}{\Gamma(1-\epsilon r)}\left[\frac{1}{r(\epsilon+r)} - \frac{r}{(\epsilon+r)(\epsilon+r-\epsilon r)}\right] t^{1-\epsilon r}, \quad t \to \infty.\end{aligned}$$

Hence

$$\lim_{t\to\infty} \frac{\langle X_t^2 \rangle}{t} = \frac{1}{r(\epsilon+r)}.$$

The diffusivity $\dfrac{1}{r(\epsilon+r)}$ can be much larger than $\dfrac{1}{r}$ in the un-perturbed model. In particular, given any $\delta > 0$, let $r_0 = \min\left\{\dfrac{1}{3}, \dfrac{1}{\delta}\right\}$, then for $r \in (0, r_0)$, $\epsilon \in \left(0, \dfrac{1}{6}\right)$,

$$\frac{1}{r(\epsilon+r)} - \frac{1}{r} = \frac{1}{r}\left(\frac{1}{\epsilon+r} - 1\right) > \frac{1}{r_0}\left(\frac{1}{\frac{1}{6}+r_0} - 1\right) \geq \delta\left(\frac{1}{\frac{1}{6}+\frac{1}{3}} - 1\right) = \delta.$$

The **new diffusive region with residual diffusivity** is the **wedge to the left of** $\gamma = 1/2$

covered by the dashed lines in Figure 3.1.

3.2.3 Residual super-diffusivity

If $\gamma = \frac{1+\epsilon r}{2}$, then

$$\begin{aligned}\langle X_t^2 \rangle =& -\frac{1}{r(\epsilon+r)}t - \frac{r\Gamma(t+1-\epsilon-r)}{(\epsilon+r)(\epsilon+r+\epsilon r)\Gamma(1-\epsilon-r)\Gamma(t)} \\ &+ \frac{1}{\Gamma(1+\epsilon r)}\left[\frac{1}{r(\epsilon+r)} + \frac{r}{(\epsilon+r)(\epsilon+r+\epsilon r)}\right]\frac{\Gamma(t+1+\epsilon r)}{\Gamma(t)},\end{aligned}$$

and

$$\begin{aligned}\langle X_t^2 \rangle \sim& -\frac{1}{r(\epsilon+r)}t - \frac{r}{(\epsilon+r)(\epsilon+r+\epsilon r)\Gamma(1-\epsilon-r)}t^{1-\epsilon-r} \\ &+ \frac{1}{\Gamma(1+\epsilon r)}\left[\frac{1}{r(\epsilon+r)} + \frac{r}{(\epsilon+r)(\epsilon+r+\epsilon r)}\right]t^{1+\epsilon r}, \quad t\to\infty.\end{aligned}$$

Thus at any $\epsilon > 0$, super-diffusion arises and

$$\lim_{t\to\infty}\frac{\langle X_t^2\rangle}{t^{1+\epsilon r}} = \frac{1}{\Gamma(1+\epsilon r)}\left[\frac{1}{r(\epsilon+r)} + \frac{r}{(\epsilon+r)(\epsilon+r+\epsilon r)}\right].$$

As $\epsilon \downarrow 0$, the super-diffusivity tends to $r^{-2} + r^{-1} > r^{-1}$ the limiting super-diffusivity of the un-perturbed model as seen from (3.12). The residual super-diffusive region is the wedge covered by lines to the right of $\gamma > 1/2$ in Figure 3.1.

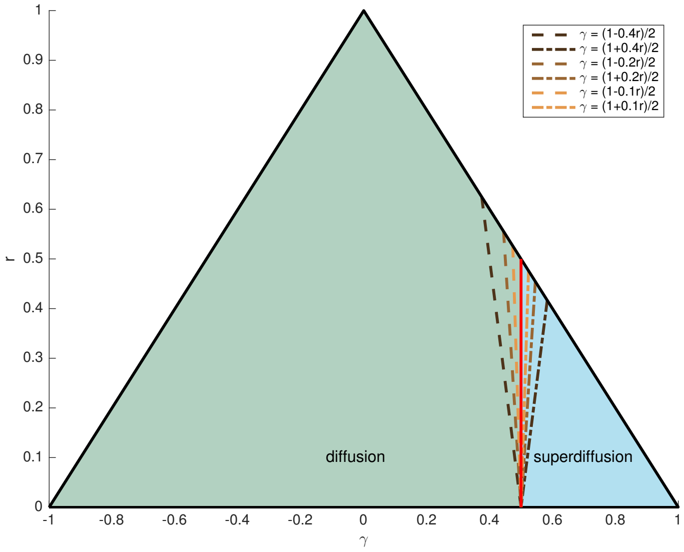


Figure 3.1: Regions of residual diffusivity and residual super-diffusivity covered by the dashed lines at $\epsilon = 0.4, 0.2, 0.1$.

3.3 2D Perturbed ERWS Model

In this section, we generalize our model to the two dimensional square lattice. Let $\mathbf{i}, \mathbf{j}$ be the standard basis in 2D. Denote the position of the walker at time t by $\boldsymbol{X}_t$,

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \boldsymbol{\sigma}_{t+1},$$

where $\boldsymbol{\sigma}_{t+1} \in \{\mathbf{i}, \mathbf{j}, -\mathbf{i}, -\mathbf{j}\}$. Let $s_i \in (0,1)$, $i = 1, \ldots, 4$ and the process is started by allowing the walker to move to the right, upward, to the left, downward with probability $s_1, \ldots, s_4$. Let $p, q, q', r, \epsilon \in (0,1)$ and $p + q + q' + r = 1$,

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

For $t \geq 1$, a random $k \in \{1, \ldots, t\}$ is chosen with uniform probability.

(i) If $|\boldsymbol{\sigma}_k| = 1$,

$$\begin{aligned}
&P\left(\boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma}_k\right) = p, \\
&P\left(\boldsymbol{\sigma}_{t+1} = -\boldsymbol{\sigma}_k\right) = q, \\
&P\left(\boldsymbol{\sigma}_{t+1} = A\boldsymbol{\sigma}_k\right) = p', \\
&P\left(\boldsymbol{\sigma}_{t+1} = A^{-1}\boldsymbol{\sigma}_k\right) = q', \\
&P\left(\boldsymbol{\sigma}_{t+1} = \mathbf{0}\right) = r.
\end{aligned}$$

(ii) If $|\boldsymbol{\sigma}_k| = 0$,

$$\begin{aligned}
&P\left(\boldsymbol{\sigma}_{t+1} = \mathbf{i}\right) = P\left(\boldsymbol{\sigma}_{t+1} = \mathbf{j}\right) = P\left(\boldsymbol{\sigma}_{t+1} = -\mathbf{i}\right) = P\left(\boldsymbol{\sigma}_{t+1} = -\mathbf{j}\right) = \epsilon/4, \\
&P\left(\boldsymbol{\sigma}_{t+1} = \mathbf{0}\right) = 1 - \epsilon.
\end{aligned}$$

Let $\gamma = p - q$, $\gamma' = p' - q'$, then for $t \geq 1$,

$$\begin{aligned}
P\left(\boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma} | \{\boldsymbol{\sigma}_t\}\right) =& \frac{1}{t} \sum_{k=1}^{t} \Big[\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma} + 1\right) \frac{p}{2} + \boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma} - 1\right) \frac{q}{2} \\
&+ \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\left(\boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma} + 1\right) \frac{p'}{2} + \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\left(\boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma} - 1\right) \frac{q'}{2} \\
&+ \left(1 - |\boldsymbol{\sigma}_k|^2\right) \frac{\epsilon}{4}\Big] \\
=& \frac{1}{2t} \sum_{k=1}^{t} \Big[\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma} \gamma + \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma} \gamma' + \left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\right)^2 (p + q) \\
&+ \left(\boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\right)^2 (p' + q') - \frac{1}{2} |\boldsymbol{\sigma}_k|^2 \epsilon\Big] + \frac{\epsilon}{4},
\end{aligned}$$

for $|\boldsymbol{\sigma}| = 1$, and

$$\begin{aligned}
P\left(\boldsymbol{\sigma}_{t+1} = \mathbf{0} \middle| \left\{\boldsymbol{\sigma}_t\right\}\right) &= \frac{1}{t}\sum_{k=1}^{t}\left[\left|\boldsymbol{\sigma}_k\right|^2 r + \left(1 - \left|\boldsymbol{\sigma}_k\right|^2\right)(1-\epsilon)\right] \\
&= \frac{1}{t}\sum_{k=1}^{t}\left|\boldsymbol{\sigma}_k\right|^2 (r + \epsilon - 1) + 1 - \epsilon.
\end{aligned}$$

The conditional mean of $\boldsymbol{\sigma}_{t+1}$ for $t \geq 1$ is

$$\begin{aligned}
\left\langle \boldsymbol{\sigma}_{t+1} \middle| \left\{\boldsymbol{\sigma}_t\right\}\right\rangle &= \sum_{|\boldsymbol{\sigma}|=1} P\left(\boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma} \middle| \left\{\boldsymbol{\sigma}_t\right\}\right)\boldsymbol{\sigma} \\
&= \frac{1}{2t}\sum_{k=1}^{t}\sum_{|\boldsymbol{\sigma}|=1}\Big[\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\gamma + \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\gamma' + \left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\right)^2 (p+q) \\
&\quad + \left(\boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\right)^2 (p' + q') - \frac{1}{2}\left|\boldsymbol{\sigma}_k\right|^2 \epsilon\Big]\boldsymbol{\sigma} \\
&= \frac{1}{2t}\sum_{k=1}^{t}\sum_{|\boldsymbol{\sigma}|=1}\left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\gamma + \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\gamma'\right)\boldsymbol{\sigma} \\
&= \frac{1}{2t}\sum_{k=1}^{t}\sum_{|\boldsymbol{\sigma}|=1}\left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\gamma + A\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\gamma'\right)\boldsymbol{\sigma} \\
&= \frac{1}{2t}\sum_{k=1}^{t} 2\left(\gamma\boldsymbol{\sigma}_k + \gamma' A\boldsymbol{\sigma}_k\right) \\
&= \frac{1}{t}\left(\gamma + \gamma' A\right)\boldsymbol{X}_t.
\end{aligned}$$

Here the symmetry of $\pm\mathbf{i}$, $\pm\mathbf{j}$ is used. Thus,

$$\left\langle \boldsymbol{X}_{t+1}\right\rangle = \left(1 + \frac{\gamma}{t} + \frac{\gamma'}{t}A\right)\left\langle \boldsymbol{X}_t\right\rangle.$$

The conditional mean of $|\boldsymbol{\sigma}_{t+1}|^2$ for $t \geq 1$ is

$$\begin{aligned}\langle\, |\boldsymbol{\sigma}_{t+1}|^2 \big| \, \{\boldsymbol{\sigma}_t\}\rangle =& \sum_{|\boldsymbol{\sigma}|=1} P\left(\boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma} \right| \{\boldsymbol{\sigma}_t\}) \, |\boldsymbol{\sigma}|^2 \\ =& \frac{1}{2t} \sum_{k=1}^{t} \sum_{|\boldsymbol{\sigma}|=1} \Big[\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma} \gamma + \boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma} \gamma' + \left(\boldsymbol{\sigma}_k \cdot \boldsymbol{\sigma}\right)^2 (p+q) \\ & + \left(\boldsymbol{\sigma}_k \cdot A\boldsymbol{\sigma}\right)^2 (p'+q') - \frac{1}{2} |\boldsymbol{\sigma}_k|^2 \epsilon \Big] + \epsilon \\ =& \frac{1}{2t} \sum_{k=1}^{t} 2\left(p+q+p'+q'-\epsilon\right) |\boldsymbol{\sigma}_k|^2 + \epsilon \\ =& \frac{1-\epsilon-r}{t} \sum_{k=1}^{t} |\boldsymbol{\sigma}_k|^2 + \epsilon.\end{aligned}$$

Similar to the 1D case,

$$\langle\, |\boldsymbol{\sigma}_{t+1}|^2 \big| \, \{\boldsymbol{\sigma}_t\}\rangle = \frac{t-1}{t} \langle\, |\boldsymbol{\sigma}_t|^2 \big| \, \{\boldsymbol{\sigma}_{t-1}\}\rangle + \frac{1-\epsilon-r}{t} |\boldsymbol{\sigma}_t|^2 + \frac{\epsilon}{t},$$

so

$$\begin{aligned}\langle |\boldsymbol{\sigma}_1|^2 \rangle &= 1, \\ \langle |\boldsymbol{\sigma}_{t+1}|^2 \rangle &= \left(1 - \frac{\epsilon + r}{t}\right) \langle |\boldsymbol{\sigma}_t|^2 \rangle + \frac{\epsilon}{t}.\end{aligned}$$

Moreover,

$$\begin{aligned}\langle\, |\boldsymbol{X}_{t+1}|^2 \big| \, \{\sigma_t\}\rangle &= |\boldsymbol{X}_t|^2 + 2\boldsymbol{X}_t \cdot \langle\, \boldsymbol{\sigma}_{t+1} | \, \{\boldsymbol{\sigma}_t\}\rangle + \langle\, \boldsymbol{\sigma}_{t+1}^2 \big| \, \{\boldsymbol{\sigma}_t\}\rangle \\ &= |\boldsymbol{X}_t|^2 + 2\boldsymbol{X}_t \cdot \frac{1}{t} \left(\gamma + \gamma' A\right) \boldsymbol{X}_t + \langle\, \boldsymbol{\sigma}_{t+1}^2 \big| \, \{\boldsymbol{\sigma}_t\}\rangle \\ &= \left(1 + \frac{2\gamma}{t}\right) |\boldsymbol{X}_t|^2 + \langle\, \boldsymbol{\sigma}_{t+1}^2 \big| \, \{\boldsymbol{\sigma}_t\}\rangle,\end{aligned}$$

hence

$$\langle |\boldsymbol{X}_{t+1}|^2 \rangle = \left(1 + \frac{2\gamma}{t}\right) \langle |\boldsymbol{X}_t|^2 \rangle + \langle \boldsymbol{\sigma}_{t+1}^2 \rangle.$$

By Proposition 3.1 and Theorem 3.1,

$$\begin{aligned} \langle |\boldsymbol{\sigma}_t|^2 \rangle &= C\frac{\Gamma\left(t-\epsilon-r\right)}{\Gamma\left(t\right)} + \frac{\epsilon}{\epsilon+r}, \\ \langle |\boldsymbol{X}_t|^2 \rangle &= \frac{\epsilon}{\left(1-2\gamma\right)\left(\epsilon+r\right)}t + \frac{C}{1-\epsilon-r-2\gamma}\frac{\Gamma\left(t+1-\epsilon-r\right)}{\Gamma\left(t\right)} + D\frac{\Gamma\left(t+2\gamma\right)}{\Gamma\left(t\right)}, \end{aligned}$$

where

$$\begin{aligned} C &= \frac{r}{\left(\epsilon+r\right)\Gamma\left(1-\epsilon-r\right)}, \\ D &= -\frac{1}{\Gamma\left(2\gamma\right)}\left[\frac{\epsilon}{\left(\epsilon+r\right)\left(1-2\gamma\right)} + \frac{r}{\left(\epsilon+r\right)\left(1-\epsilon-r-2\gamma\right)}\right]. \end{aligned}$$

Due to the above moment formulas, the residual diffusivity results in 1D extend verbatim to the 2D model.

Chapter 4

Curvature Dependent Flame Speed in Shear Flow

4.1 Curvature Dependent Flame Speed

Let $V : \mathbf{R}^n \to \mathbf{R}^n$ be smooth, $\mathbf{Z}^n$-periodic and incompressible (i.e. $\mathrm{div} V = 0$), the G-equation with mean curvature is

$$G_t + V(x) \cdot DG + |DG| - \tilde{d}\, |DG|\, \mathrm{div} \left(\frac{DG}{|DG|} \right) = 0, \tag{4.1}$$

where $\tilde{d} > 0$ is the Markstein length proportional to the flame thickness.

Turbulent combustion usually involves small scales. As a simplified model, we rescale V as $V = V(\frac{x}{\epsilon})$ and write $\tilde{d} = d\epsilon$. Here ϵ denotes the Kolmogorov scale (the small scale in the flow). The diffusivity constant $d > 0$ is called the Markstein number. The dimensionless Markstein number is $d \cdot \frac{\delta_L}{\epsilon}$ with δ_L denoting the flame thickness [56]. In the thin reaction zone regime, $\delta_L = \mathcal{O}(\epsilon)$, see Equation (2.28) and Figure 2.8 of [56]. Without loss of generality, let

$\frac{\delta_L}{\epsilon} = 1$. Then (4.1) becomes

$$G^\epsilon_t + V(\frac{x}{\epsilon}) \cdot DG^\epsilon + |DG^\epsilon| - d\,\epsilon\,|DG^\epsilon| \text{ div}\left(\frac{DG^\epsilon}{|DG^\epsilon|}\right) = 0. \tag{4.2}$$

Since $\epsilon \ll 1$, it is natural to look at $\lim_{\epsilon\to 0} G^\epsilon$, i.e. the homogenization limit. If for any $p \in \mathbf{R}^n$, there exists a unique number $\overline{H}_d(p)$ such that the following cell problem has (approximate) $\mathbf{Z}^n$-periodic viscosity solutions in $\mathbf{R}^n$:

$$-d\,|p+Dw| \text{ div}\left(\frac{p+Dw}{|p+Dw|}\right) + |p+Dw| + V(y)\cdot(p+Dw) = \overline{H}_d(p), \tag{4.3}$$

then standard tools in the homogenization theory imply that

$$\lim_{\epsilon\to 0} G^\epsilon(x,t) = \bar{G}(x,t) \quad \text{locally uniformly in } \mathbf{R}^n \times [0,+\infty).$$

Here $\bar{G}$ is the unique solution to the following effective equation, which captures the propagation of the mean flame front (see Figure 4.1 below).

$$\begin{cases} \bar{G}_t + \overline{H}_d(D\bar{G}) = 0, \\ \bar{G}(x,0) = G_0(x). \end{cases} \tag{4.4}$$

Solution to the cell problem (4.3) formally describes fluctuations around the mean flame front,

$$G(x,t) = \bar{G}(x,t) + \epsilon w(x,\frac{x}{\epsilon}) + O(\epsilon^2).$$

Here for fixed location-time (x,t) and $p = D\bar{G}(x,t)$, $w(x,\cdot)$ is a solution to (4.3) with mean zero, i.e. $\int_0^1 w(x,y)\,dy = 0$. The quantity $\overline{H}_d(p)$, if it exists, can be viewed as the turbulent

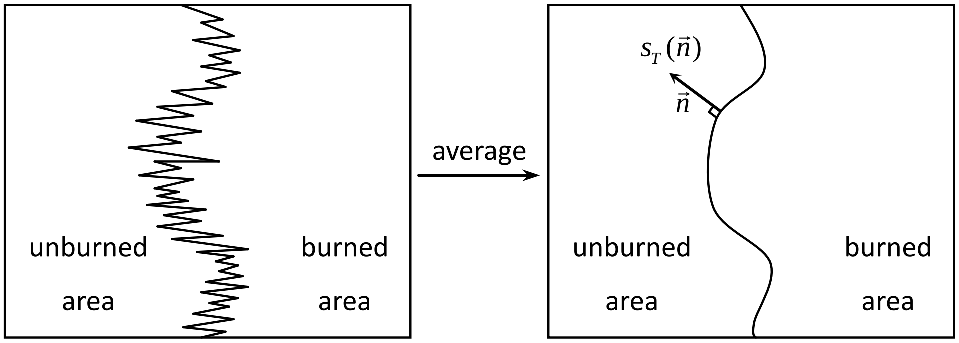


Figure 4.1: Average of fluctuations in the homogenization limit.

flame speed $s_T(p)$ along a given direction p.

For general V, we do not even know the existence of $\overline{H}_d(p)$, i.e. the well-posedness of (4.3). In fact, given the counter-example in [6] for a coercive mean curvature type equation, the cell problem (4.3) and the homogenization in our non-coercive setting is very likely not well-posed in general. To avoid this existence issue, we consider the shear flow

$$V(x) = (v(x_2), 0) \quad \text{for } x = (x_1, x_2) \in \mathbf{R}^2,$$

where $v : \mathbf{R} \to \mathbf{R}$ is smooth and periodic. For $p = (\gamma, \mu) \in \mathbf{R}^2$, the cell problem (4.3) is reduced to the following ODE:

$$-\frac{d\gamma^2 w''}{\gamma^2 + (\mu + w')^2} + \sqrt{\gamma^2 + (\mu + w')^2} + \gamma v(y) = \overline{H}_d(p). \tag{4.5}$$

It is then easy to show that there exists a unique number $\overline{H}_d(p)$ such that the ODE (4.5) has a C^2 periodic solution. Throughout this chapter, we denote w as the unique solution satisfying $w(0) = 0$. To simplify notations, we omit the dependence of w on d.

4.2 Slowdown of Flame Propagation

4.2.1 Key inequalities

Let $n \in \mathbf{N}$, $\{b_{ik}\}_{1\le i,k\le n}$ and $\{\tilde{b}_{ik}\}_{1\le i,k\le n}$ be two sequences of positive numbers such that

$$\sum_{l=1}^{i} b_{il} + \sum_{l=i}^{n} \tilde{b}_{il} = \sum_{l=k}^{n} b_{lk} + \sum_{l=1}^{k} \tilde{b}_{lk} = c, \quad \forall i, k,$$

where c is a constant independent of i and k,

$$\min\{\min_{1\le k\le i\le n} b_{ik}, \min_{1\le i\le k\le n} \tilde{b}_{ik}\} \ge \tau > 0. \tag{4.6}$$

Lemma 4.1. *Assume $L > 0$ and $g \in C((0, L])$ satisfies*

$$g'(a) \le -\theta \quad \textit{for some } \theta \ge 0,$$

then

$$\sum_{i=1}^{n} a_i \sum_{k=1}^{i} g(a_k) b_{ik} + \sum_{i=1}^{n} a_i \sum_{k=i}^{n} g(a_k)\tilde{b}_{ik} \ge c\sum_{i=1}^{n} a_i g(a_i) + \frac{\theta\tau}{2} \sum_{1\le i,k\le n} (a_i - a_k)^2.$$

for all $(a_1, a_2, ..., a_n) \in (0, L]^n$. Here τ is from (4.6). Moreover, if $\theta > 0$, the equality holds if and only if $a_1 = a_2 = \cdots = a_n$.

Proof. By approximation, we may assume that $\theta > 0$. Define

$$W(a_1, a_2, \dots, a_n) = \sum_{i=1}^{n} a_i \sum_{k=1}^{i} g(a_k) b_{ik} + \sum_{i=1}^{n} a_i \sum_{k=i}^{n} g(a_k)\tilde{b}_{ik},$$

$$H(a_1, a_2, \dots, a_n) = c\sum_{i=1}^{n} a_i g(a_i) + \frac{\theta\tau}{2} \sum_{1\le i,k\le n} (a_i - a_k)^2.$$

It suffices to show that for any fixed $r \in (0, L)$

$$\min_{[r,L]^n}(W - H) = 0, \tag{4.7}$$

and the minimum is attained when all a_i are the same.

Choose $(\hat{a}_1, \hat{a}_2, \hat{a}_3, \dots \hat{a}_n) \in [r, L]^n$ such that

$$W(\hat{a}_1, \hat{a}_2, \hat{a}_3, \dots \hat{a}_n) - H(\hat{a}_1, \hat{a}_2, \hat{a}_3, \dots \hat{a}_n) = \min_{[r,L]^n}(W - H).$$

Let $\hat{a}_j = \max_{1 \le i \le n}\{\hat{a}_i\}$. If $\hat{a}_j = r$, then $\hat{a}_1 = \hat{a}_2 = \cdots = \hat{a}_n = r$ and (4.7) holds. Assume $\hat{a}_j > r$, then

$$W_{a_j} - H_{a_j} \le 0 \quad \text{at } (\hat{a}_1, \hat{a}_2, \hat{a}_3, ..\hat{a}_n).$$

Here we include "< 0" since $\hat{a}_j$ might be equal to L. Accordingly,

$$\begin{aligned}&\sum_{k=1}^{j} g(\hat{a}_k) b_{jk} + \sum_{k=j}^{n} g(\hat{a}_k)\tilde{b}_{jk} + g'(\hat{a}_j)\sum_{k=j}^{n} \hat{a}_k b_{kj} + g'(\hat{a}_j)\sum_{k=1}^{j} \hat{a}_k \tilde{b}_{kj} \\ &\le c(g(\hat{a}_j) + \hat{a}_j g'(\hat{a}_j)) + \sum_{k \ne j} \theta\tau(\hat{a}_j - \hat{a}_k).\end{aligned}$$

On the other hand, since $g' \le -\theta < 0$,

$$\begin{aligned}&\sum_{k=1}^{j} g(\hat{a}_k) b_{jk} + \sum_{k=j}^{n} g(\hat{a}_k)\tilde{b}_{jk} + g'(\hat{a}_j)\sum_{k=j}^{n} \hat{a}_k b_{kj} + g'(\hat{a}_j)\sum_{k=1}^{j} \hat{a}_k \tilde{b}_{kj} \\ &\ge c(g(\hat{a}_j) + \hat{a}_j g'(\hat{a}_j)) + \sum_{k \ne j} \theta\tau(\hat{a}_j - \hat{a}_k).\end{aligned}$$

Hence all equalities should hold and $\hat{a}_1 = \hat{a}_2 = \cdots = \hat{a}_n$ follows from that g is strictly decreasing. Therefore $W(\hat{a}_1, \hat{a}_2, \hat{a}_3, \dots \hat{a}_n) - H(\hat{a}_1, \hat{a}_2, \hat{a}_3, \dots \hat{a}_n) = 0$. □

Theorem 4.1. *Let $T > 0$ and $f \in C([0, T])$ be a continuous positive function. Suppose that*

$g \in C^1((0, L])$ for $L = \max_{[0,T]} f$.

(1) If $g' \le -\theta$ for some $\theta \ge 0$, then

$$\begin{aligned} &e^T \int_0^T f(x)e^{-x} \int_0^x g(f(y))e^y \, dydx + \int_0^T f(x)e^{-x} \int_x^T g(f(y))e^y \, dydx \\ &\ge (e^T - 1) \int_0^T f(x)g(f(x))) \, dx + \frac{\theta}{2} \int_{[0,T]^2} |f(x) - f(y)|^2 \, dxdy. \end{aligned}$$

(2) If $g' \ge \theta$ for some $\theta \ge 0$, then

$$\begin{aligned} &e^T \int_0^T f(x)e^{-x} \int_0^x g(f(y))e^y \, dydx + \int_0^T f(x)e^{-x} \int_x^T g(f(y))e^y \, dydx \\ &\le (e^T - 1) \int_0^T f(x)g(f(x))) \, dx - \frac{\theta}{2} \int_{[0,T]^2} |f(x) - f(y)|^2 \, dxdy. \end{aligned}$$

Proof. (1) For $n \in \mathbf{N}$, let $x_i = \dfrac{iT}{n}$ for $i = 1, 2, \dots, n$. Note that for $i, k = 1, 2, 3, \dots, n$,

$$\sum_{l=1}^{i} e^{T - x_i + x_l} + \sum_{l=i}^{n} e^{x_l - x_i} = \frac{e^{T + \frac{T}{n}} - 1}{e^{\frac{T}{n}} - 1} = \sum_{l=k}^{n} e^{T - x_l + x_k} + \sum_{l=1}^{k} e^{x_k - x_l}.$$

Then desired inequality in (1) follows from Lemma 4.1 and Riemann sum approximation by taking $a_i = f(x_i)$, $c = \dfrac{e^{T+\frac{T}{n}} - 1}{e^{\frac{T}{n}} - 1}$, $\tau = 1$,

$$b_{ik} = e^{T - x_i + x_k} \quad \text{and} \quad \tilde{b}_{ik} = e^{x_k - x_i} \quad \text{for } 1 \le i, k \le n.$$

(2) follows immediately from (1) by considering $-g$. □

4.2.2 Strict decreasing of $\overline{H}_d(p)$

Lemma 4.2. *Let $d > 0$ and ϕ be a non-constant C^1 periodic function. If the following equation has a mean-zero, periodic solution F*

$$-dF' + b(x)F = \phi' + \alpha(1+\phi^2) \quad \textit{in } \mathbf{R}$$

for some $\alpha \in \mathbf{R}$ and

$$b(x) = \frac{2d\phi'\phi}{1+\phi^2} + \phi\sqrt{1+\phi^2},$$

then

$$\alpha < 0.$$

Proof. It suffices to prove this for $d = 1$. The proof for other d's is similar. F can be solved in terms of ϕ and α. Since F is periodic and mean zero (i.e. $F(0) = F(1)$ and $\displaystyle\int_0^1 F(s)\,ds = 0$),

$$\alpha = -\frac{e^{g(1)}\int_0^1 \phi' e^{-g(x)}\,dx \int_0^1 e^{g(x)}\,dx - (e^{g(1)}-1)\int_0^1 e^{g(x)}\int_0^x \phi' e^{-g(y)}\,dydx}{e^{g(1)}\int_0^1 (1+\phi^2)e^{-g(x)}\,dx \int_0^1 e^{g(x)}\,dx - (e^{g(1)}-1)\int_0^1 e^{g(x)}\int_0^x (1+\phi^2)e^{-g(y)}\,dydx}.$$

Here

$$g(x) = \int_0^x b(y)\,dy = \log(1+\phi^2(x)) - \log(1+\phi^2(0)) + \int_0^x \phi\sqrt{1+\phi^2}\,dx.$$

In particular, $g(1) = \displaystyle\int_0^1 \phi\sqrt{1+\phi^2}\,dx$. The denominator is obviously positive. Hence $\alpha < 0$ is equivalent to proving the inequality

$$e^{g(1)}\int_0^1 \phi' e^{-g(x)}\,dx \int_0^1 e^{g(x)}\,dx > (e^{g(1)}-1)\int_0^1 e^{g(x)}\int_0^x \phi' e^{-g(y)}\,dydx.$$

for every non-constant C^1 periodic function ϕ. Denote that

$$h(x) = \int_0^x \phi\sqrt{1+\phi^2}\, dy,$$

then it is equivalent to showing that

$$\begin{aligned} & e^{h(1)} \int_0^1 \frac{\phi'}{1+\phi^2} e^{-h(x)}\, dx \int_0^1 (1+\phi^2)e^{h(x)}\, dx \\ > \; & (e^{h(1)}-1) \int_0^1 (1+\phi^2)e^{h(x)} \int_0^x \frac{\phi'}{1+\phi^2} e^{-h(y)}\, dydx. \end{aligned}$$

Write $\lambda(\phi) = \arctan\phi$. Using integration by parts and $\phi(0)=\phi(1)$, we have

$$\begin{aligned} LHS =& e^{h(1)} \left(\lambda(\phi(1))e^{-h(1)} - \lambda(\phi(1)) + \int_0^1 \lambda(\phi)e^{-h(x)}\phi\sqrt{1+\phi^2} dx \right) \int_0^1 (1+\phi^2)e^{h(x)} dx, \\ RHS =& (e^{h(1)}-1) \left(\int_0^1 \lambda(\phi)(1+\phi^2)\, dx - \lambda(\phi(1)) \int_0^1 (1+\phi^2)e^{h(x)}\, dx \right) \\ & + (e^{h(1)}-1) \left(\int_0^1 (1+\phi^2)e^{h(x)} \int_0^x \lambda(\phi)e^{-h(y)}\phi\sqrt{1+\phi^2}\, dydx \right). \end{aligned}$$

By Fubini Theorem,

$$\begin{aligned} & \int_0^1 (1+\phi^2)e^{h(x)} \int_0^x \lambda(\phi)e^{-h(y)}\phi\sqrt{1+\phi^2}\, dydx \\ = & \int_0^1 \lambda(\phi)e^{-h(x)}\phi\sqrt{1+\phi^2} \int_x^1 (1+\phi^2)e^{h(y)}\, dydx, \end{aligned}$$

then $LHS - RHS$ is $A + B - C$ where

$$\begin{aligned} A(\phi) &= e^{h(1)} \int_0^1 \lambda(\phi)e^{-h(x)}\phi\sqrt{1+\phi^2} \int_0^x (1+\phi^2)e^{h(y)}\, dydx, \\ B(\phi) &= \int_0^1 \lambda(\phi)e^{-h(x)}\phi\sqrt{1+\phi^2} \int_x^1 (1+\phi^2)e^{h(y)}\, dydx, \\ C(\phi) &= (e^{h(1)}-1) \int_0^1 \lambda(\phi)(1+\phi^2)\, dx. \end{aligned}$$

If $h(1) = 0$, then $A + B - C = A + B \geq 0$ since $s\lambda(s) \geq 0$. Cleary, "$= 0$" if and only if $\phi \equiv 0$. Hence we assume

$$h(1) \neq 0.$$

Also, for $\tilde{\phi}(x) = -\phi(-x)$, the correspsonding

$$\tilde{b}(x) = \frac{2\tilde{\phi}'\tilde{\phi}}{1 + \tilde{\phi}^2} + \tilde{\phi}\sqrt{1 + \tilde{\phi}^2} = -b(-x)$$

and $\tilde{F}(x) = -F(-x)$ satisfy

$$-\tilde{F}' + \tilde{b}(x)\tilde{F} = \tilde{\phi}' + \alpha(1 + \tilde{\phi}^2).$$

Without lost of generality, we may further assume

$$h(1) > 0.$$

Let $\phi_+ = \max\{\phi, 0\}$, $\phi_- = \min\{\phi, 0\}$ and

$$h^{\pm}(x) = \int_0^x \phi_{\pm}\sqrt{1 + \phi_{\pm}^2}\, dy,$$

then $h(x) = h^+ + h^-$.

Moreover,

$$A(\phi) + B(\phi) - C(\phi) \geq e^{h^-(1)} \left(A(\phi_+) + B(\phi_+) - C(\phi_+)\right), \tag{4.8}$$

and equality holds if only if $\phi \geq 0$, i.e. $\phi_- = 0$. In fact,

$$\begin{aligned}
A(\phi) &\geq e^{h(1)} \int_0^1 \lambda(\phi_+) e^{-h(x)} \phi_+ \sqrt{1+\phi_+^2} \int_0^x (1+\phi_+^2) e^{h(y)} \, dydx \\
&= e^{h(1)} \int_0^1 \lambda(\phi_+) e^{-h^+(x)} \phi_+ \sqrt{1+\phi_+^2} \int_0^x (1+\phi_+^2) e^{h^+(y)} e^{h^-(y)-h^-(x)} \, dydx \\
&\geq e^{h^-(1)} A(\phi_+), \quad (\text{since } h^-(x) \leq h^-(y) \text{ for } x \geq y) \\
B(\phi) &\geq \int_0^1 \lambda(\phi_+) e^{-h(x)} \phi_+ \sqrt{1+\phi_+^2} \int_x^1 (1+\phi_+^2) e^{h(y)} \, dydx \\
&= e^{h^-(1)} \int_0^1 \lambda(\phi_+) e^{-h^+(x)} \phi_+ \sqrt{1+\phi_+^2} \int_x^1 (1+\phi_+^2) e^{h^+(y)} e^{h^-(y)-h^-(1)} e^{-h^-(x)} \, dydx \\
&\geq e^{h^-(1)} B(\phi_+), \quad (\text{since } 0 \geq h^-(y) \geq h^-(1) \text{ for } y \in [0,1]) \\
C(\phi) &\leq (e^{h(1)} - 1) \int_0^1 \lambda(\phi_+)(1+\phi_+^2) \, dx \\
&= \frac{(e^{h(1)} - 1)}{(e^{h^+(1)} - 1)} C(\phi_+) \\
&\leq e^{h^-(1)} C(\phi_+),
\end{aligned}$$

and for all inequalities to hold, we must have $h^- \equiv 0$ and $\phi_- \equiv 0$.

Since $h(1) > 0$, that ϕ is not constant implies ϕ_+ is not constant either. By a small perturbation like $\phi_+ + \epsilon$, we may assume $\phi_+ > 0$. Then h^+ is strictly increasing. After changing of variables $h^+(x) \to x$ and writing $\psi(h^+(x)) = \phi_+(x)$ and $T = h^+(1)$, we obtain that

$$\begin{aligned}
A(\phi_+) &= A_{T,\psi} = e^T \int_0^T \lambda(\psi) e^{-x} \int_0^x \frac{\sqrt{1+\psi^2}}{\psi} \, e^y \, dydx, \\
B(\phi_+) &= B_{T,\psi} = \int_0^T \lambda(\psi) e^{-x} \int_x^T \frac{\sqrt{1+\psi^2}}{\psi} \, e^y \, dydx, \\
C(\phi_+) &= C_{T,\psi} = (e^T - 1) \int_0^T \lambda(\psi) \frac{\sqrt{1+\psi^2}}{\psi} \, dx.
\end{aligned}$$

Hence

$$\begin{aligned}A_{T,\psi}+B_{T,\psi}-C_{T,\psi}=&e^T\int_0^T\lambda(\psi)e^{-x}\int_0^x\frac{\sqrt{1+\psi^2}}{\psi}\,e^y\,dydx\\&+\int_0^T\lambda(\psi)e^{-x}\int_x^T\frac{\sqrt{1+\psi^2}}{\psi}\,e^y\,dydx\\&-(e^T-1)\int_0^T\lambda(\psi)\frac{\sqrt{1+\psi^2}}{\psi}\,dx.\end{aligned}$$

Let $M=\max_{[0,T]}\psi=\max_{[0,1]}\phi_+>0$. According to Theorem 4.1 by taking $f(x)=\lambda(\psi)=\arctan(\psi)$, $g(y)=\dfrac{1}{\sin y}$, $L=\arctan(M)$ and $\theta=\dfrac{1}{\sqrt{1+M^2}}$, we have $\dfrac{\sqrt{1+\psi^2}}{\psi}=g(f)$ and

$$\begin{aligned}A_{T,\psi}+B_{T,\psi}-C_{T,\psi}&\geq\frac{1}{2\sqrt{1+M^2}}\int_{[0,T]^2}|\lambda(\psi(x))-\lambda(\psi(y))|^2\,dxdy\\&=\frac{1}{2\sqrt{1+M^2}}\int_{[0,1]^2}|\lambda(\phi_+(x))-\lambda(\phi_+(y))|^2J(x)J(y)\,dxdy\\&>0,\end{aligned}$$

since ϕ_+ is not constant. Here $J(x)=\phi_+(x)\sqrt{1+\phi_+^2}$. It follows from (4.8) that $A(\phi)+B(\phi)-C(\phi)>0$. □

The following is our main result.

Theorem 4.2. *Assume that $v=v(y)$ is not a constant function.*

(1) $\overline{H}_d(0,\pm\mu)=|\mu|$.

(2) **(Major Part).** *If $\gamma\neq 0$, then*

$$\frac{\partial\overline{H}_d(p)}{\partial d}<0,$$

so $\overline{H}_d$ is strictly decreasing with respect to the Markstein number d.

(3) $\lim\limits_{d\to 0^+}\overline{H}_d = \overline{H}_0$. *Here* $\overline{H}_0(p)$ *is the unique number (effective Hamiltonian) such that the following inviscid equation admits periodic viscosity solutions*

$$\sqrt{\gamma^2 + (\mu + w_0')^2} + \gamma v(y) = \overline{H}_0(p) \quad \textit{in } \mathbf{R}.$$

(4) $\lim\limits_{d\to+\infty}\overline{H}_d = |p| + \gamma\int_0^1 v(y)\,dy$ *and* $\lim\limits_{d\to+\infty} w = 0$ *uniformly in* $\mathbf{R}$.

Proof. (1) is trivial.

(2) Fix (γ,μ). Denote $\phi = \dfrac{\mu + w'}{\gamma}$. Then ϕ is the unique periodic solution to

$$-\frac{d\phi'}{1+\phi^2} + \sqrt{1+\phi^2} + v(y) = E(d) = \frac{\overline{H}_d(p)}{\gamma} \quad \text{in } \mathbf{R}$$

subject to $\int_0^1 \phi(x)\,dx = \dfrac{\mu}{\gamma}$. To prove (2) is equivalent to showing that

$$E'(d) < 0.$$

Taking derivative on both sides of the above equation with respect to d, we obtain

$$-dF' + b(x)F = E'(d)(1+\phi^2) + \phi',$$

where $b(x) = \dfrac{2d\phi'\phi}{1+\phi^2} + \phi\sqrt{1+\phi^2}$ and $F(x) = \phi_d(x)$, i.e. the derivative of ϕ with respect to d. Clearly, F is periodic and has zero mean, i.e. $\int_{[0,1]} F = 0$. Note that v is not constant is equivalent to saying the ϕ is not constant. Then (2) follows immediately from Lemma 4.2.

(3) Integrating both sides of (4.5), we have

$$\overline{H}_d(p) = \int_0^1 \sqrt{\gamma^2 + (\mu + w')^2}\,dy + \gamma\int_0^1 v(y)\,dy. \tag{4.9}$$

Due to the convexity of $s(t) = \sqrt{\gamma^2 + t^2}$,

$$\overline{H}_d(p) \geq |p| + \gamma \int_0^1 v(y)\, dy.$$

Also, by maximum principle,

$$\overline{H}_d(p) \leq |p| + \max_{\mathbf{R}} \gamma v$$

and

$$\max_{\mathbf{R}} |\mu + w'| \leq \overline{H}_d(p) - \min_{\mathbf{R}} \gamma v \leq |p| + 2 \max_{\mathbf{R}} |\gamma v|.$$

Hence, up to a sequence, we may assume that

$$\lim_{d \to 0} \overline{H}_d = \overline{H}_0 \quad \text{and} \lim_{d \to 0^+} w = w_0 \quad \text{uniformly in } \mathbf{R}.$$

The stability of viscosity solution immediately implies that w_0 is a continuous periodic viscosity solution to

$$\sqrt{\gamma^2 + \left(\mu + w_0'\right)^2} + \gamma v(y) = \overline{H}_0(p) \quad \text{in } \mathbf{R}.$$

Note that $\overline{H}_0(p)$ is unique number such that the above equation has a periodic viscosity solutions w_0 although w_0 might not be unique.

(4) If $\gamma = 0$, this is trivial. We assume $\gamma \neq 0$. Note that estimates of $\overline{H}_d$ and $\mu + w'$ in (3) are independent of d. Since

$$w'' = \frac{1}{d\gamma^2}(\gamma^2 + (\mu + w')^2) \left(\sqrt{\gamma^2 + (\mu + w')^2} + v - \overline{H}_d(\mu) \right),$$

$$\max_{\mathbf{R}} |w''| \leq \frac{C}{d}$$

for a constant C independent of d. Due to the periodicity of w and $w(0) = 0$, it is obvious that

$$\lim_{d \to +\infty} w = \lim_{d \to +\infty} w' = 0 \quad \text{uniformly in } \mathbf{R}.$$

(4) follows from (4.9). □

4.3 Selection of Physical Fluctuations as $d \to 0$

To have a more complete picture, it is also interesting to ask what is the limit of solutions of (4.5) as $d \to 0^+$ (the vanishing curvature limit). When $d = 0$, equation (4.2) becomes the inviscid G-equation

$$G^\epsilon_t + V(\frac{x}{\epsilon}) \cdot DG^\epsilon + |DG^\epsilon| = 0.$$

It is proved in [69] and [10] independently that there exists a unique $\overline{H}_0(p)$ such that the corresponding cell problem

$$|p + Dw| + V(y) \cdot (p + Dw) = \overline{H}_0(p) \quad \text{in } \mathbf{R}^n \tag{4.10}$$

admits a periodic (approximate) viscosity solution, which implies

$$\lim_{\epsilon \to 0} G^\epsilon(x,t) = \bar{G}(x,t) \quad \text{locally uniformly in } \mathbf{R} \times [0, +\infty).$$

As in the curvature case, $\bar{G}$ is the unique solution to the following effective equation, which captures the propagation of the mean flame front:

$$\begin{cases} \bar{G}_t + \overline{H}_0(D\bar{G}) = 0 \\ \bar{G}(x,0) = G_0(x) \quad \text{initial flame front.} \end{cases}$$

The formal two-scale expansion says that

$$G_\epsilon(x,t) = \bar{G}(x,t) + \epsilon w(x, \frac{x}{\epsilon}) + O(\epsilon^2),$$

where the fluctuation $w(x,\cdot)$ is a solution to (4.10) with $p = D\bar{G}(x,t)$ for fixed (x,t). Nevertheless, solutions to (4.10) are in general not unique even up to a constant. To find the physical solution that captures the fluctuation of flame front, a natural approach is to look at the limit of solutions to (4.3) (if it exists uniquely) as $d \to 0$. The limit is however, very challenging and unknown in general. In this section, we identify the limit for the equation (4.5) under some non-degeneracy conditions.

It is easy to show that as $d \to 0^+$, the solution w to (4.5), up to a subsequence, converges to a periodic viscosity solution w_0 of

$$\sqrt{\gamma^2 + (\mu + w_0')^2} + \gamma v(y) = \overline{H}_0(p) \quad \text{in } \mathbf{R}. \tag{4.11}$$

When $\gamma = 0$, $w = w_0 \equiv 0$. Without loss of generality, we set $\gamma = 1$ in this section and denote

$$\overline{H}_0(\mu) = \overline{H}_0(p).$$

We also assume

$$\max_{\mathbf{R}} v = 0.$$

4.3.1 Uniqueness case

If $|\mu| \geq \int_0^1 \sqrt{(1-v)^2 - 1}\,dy$, $\overline{H}(\mu) \geq 1$ is the unique number such that

$$|\mu| = \int_0^1 \sqrt{(\overline{H}(\mu) - v(y))^2 - 1}\,dy.$$

Also, the inviscid equation (4.11) has a unique solution up to a constant, i.e.

$$w_0(x) = (sign(\mu)) \int_0^x \sqrt{(\overline{H}(\mu) - v(y))^2 - 1}\,dy - \mu x + c$$

for some $c \in \mathbf{R}$ since $w_0' + \mu$ can not change signs. By $w(0) = 0$,

$$\lim_{d \to 0^+} w = (sign(\mu)) \int_0^x \sqrt{(\overline{H}(\mu) - v(y))^2 - 1}\,dy - \mu x.$$

4.3.2 Non-uniqueness case

When $|\mu| < \int_0^1 \sqrt{(1-v)^2 - 1}\,dy$, $\overline{H}_d(\mu) = 1$. Solutions to the inviscid equation (4.11) are not unique if the set

$$\mathcal{M}_0 = \{x \in [0,1) |\ v(x) = \max_{\mathbf{R}} v = 0\}$$

has multiple points. For example, assume that $x_i \in \mathcal{M}_0$ for $i = 1, 2$. Choose $x_{\mu,i} \in (x_i, x_i{+}1)$ such that

$$\int_{x_i}^{x_{\mu,i}} \sqrt{(1-v)^2 - 1}\,dy - \int_{x_{\mu,i}}^{x_i+1} \sqrt{(1-v)^2 - 1}\,dy = \mu,$$

then

$$w_i(x) = \begin{cases} \displaystyle\int_{x_i}^{x} \sqrt{(1-v(y))^2-1}\,dy - \mu x, \ \forall x \in [x_i, x_{\mu,i}] \\ \displaystyle\int_{x_i}^{x_{\mu,i}} \sqrt{(1-v(y))^2-1}\,dy - \int_{x_{\mu,i}}^{x} \sqrt{(1-v(y))^2-1}\,dy - \mu x, \ \forall x \in [x_{\mu,i}, x_i+1] \end{cases}$$

(extended periodically) are both viscosity solutions to (4.11) and $w_1 - w_2$ is not a constant. Hence a very interesting problem is to identify the solution selected by the limiting process, i.e. the physical fluctuation associated with the inviscid G-equation model. We assume that

$$\mathcal{M}_0 \text{ is finite and } v''(x) \text{ is distinct for } x \in \mathcal{M}_0 \tag{4.12}$$

Choose the unique $\bar{x} \in \mathcal{M}_0$ then $x_\mu \in (\bar{x}, \bar{x}+1)$ such that

$$v''(\bar{x}) = \min_{x \in \mathcal{M}_0} \{-v''(x)\},$$
$$\int_{\bar{x}}^{x_\mu} \sqrt{(1-v)^2-1}\,dy - \int_{x_\mu}^{\bar{x}+1} \sqrt{(1-v)^2-1}\,dy = \mu.$$

Clearly, such x_μ is unique.

Theorem 4.3.

$$\lim_{d \to 0^+} w = w_0(x) - w_0(0) \quad \textit{uniformly in } \mathbf{R}.$$

Here

$$w_0(x) = \begin{cases} \displaystyle\int_{\bar{x}}^{x} \sqrt{(1-v)^2-1}\,dy - \mu x, \ \forall x \in [\bar{x}, x_\mu] \\ \displaystyle\int_{\bar{x}}^{x_\mu} \sqrt{(1-v)^2-1}\,dy - \int_{x_\mu}^{x} \sqrt{(1-v)^2-1}\,dy - \mu x, \ \forall x \in [x_\mu, \bar{x}+1]. \end{cases} \tag{4.13}$$

We would like to point out that selection problems of similar spirit have been studied for the vanishing viscosity limit ([33], [1], [2], etc), after which the viscosity solution was originally named. The authors aim to identify $\lim_{\epsilon\to 0^+} v_\epsilon$, where v_ϵ is the unique smooth solution to

$$-\epsilon\Delta v_\epsilon + H(p+Dv_\epsilon, x) = \overline{H}(p,\epsilon) \quad \text{in } \mathbf{R}^n.$$

The most important case is the mechanical Hamiltonian $H(p,x) = |p|^2 + G(x)$ with a potential function G. The limiting process resembles the passage from quantum mechanics to classical mechanics ([1], [24]). The works [1] and [2] deal with some special cases in high dimensions by employing advanced tools from dynamical systems and random perturbations. Assumptions therein are very hard to check. The method in [33] is purely 1D. Based on simple comparison principles of PDEs/ODEs, our arguments are simpler and more robust. In particular, they can be easily extended to handle certain cases in high dimensions.

Lemma 4.3. *Assume that $\mathcal{M}_0 = \{\bar{x}\}$, i.e. it contains a single element, then*

$$\lim_{d\to 0^+} \frac{\overline{H}_d(\mu) - 1}{d} = -\sqrt{-v''(\bar{x})}.$$

Proof. Since $\mathcal{M}_0$ has only one element, $1 - v > 1$ in $(\bar{x}, \bar{x}+1)$. It is easy to see that periodic viscosity solutions to

$$\sqrt{1 + (\mu + w_0')^2} + v(y) = 1 \quad \text{in } \mathbf{R}$$

are unique up to a constant. Hence, since $w(0) = 0$,

$$\lim_{d\to 0^+} w = w_0(x) - w_0(0) \quad \text{uniformly in } \mathbf{R}. \tag{4.14}$$

Here w_0 is given by (4.13). Fix $\delta > 0$ and let

$$u_{\delta,\pm}(x) = \begin{cases} \displaystyle\int_{\bar{x}}^{x} \sqrt{(1-(1\pm\delta)v)^2 - 1}\,dy & \text{for } x \geq \bar{x}, \\ \displaystyle\int_{x}^{\bar{x}} \sqrt{(1-(1\pm\delta)v)^2 - 1}\,dy & \text{for } x \leq \bar{x}. \end{cases}$$

Apparently,

$$u_{\delta,-}(x) < u_0(x) = w_0(x) + \mu x < u_{\delta,+}(x) \quad \text{for } x \in [x_\mu - 1, x_\mu] \backslash \{\bar{x}\}$$

and $u_{\delta,-}(\bar{x}) = u_0(\bar{x}) = u_{\delta,+}(\bar{x}) = 0$. See Figure 4.2. Denote

$$e_\delta = \min_{x = x_\mu \text{ or } x_\mu - 1} \{u_0(x) - u_{\delta,-}(x),\ u_{\delta,+}(x) - u_0(x)\} > 0,$$

and let

$$u_{d,\delta,\pm}(x) = w(x) - w(\bar{x}) + \mu(x - \bar{x}) \pm \frac{1}{2} e_\delta.$$

By (4.14), when d is small enough, there exist $x_{d,\delta,\pm} \in (x_\mu - 1, x_\mu)$ such that

$$u_{d,\delta,+}(x_{d,\delta,+}) - u_{\delta,+}(x_{d,\delta,+}) \geq u_{d,\delta,+}(x) - u_{\delta,+}(x) \quad \forall x \in (x_\mu - 1, x_\mu),$$

$$u_{d,\delta,-}(x_{d,\delta,-}) - u_{\delta,-}(x_{d,\delta,-}) \leq u_{d,\delta,-}(x) - u_{\delta,-}(x) \quad \forall x \in (x_\mu - 1, x_\mu).$$

Maximum principle implies that

$$-\frac{du''_{\delta,+}}{1 + (u'_{\delta,+})^2} + \sqrt{1 + (u'_{\delta,+})^2} + v \leq \overline{H}_d(\mu) \quad \text{at } x_{d,\delta,+}.$$

Hence

$$-\frac{du''_{\delta,+}}{1 + (u'_{\delta,+})^2} \leq \overline{H}_d(\mu) - 1 + \delta v \leq \overline{H}_d(\mu) - 1 \quad \text{at } x_{d,\delta,+}.$$

Let $d \to 0$ first then $\delta \to 0$, we have $x_{d,\delta,+} \to \bar{x}$ and

$$\liminf_{d \to 0^+} \frac{\overline{H}_d(\mu) - 1}{d} \geq -\sqrt{-v''(\bar{x})}.$$

Similarly with $x_{d,\delta,-}$, we can obtain that

$$\limsup_{d \to 0^+} \frac{\overline{H}_d(\mu) - 1}{d} \leq -\sqrt{-v''(\bar{x})}.$$

□

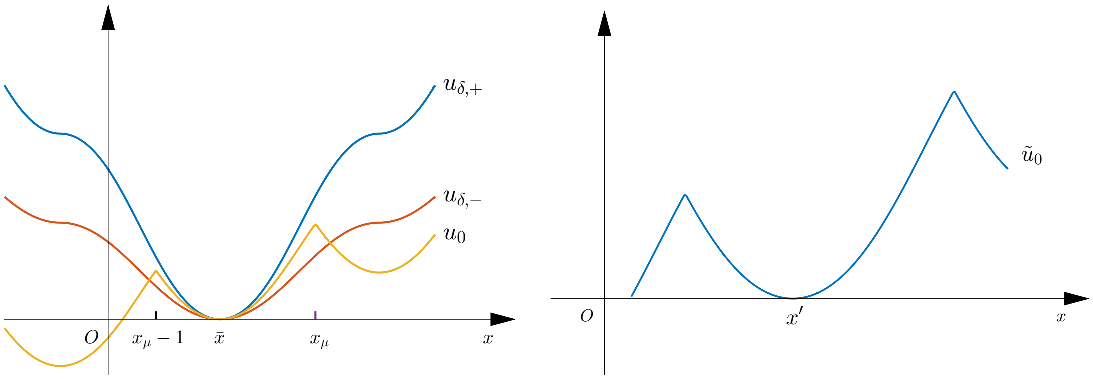


Figure 4.2: Graphes of $u_{\delta,\pm}$, u_0, and turning point.

Remark 4.1. *The above proof based on comparison and maximum principle also shows that for any subsequence $\{d_m\} \to 0$, if*

$$\lim_{d_m \to 0^+} w(x) = \tilde{w}_0(x)$$

and $\tilde{u}_0 = \mu x + \tilde{w}_0(x)$ has turning point at some $x' \in \mathcal{M}$, i.e there exists $\tau > 0$ (see Figure 4.2) such that

$$\tilde{u}_0(x) - \tilde{u}_0(x') = \begin{cases} \int_{x'}^{x} \sqrt{(1-v)^2 - 1}\, dy & \textit{for } x \in [x', x' + \tau], \\ \int_{x}^{x'} \sqrt{(1-v)^2 - 1}\, dy & \textit{for } x \in [x' - \tau, x'], \end{cases}$$

then

$$\lim_{m\to+\infty} \frac{\overline{H}_{d_m}(\mu)-1}{d_m} = -\sqrt{-v''(x')}.$$

Lemma 4.4. *Suppose that $\tilde{w}$ is a periodic viscosity solution to the inviscid equation*

$$\sqrt{1+(\mu+\tilde{w}')^2}+v=1 \quad \textit{in } \mathbf{R},$$

then x_0 is a turning point of $\tilde{u}(x)=\mu x+\tilde{w}$ if and only if $\tilde{u}(x)$ attains local minimum at x_0.

Proof. "⇒" is obvious. We only need to show that any local minimum point x_0 must be a turning point. By the definition of viscosity solutions,

$$1+v(x_0)\geq 1,$$

so $v(x_0)=0$ and $x_0\in\mathcal{M}_0$. Choose $\tau>0$ such that $(x_0,x_0+\tau)\cap\mathcal{M}_0=\emptyset$ and $\tilde{u}'(x_0+\tau)=p+\tilde{w}'(x_0+\tau)>0$, then

$$\tilde{u}'(y)>0 \quad \text{for any } y\in(x_0,x_0+\tau) \text{ where } \tilde{u}' \text{ exists.}$$

Otherwise there will be a local mimimum point in $(x_0,x_0+\tau)$. Since any local minimum point belongs to $\mathcal{M}_0$, that contradicts to the choice of τ. Thus,

$$\tilde{u}'=\sqrt{(1-v)^2-1} \quad \text{in } (x_0,x_0+\tau).$$

Similarly,

$$\tilde{u}'=-\sqrt{(1-v)^2-1} \quad \text{in } (x_0-\tau',x_0),$$

for some $\tau' > 0$. □

Proof of Theorem 4.3. We first show that

$$\liminf_{d\to 0^+} \frac{\overline{H}_d(\mu) - 1}{d} \geq -\sqrt{-v''(\bar{x})}. \tag{4.15}$$

In fact, let $h(x)$ be a smooth periodic function such that $h(\bar{x}) = 0$ and $h(x) > 0$ for $x \notin \bar{x} + \mathbf{Z}$. For $\epsilon > 0$, let

$$v_\epsilon(x) = v(x) - \epsilon h(x)$$

and $\overline{H}_{d,\epsilon}(p)$ from the cell problem (4.5) with $\gamma = 1$ and v replaced by v_ϵ. Clearly,

$$\overline{H}_d(\mu) \geq \overline{H}_{d,\epsilon}(\mu).$$

Choose ϵ small enough such that

$$|\mu| < \int_0^1 \sqrt{(1 - v_\epsilon)^2 - 1}\, dx,$$

then $\max\limits_{\mathbf{R}} v_\epsilon = 0$ and the maximum is only obtained at $\bar{x} + \mathbf{Z}$. (4.15) follows immediately from Lemma 4.3.

Suppose $\tilde{u} = \mu x + \tilde{w}$ is the limit of a subsequence of $\mu x + w$ as $d \to 0$. Combining with the above Remark 4.1 and assumption (4.12), (4.15) implies that $\tilde{u}$ can only have a turning point at $\bar{x}$. By Lemma 4.4, $\tilde{u}$ does not have local minimum points in $(\bar{x}, \bar{x} + 1)$. Since $|\mu| < \int_0^1 \sqrt{(1 - v)^2 - 1}\, dx$, there exists a unique $x_\mu \in (\bar{x}, \bar{x} + 1)$ such that $\tilde{u}$ is increasing in $(\bar{x}, x_\mu)$ and is decreasing in $(x_\mu, \bar{x} + 1)$. Hence $\tilde{w}$ is uniquely given by the formula (4.13). □

Chapter 5

Adaptive Basis Learning

5.1 Orthogonal Adaptive Basis Learning

5.1.1 Snapshots of periodic solutions and SVD

Recall the truncated ODE system (2.5) in Chapter 2,

$$\frac{dw_k^N}{dt} + D_0 \left|k\right|^2 w_k^N + i \sum_{\|k-j\| \leq N} \left[\left(k_1 - j_1\right) v_j \left(t\right) + \left(k_2 - j_2\right) \tilde{v}_j \left(t\right) \right] w_{k-j}^N = -v_k \left(t\right), \tag{5.1}$$

and its matrix form

$$\frac{d\mathbf{w}}{dt} = A\left(t\right)\mathbf{w} + \mathbf{v}\left(t\right). \tag{5.2}$$

For some fixed N, let $\{\hat{\mathbf{w}}_n^*\}_{n=0}^{N_t}$ be a numerical periodic solution to (5.1) for some D_0^*. Define the solution matrix

$$W = \begin{bmatrix} \hat{\mathbf{w}}_0^* & \hat{\mathbf{w}}_1^* & \dots & \hat{\mathbf{w}}_{N_t}^* \end{bmatrix}, \tag{5.3}$$

and apply singular value decomposition to W,

$$W = U\Sigma V.$$

Consider SVD of numerical solutions for the time periodic cellular flow

$$\begin{aligned} v\left(x,t\right) &= \cos\left(x_2\right) + \sin\left(x_2\right)\cos\left(t\right), \\ \tilde{v}\left(x,t\right) &= \cos\left(x_1\right) + \sin\left(x_1\right)\cos\left(t\right). \end{aligned} \tag{5.4}$$

Snapshots of numerical solutions to (2.3)

$$w_t + \left(\boldsymbol{v}\cdot\boldsymbol{\partial}\right)w - D_0\partial^2 w = -v.$$

at $D_0^* = 10^{-3}, 10^{-4}$ are shown in Figure 5.1-5.2 where we see thinner layered structures arise as D_0 becomes smaller. Singular values of W for several D_0's are plotted in Figure 5.3 which shows rapid decay beyond 250 out of 2500 modes, uniformly as $D_0 \downarrow 0$.

5.1.2 ODEs from adaptive basis and Poincaré map

Denote by $\mathbf{u}_j$ the jth column of U. For $m > 0$, the adaptive orthogonal basis consists of the columns of the matrix:

$$U_m = \left[\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_m\right].$$

Figure 5.4-5.5 are visualizations of $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_5$, $\mathbf{u}_6$ for the flow (5.4).

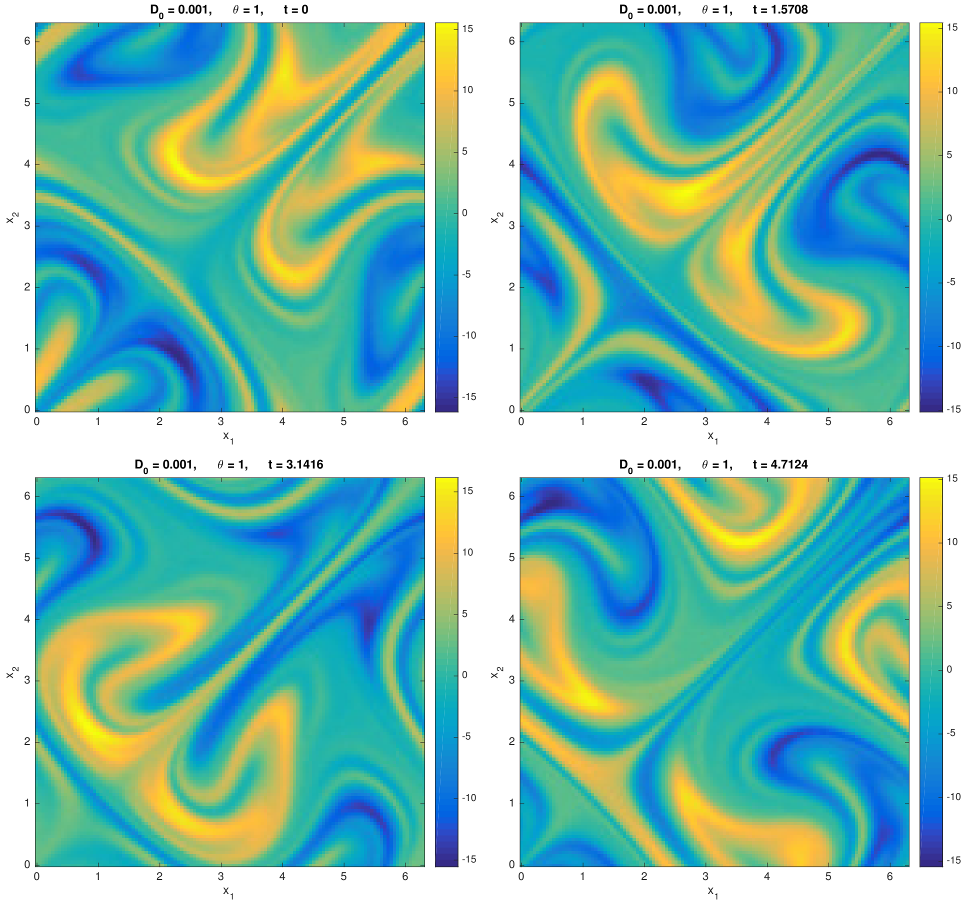


Figure 5.1: Sampled snapshots of solution to (2.3) with $D_0^* = 10^{-3}$, appearance of layered structures.

Given $D_0 > 0$, let us write the solution to (5.2) in the orthogonal adaptive basis as:

$$\mathbf{w}\left(t\right) = U_m \mathbf{a}\left(t\right),$$

where $\mathbf{a}\left(t\right) = \left[a_1\left(t\right), a_2\left(t\right), \ldots, a_m\left(t\right)\right]^T$ is periodic, then

$$\frac{d\mathbf{a}}{dt} = \bar{U}_m^T A\left(t\right) U_m \mathbf{a} + \bar{U}_m^T \mathbf{v}\left(t\right). \tag{5.5}$$

Hence an approximation of solution to (5.2) can be obtained by solving (5.5).

Similar to the approach used for the Fourier modes, let us define the Poincaré map associated

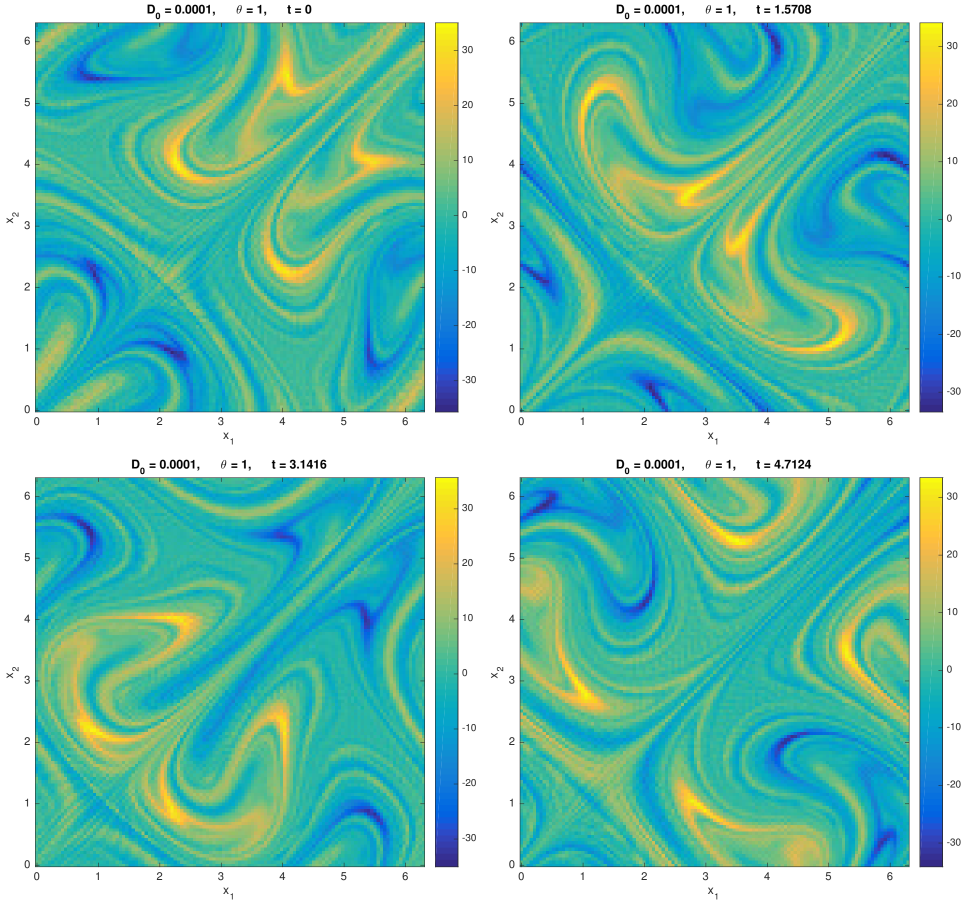


Figure 5.2: Sampled snapshots of solution to (2.3) with $D_0^* = 10^{-4}$, formation of thin layers.

to the ODE (5.5)

$$P\left(x\right) = Mx + b, \quad x \in \mathbf{R}^m,$$

where M is an $m \times m$ matrix and b is an $m \times 1$ vector. Denote the RK4 operator by $\mathcal{L}$, the algorithm is shown below.

Reorder $\hat{\mathbf{w}}_n^N$ as Fourier modes $\left\{\hat{w}_{k,n}^N\right\}_{\|\boldsymbol{k}\| \leq N}$, then $D_{11,N}^E$ is estimated by

$$\hat{D}_{11,N}^{E,a} = D_0 \left(1 + \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{\|k\| \leq N} \left|\boldsymbol{k}\right|^2 \left|\hat{w}_{k,n}^N\right|^2 \right).$$

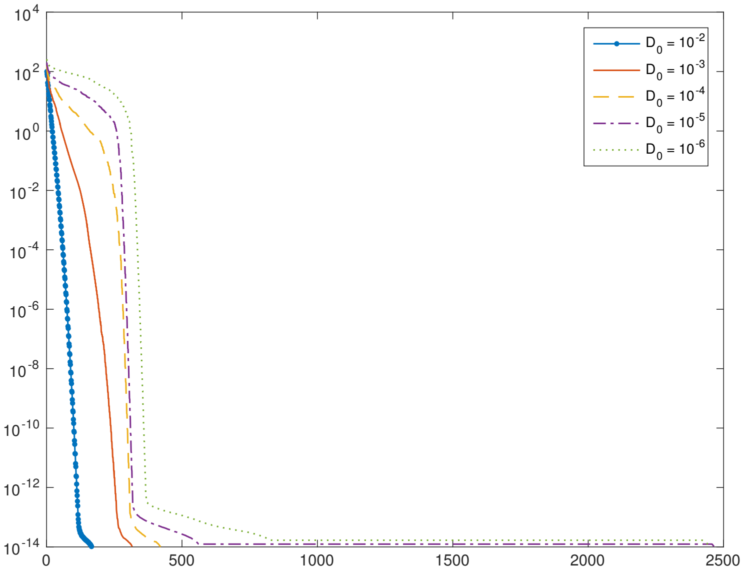


Figure 5.3: Singular values of numerical periodic solution matrices, rapid decay uniformly in $D_0 \downarrow 0$.

5.1.3 Experimental results of orthogonal adaptive basis

We show computational results on residual diffusion from the orthogonal adaptive basis on time periodic cellular flows. The main goal is to maintain enough accuracy at low costs.

Two-dimensional time-dependent flow (5.4)

- $D_0^* = 10^{-3}, N = 60, N_t = 1500$ with $m = 100$ (the number of adaptive basis functions).

In Table 5.1, $D_{11,N}^E$ from the Fourier basis (with $N_t = 2000$) for flow (2.10) at varied D_0's are shown along with those from the orthogonal adaptive basis, denoted by $\hat{D}_{11,N}^{E,a}$. To measure the reduction in the number of basis functions, we define $r = m/\left(2N+1\right)^2$ as the ratio of the number of adaptive basis functions and that of the Fourier basis functions. The estimates by adaptive basis are close to those from the Fourier basis when D_0 is not far from $D_0^* = 10^{-3}$ (the D_0 value where the adaptive basis is constructed or trained). The robustness of adaptive

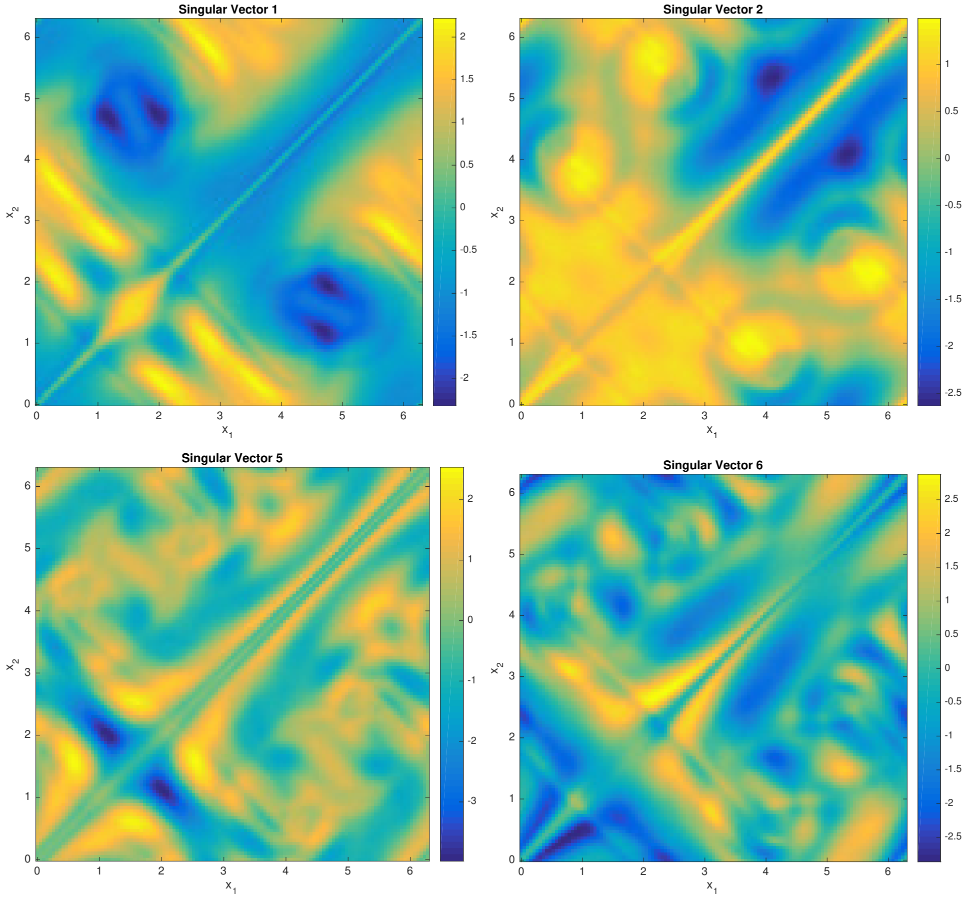


Figure 5.4: Sampled singular vectors with $D_0^* = 10^{-3}, N = 60, N_t = 1500$.

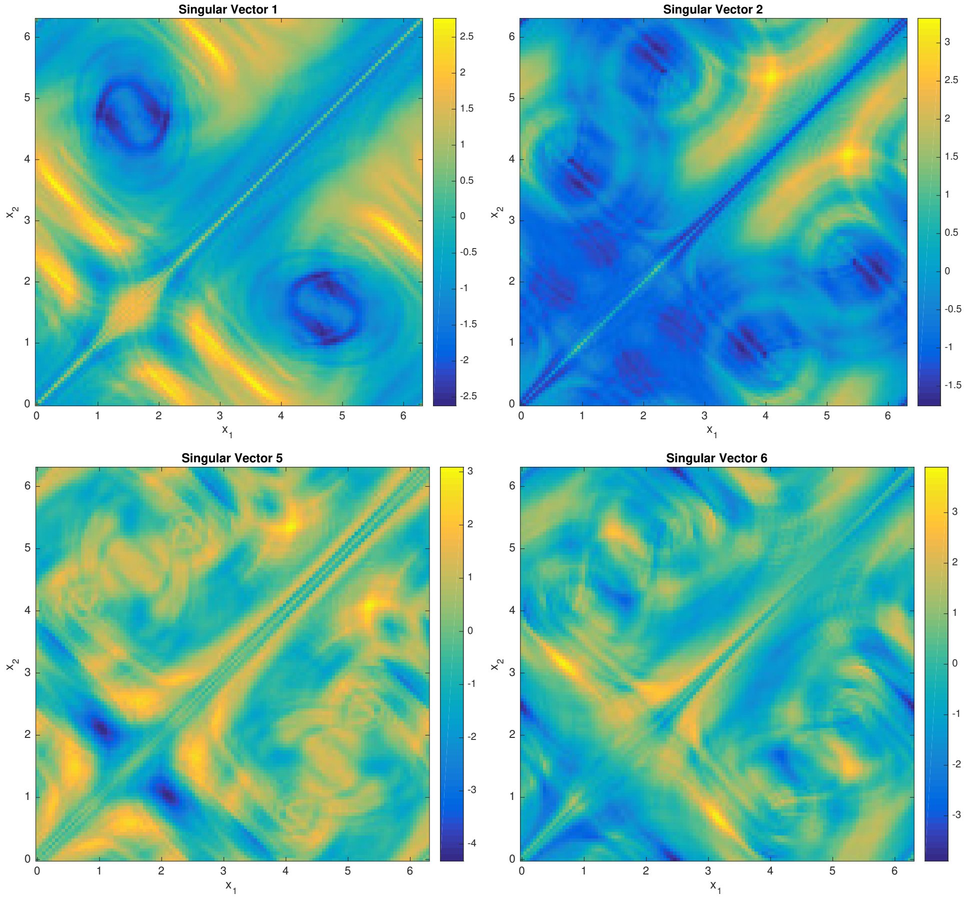
basis hinges on how fast the error grows as the testing occurs at a D_0 value deviating from the training value D_0^*.

The energy of a truncated Fourier expansion is:

$$\mathcal{E}\left(\sum_{\|k\|\leq N} z_k^N\left(t\right)e^{ik\cdot x}\right) = D_0 \sum_{\|k\|\leq N} |k|^2 \left\langle z_k^N \overline{z}_k^N \right\rangle .$$

Let $\left\{\hat{z}_{\boldsymbol{k},n}^N\right\}_{n=1}^{N_t+1}$ be the numerical approximation of $z_{\boldsymbol{k}}^N\left(t\right)$, then the energy for $\sum\limits_{\|\boldsymbol{k}\|\leq N} z_{\boldsymbol{k}}^N\left(t\right)e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$ can be approximated by

$$\mathcal{E}\left(\sum_{\|k\|\leq N} z_k^N\left(t\right)e^{ik\cdot x}\right) \approx \frac{D_0}{N_t}\sum_{n=1}^{N_t}\sum_{\|k\|\leq N} |k|^2 \left|\hat{z}_{k,n}^N\right|^2 .$$


Singular Vector 1

Singular Vector 2

Singular Vector 5

Singular Vector 6

Figure 5.5: Sampled singular vectors with $D_0^* = 10^{-4}, N = 60, N_t = 2000$.

Figure 5.6 shows the energy vs. the number of modes in the solutions solved by Fourier basis (solid, blue) and the learned orthogonal adaptive basis (dashdot, red). Clearly, a much smaller number of basis functions is needed to represent the same level of energy by the adaptive basis than by the Fourier basis.

- $D_0^* = 10^{-4}, N = 60, N_t = 2000$ with $m = 200$

Computations of $D_{11,N}^E$ and $\hat{D}_{11,N}^{E,a}$ for the flow (2.10) at smaller D_0's are shown in Table 5.2. The comparisons of energy growth vs. the number of adaptive (dashdot, red) and Fourier (solid, blue) basis functions are shown in Figure 5.7. Here $N_t = 2500$ in computation of $\hat{D}_{11,N}^E$ with the Fourier basis. Interestingly, the relative errors of solutions via the orthogonal

Algorithm 2 Solving cell problem with SVD.

```
1. Set $\hat{X}_0 = [e_1 \quad e_2 \quad \dots \quad e_m]$ to be the matrix of standard basis of $\mathbf{R}^m$.
**for** n = 0, 1,..., $N_t - 1$ **do**
    $\hat{X}_{n+1} = \mathcal{L}\left(\bar{U}_m^T A^N U_m, \mathbf{0}; \hat{X}_n, t_n\right)$
**end for**
$\hat{M} = \hat{X}_{N_t}$.
2. Set $\hat{\mathbf{x}}_0 = \mathbf{0}$.
**for** n = 0, 1,..., $N_t - 1$ **do**
    $\hat{\mathbf{x}}_{n+1} = \mathcal{L}\left(\bar{U}_m^T A^N U_m, \bar{U}_m^T \mathbf{v}^N; \hat{\mathbf{x}}_n, t_n\right)$
**end for**
$\hat{b} = \hat{\mathbf{x}}_{N_t}$.
3. Solve $\hat{x} = \hat{M}\hat{x} + \hat{b}$.
Set $\hat{\mathbf{a}}_0 = \hat{x}$.
**for** n = 0, 1,..., $N_t - 1$ **do**
    $\hat{\mathbf{a}}_{n+1} = \mathcal{L}\left(A, \mathbf{v}; \hat{\mathbf{a}}_n, t_n\right)$
**end for**
4. $\hat{\mathbf{w}}_n^N = U_m \hat{\mathbf{a}}_n^N$ for $n = 0, 1, \dots, N_t$.
```

D_0	10^{-3}	9×10^{-4}	8×10^{-4}	7×10^{-4}	6×10^{-4}	5×10^{-4}	4×10^{-4}	3×10^{-4}	2×10^{-4}	10^{-4}
$\hat{D}_{11,N}^{E,a}$	1.3772	1.4050	1.4337	1.4632	1.4931	1.5229	1.5515	1.5775	1.6047	1.7191
$\hat{D}_{11,N}^{E}$	1.3772	1.3765	1.3763	1.3772	1.3796	1.3847	1.3940	1.4105	1.4395	1.4951
relative error	**0**	**2.1%**	**4.2%**	**6.3%**	**8.2%**	**10.0%**	**11.3%**	**11.8%**	**11.5%**	**15.0%**

Table 5.1: $\hat{D}_{11,N}^{E,a}$ and $\hat{D}_{11,N}^{E}$ for flow (5.4) with $D_0^* = 10^{-3}$, $r = 0.68\%$.

adaptive basis drop considerably at smaller D_0, suggesting that the basis learning is effective for computing residual diffusion.

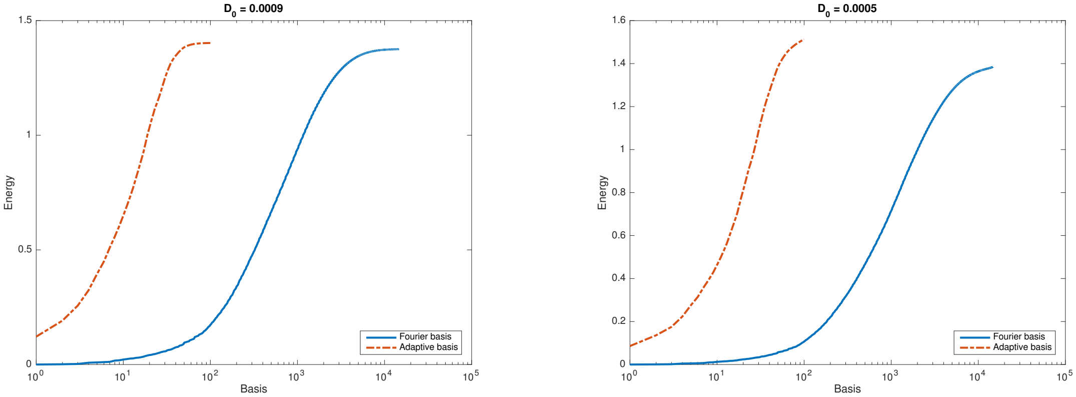


Figure 5.6: Energy growth vs. the number of adaptive and Fourier basis functions for $D_0 = 9 \times 10^{-4}$ and $D_0 = 5 \times 10^{-4}$.

D_0	10^{-4}	9×10^{-5}	8×10^{-5}	7×10^{-5}	6×10^{-5}
$\hat{D}^{E,a}_{11,N}$	1.4951	1.5042	1.5129	1.5208	1.5272
$\hat{D}^{E}_{11,N}$	1.4951	1.5036	1.5131	1.5236	1.5355
relative error	**0**	**0**	**0**	**0.2%**	**0.5%**

D_0	5×10^{-5}	4×10^{-5}	3×10^{-5}	2×10^{-5}	10^{-5}
$\hat{D}^{E,a}_{11,N}$	1.5314	1.5313	1.5242	1.5107	1.5243
$\hat{D}^{E}_{11,N}$	1.5492	1.5649	1.5834	1.6052	1.6301
relative error	**1.1%**	**2.1%**	**3.7%**	**5.9%**	**6.5%**

Table 5.2: $\hat{D}^{E,a}_{11,N}$ and $\hat{D}^{E}_{11,N}$ for flow (5.4) with $D_0^* = 10^{-4}$, $r = 1.37\%$.

Two-dimensional time-dependent flow with θ

Recall the time-dependent flow with $\theta \in (0, 1]$

$$\begin{aligned} v\left(x,t\right) &= \cos\left(x_2\right) + \theta \sin\left(x_2\right) \cos\left(t\right), \\ \tilde{v}\left(x,t\right) &= \cos\left(x_1\right) + \theta \sin\left(x_1\right) \cos\left(t\right). \end{aligned} \tag{5.6}$$

Adaptive orthogonal basis can also be trained from a periodic solution at a θ value and applied to another flow at a nearby θ value. For instance, assemble W in Section 5.1.1 with snapshots of a periodic solution for some D_0^* and flow (5.6) with parameter θ^*. Periodic solutions as well as effective diffusivities at the same D_0^* but different θ's can be approximated

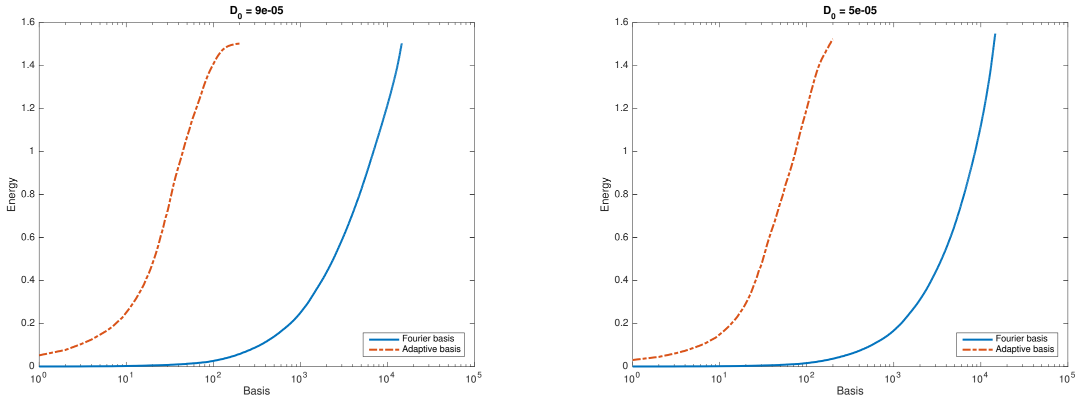


Figure 5.7: Energy growth vs. the number of adaptive and Fourier basis functions for $D_0 = 9 \times 10^{-5}$ and $D_0 = 5 \times 10^{-5}$.

as in Section 5.1.2. However, the dependence of D_{11}^E on θ seems very sensitive especially at small D_0, as the test results indicate below.

In the following experiments, $N = 60$, $m = 100$ and $r = 0.68\%$.

- $\theta^* = 0.7, D_0 = 10^{-3}, 10^{-4}$ with $m = 100$

Estimates of $D_{11,N}^E$ for flow (5.6) with $D_0 = 10^{-3}, 10^{-4}$ and varied θ's by reduced basis trained with $\theta^* = 0.7$, denoted by $\hat{D}_{11,N}^{E,a}$, as well as results from Fourier basis, are presented in Tables 5.3-5.4.

θ	0.7	0.71	0.72	0.73	0.74	0.75
$\hat{D}_{11,N}^{E,a}$	0.2177	0.2138	0.2101	0.2065	0.2029	0.1993
$\hat{D}_{11,N}^{E}$	0.2177	0.2251	0.2368	0.2509	0.2715	0.2978
relative error	**0**	**5.0%**	**10.9%**	**17.7%**	**25.3%**	**33.1%**

Table 5.3: $\hat{D}_{11,N}^{E,a}$ and $\hat{D}_{11,N}^{E}$ for flow (5.6) with $D_0 = 10^{-3}$ and $\theta^* = 0.7$, $N_t = 1500$.

θ	0.7	0.71	0.72	0.73	0.74	0.75
$\hat{D}_{11,N}^{E,a}$	0.1725	0.1636	0.1571	0.1536	0.1518	0.1491
$\hat{D}_{11,N}^{E}$	0.1708	0.1838	0.1957	0.2063	0.2360	0.2849
relative error	**1.0%**	**10.1%**	**19.7%**	**25.5%**	**35.7%**	**47.7%**

Table 5.4: $\hat{D}_{11,N}^{E,a}$ and $\hat{D}_{11,N}^{E}$ for flow (5.6) with $D_0 = 10^{-4}$ and $\theta^* = 0.7$, $N_t = 2000$.

- $\theta^* = 0.4, D_0 = 10^{-3}, 10^{-4}$ with $m = 100$

Estimates of $D^E_{11,N}$ for flow (5.6) with $D_0 = 10^{-3}, 10^{-4}$ and varied θ's by reduced basis trained with $\theta^* = 0.4$ as well as results from Fourier basis are shown in Tables 5.5-5.6.

θ	0.4	0.41	0.42	0.43	0.44	0.45
$\hat{D}^{E,a}_{11,N}$	0.3921	0.3700	0.3523	0.3380	0.3261	0.3161
$\hat{D}^{E}_{11,N}$	0.3921	0.3772	0.3637	0.3528	0.3451	0.3405
relative error	**0**	**2.0%**	**3.1%**	**4.2%**	**5.5%**	**7.2%**

Table 5.5: $\hat{D}^{E,a}_{11,N}$ and $\hat{D}^{E}_{11,N}$ for flow (5.6) with $D_0 = 10^{-3}$ and $\theta^* = 0.4$, $N_t = 1500$.

θ	0.4	0.41	0.42	0.43	0.44	0.45
$\hat{D}^{E,a}_{11,N}$	0.3888	0.3795	0.3788	0.3810	0.3823	0.3792
$\hat{D}^{E}_{11,N}$	0.3887	0.3516	0.3187	0.3027	0.3041	0.3195
relative error	**0**	**8.0%**	**18.9%**	**25.9%**	**25.7%**	**18.7%**

Table 5.6: $\hat{D}^{E,a}_{11,N}$ and $\hat{D}^{E}_{11,N}$ for flow (5.6) with $D_0 = 10^{-4}$ and $\theta^* = 0.4$, $N_t = 2000$.

5.2 Construction of Adaptive Basis via DNNs

5.2.1 Learning thinner structures

Consider the numerical solution to the problem with flow (5.4). It can be seen from Figure 5.1-5.2 that layers in snapshots get thinner as D_0 becomes smaller. In the prediction of the singular solutions at D_0 much smaller than D_0^*, it would be helpful to have the adaptive basis learned at D_0^* demonstrate the thinner layered structures. Particularly, given the solution matrix W at D_0^* as (5.3), we are looking for a transform $\mathcal{T}$ such that snapshots of Fourier coefficients $\mathcal{T}(W)$ have sharpened layers.

Suppose $D_0^1 > D_0^2$ and W^i is the solution matrix at D_0^i for $i = 1, 2$. Let $\mathcal{F}$ be column-wise Fourier transform on matrices, then columns of $\mathcal{F}^{-1}(W^i)$ are snapshots of solution W^i.

When W^i's are known, we may use certain DNN to train a map $\mathcal{T}$ for the following regression problem

$$\mathcal{T}: \quad \mathcal{F}^{-1}\left(W^1\right) \to \mathcal{F}^{-1}\left(W^2\right).$$

$D_0^2 < D_0^1$ implies $\mathcal{F}^{-1}(W^2)$ has thinner layered structures than $\mathcal{F}^{-1}(W^1)$ and so does $\mathcal{T}(\mathcal{F}^{-1}(W^1))$. Hence $\mathcal{T}$ can be applied to solution matrix W at some D_0^* and $\mathcal{T}(\mathcal{F}^{-1}(W))$ is expected to have thinner structures. Thus the adaptive basis with thinner structures will be obtained from SVD of

$$\mathcal{F}\left(\mathcal{T}\left(\mathcal{F}^{-1}(W)\right)\right).$$

5.2.2 Adversarial network

We apply the super-resolution generative adversarial network (SRGAN) [37] to the construction of map $\mathcal{T}$. As a generative adversarial network (GAN), SRGAN consists of a generator network G and a discriminator network D. The two networks are competing in a way that D is trained to distinguish the real high-resolved images and images generated from low-resolved images, while G is trained to create fake high-resolved images from low-resolved images to fool D.

We train the SRGAN with $\mathcal{F}^{-1}(W^1)$ as input data and $\mathcal{F}^{-1}(W^2)$ as target data so that the generator G can learn to generate thinner structures when it is fed with $\mathcal{F}^{-1}(W)$. With that approach we may implement $\mathcal{T}$ by G.

Network architecture

As shown in Figure 5.8, the generator network G starts with a convolutional block with kernel size 9×9, followed by a few residual blocks. Here a convolutional block consists of a convolutional layer and a PReLU layer, a residual block is a convolutional block with kernel size 3×3 followed by a convolutional layer of the same kernel size and a shortcut from the input to output. There are two more convolutional layers with kernel size 3×3 and 9×9 after the residual blocks at the end of the network. The number of filters in all convolutional blocks are the same except for the last one. Note that we remove the two upscale layers in [37] since the sizes of slides of $\mathcal{F}^{-1}\left(W^1\right)$ and $\mathcal{F}^{-1}\left(W^2\right)$ are the same.

The discriminator network D is defined by the architectural guidelines summarized in [52], see Figure 5.8. It has eight convolutional blocks with PReLU layers replaced by LeakyReLU layers with $\alpha = 0.2$. Moreover, there is a batch normalization layer before each LeakyReLU in the convolutional blocks. The kernel size is 3×3 in all convolutional blocks and the number of filters is doubled in the 3rd, 5th and 7th block. Those blocks are followed by a fully connected layer, a LeakyReLU layer and one more fully connected layer. Finally the feature map is fed in a sigmoid layer which gives the probability of real high-resolved image and reconstructed one.

Loss function of generator network

As a binary classifier, the discriminator network is equipped with the cross entropy loss. We are focusing on the loss function of the generator network.

Suppose $\mathcal{F}^{-1}\left(W^1\right)$ and $\mathcal{F}^{-1}\left(W^2\right)$ are real matrices of dimension $(2N+1)^2 \times N_t$, columns of $\mathcal{F}^{-1}\left(W^1\right)$ and $\mathcal{F}^{-1}\left(W^2\right)$ are x_i and y_i, $i = 1, 2, \ldots, N_t$. Following the formulation in [37],

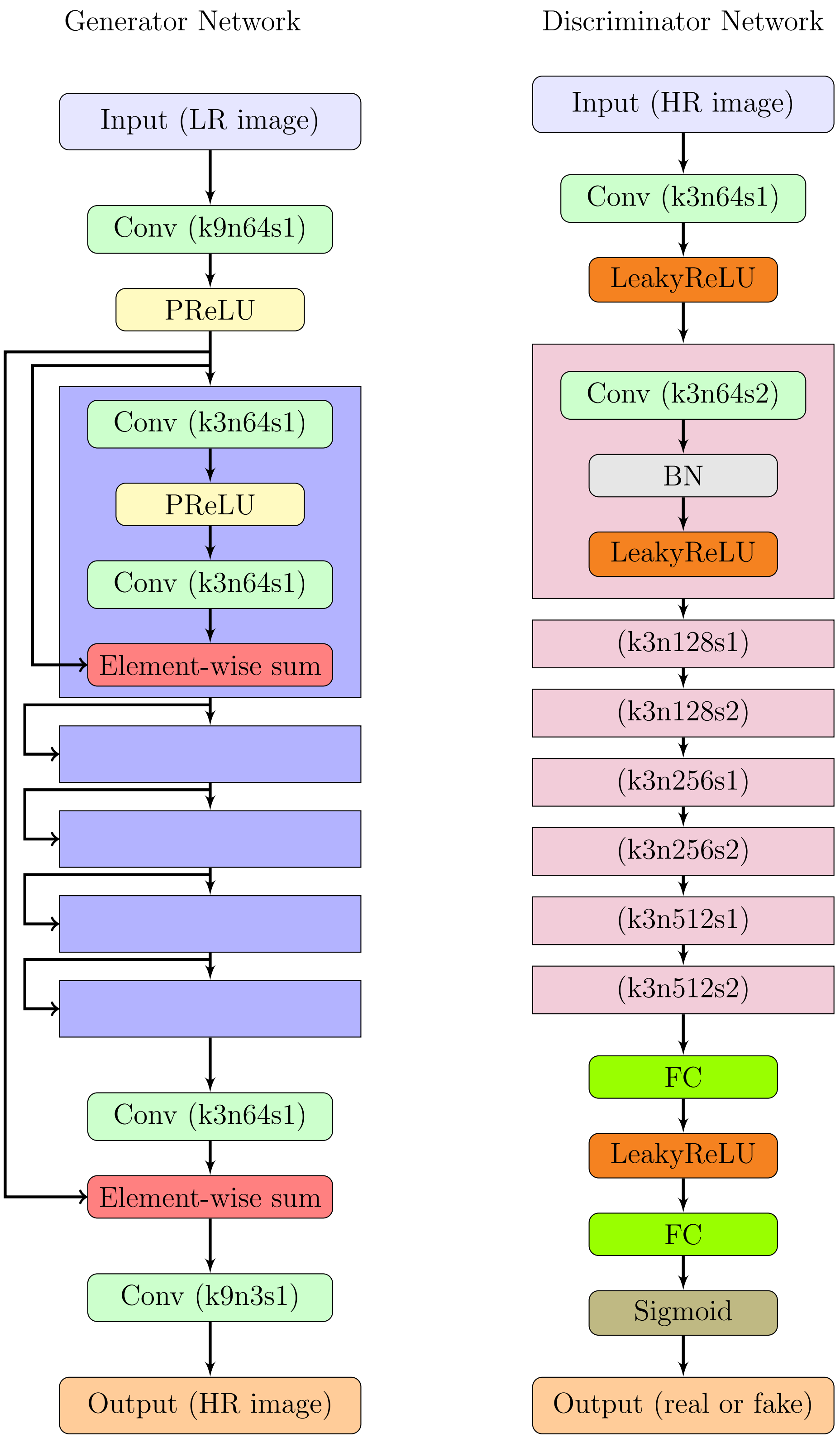


Figure 5.8: Architecture of the generator (left) and discriminator (right) network.

we define the loss function of the generator network as

$$l\left(G\right) = l_{MSE}\left(G\right) + 10^{-2} l_{VGG}\left(G\right) + 10^{-3} l_{Gen}\left(G\right). \tag{5.7}$$

In (5.7), l_{MSE} is the pixel-wise **MSE loss** defined as the sum of the squares of error at each pixel,

$$l_{MSE}\left(G\right) = \sum_{i=1}^{N_t} \left\| y_i - G\left(x_i\right) \right\|_2^2.$$

l_{VGG} is the **VGG loss** based on layers of the pre-trained VGG-19 network [65]. Let ϕ be a feature map of VGG-19 and s_ϕ be its size, then the VGG loss is the average of squares of Euclidean distances between the feature representations of y_i and $G\left(x_i\right)$

$$l_{VGG}\left(G\right) = \frac{1}{s_\phi} \sum_{i=1}^{N_t} \left\| \phi\left(y_i\right) - \phi\left(G\left(x_i\right)\right) \right\|_2^2.$$

The generator network is expected to fool the discriminator network, so (5.7) contains l_{Gen} called **generative loss**. l_{Gen} is defined based on the cross-entropy loss of the discriminator network

$$\sum_{i=1}^{N_t} \log\left[1 - D\left(G\left(x_i\right)\right)\right]. \tag{5.8}$$

Here $D\left(G\left(x_i\right)\right)$ means the binary classification result of the reconstructed high-resolved image by the generator network G in (5.8). In practice, we define

$$l_{Gen}\left(G\right) = \sum_{i=1}^{N_t} -\log D\left(G\left(x_i\right)\right),$$

for better gradient behavior.

5.2.3 Experimental results of adaptive basis from SRGAN

Let $D_0^1 = 10^{-2}$, $D_0^2 = 10^{-3}$. We solved for both W^1 and W^2 with $N = 50$ and $N_t = 1500$, then train SRGAN with input data $\mathcal{F}^{-1}(W^1)$ and target data $\mathcal{F}^{-1}(W^2)$. We set $D_0^* = D_0^2$ in the following experiments.

Figure 5.9 shows a time slide of the input $\mathcal{F}^{-1}(W^1)$ (top left), target $\mathcal{F}^{-1}(W^2)$ (top right), output $G(\mathcal{F}^{-1}(W^1))$ (bottom left) and $G(\mathcal{F}^{-1}(W^2))$ (bottom right). It can be seen that thinner layers are created by the network G.

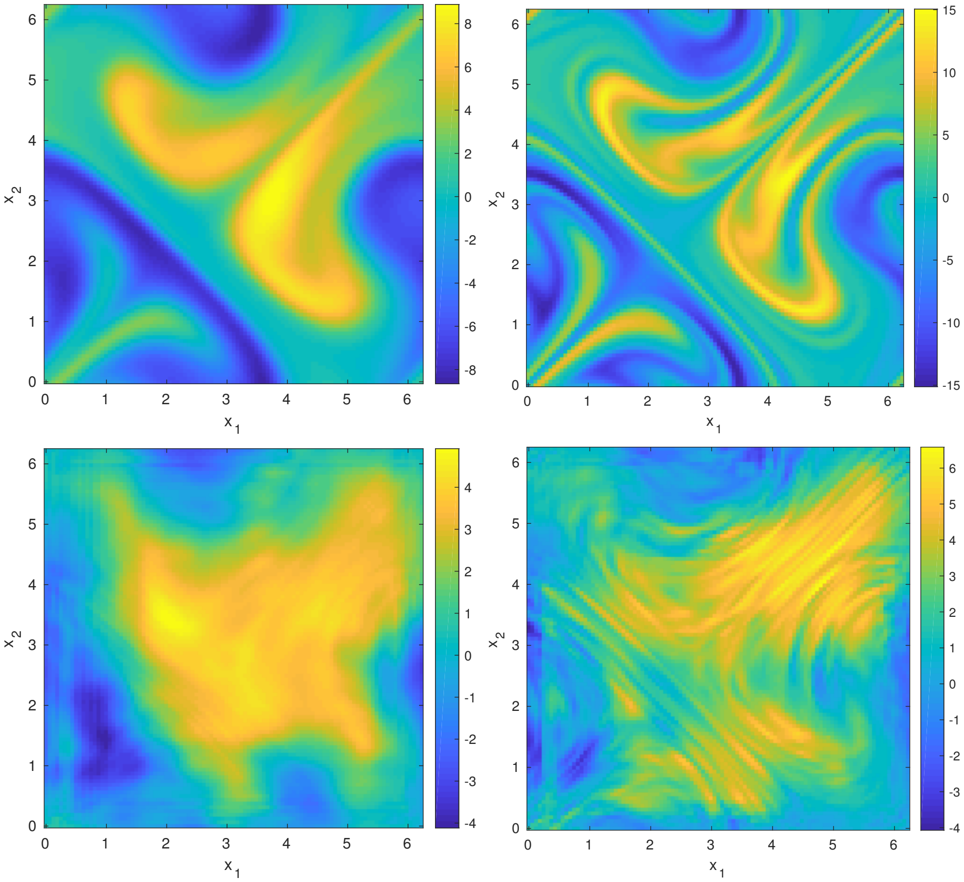


Figure 5.9: Input and Output of the SRGAN with $D_0^1 = 10^{-2}$, $D_0^2 = 10^{-3}$.

Set $D_0^1 = 10^{-2}$ and $D_0^2 = 10^{-3}$, the comparison of predictions of D_{11}^E by direct SVD and SRGAN assisted SVD is shown in Table 5.7. The number of adaptive basis used is $m = 100$ for both method.

D_0		5×10^{-4}	4×10^{-4}	3×10^{-4}	2×10^{-4}	10^{-4}
$\hat{D}^{E}_{11,60}$		1.3847	1.3940	1.4105	1.4395	1.4951
$\hat{D}^{E,a}_{11,50}$	SVD	1.5258	1.5597	1.5969	1.6381	1.6854
	SRGAN	1.2429	1.2663	1.3056	1.3786	1.5293
relative error	SVD	10.2%	11.9%	13.2%	13.8%	12.7%
	SRGAN	10.2%	**9.2%**	**7.4%**	**4.2%**	**2.3%**

Table 5.7: Comparison of $\hat{D}^{E,a}_{11,N}$ for flow (5.4) with $D^1_0 = 10^{-2}$, $D^2_0 = 10^{-3}$.

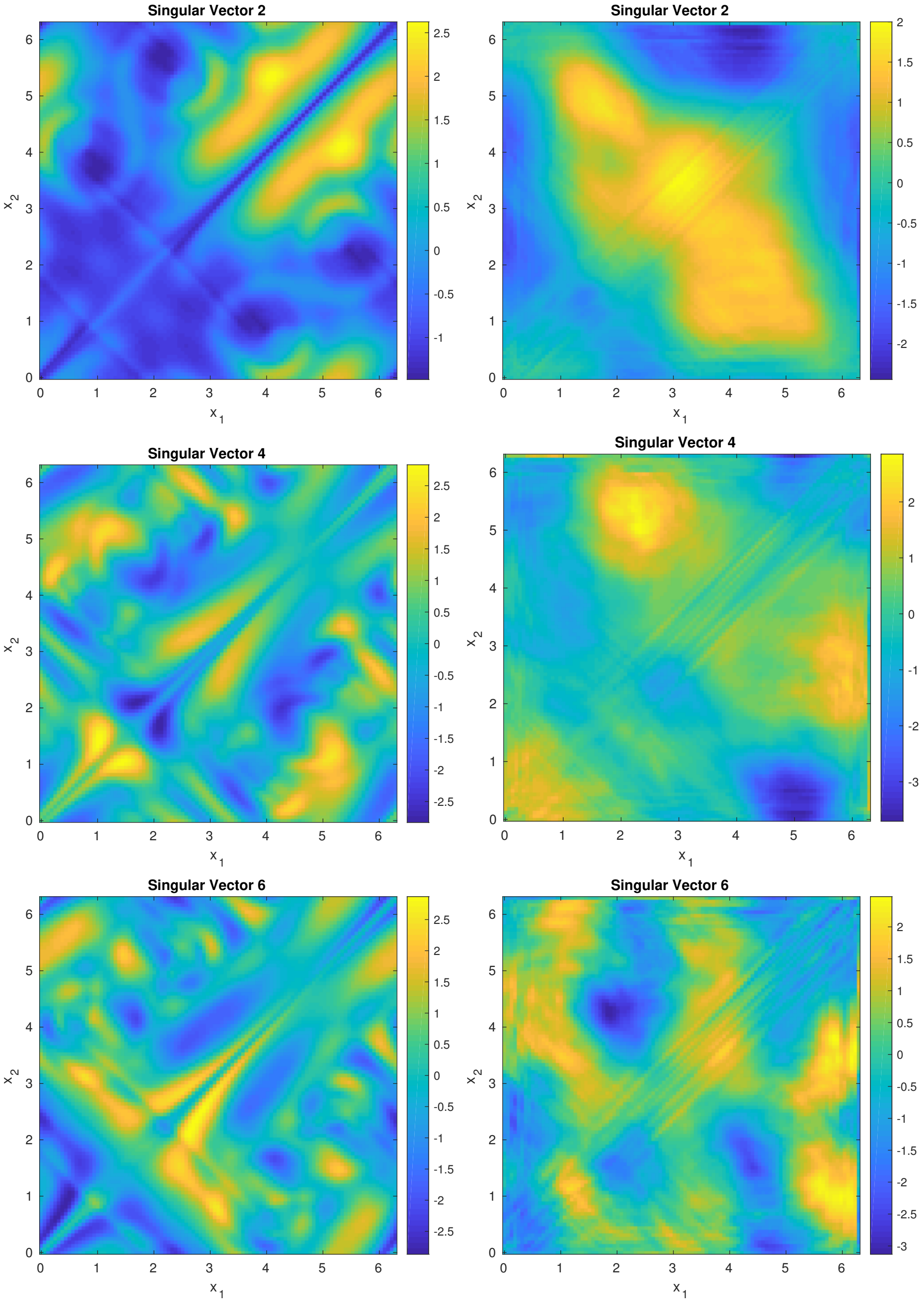
When D^1_0 is closer to D^2_0, SRGAN assisted SVD may have even better predictions at smaller D_0. In Table 5.8, $D^1_0 = 5 \times 10^{-3}$, $D^2_0 = 10^{-3}$ and $N = 50$ and we predict the $\hat{D}^{E}_{11,60}$ for $D_0 = 3 \times 10^{-4}$, 2×10^{-4} and 10^{-4}. For $D^1_0 = 5 \times 10^{-3}$, $D^2_0 = 10^{-3}$, singular vectors of $\mathcal{F}\left(G\left(\mathcal{F}^{-1}\left(W^2\right)\right)\right)$ also have thinner structures than that of W^2, as shown in right column and left column of Figure 5.10 respectively. Table 5.9 summarizes predictions for $D_0 = 2 \times 10^{-5}$ and 10^{-5} from $D^1_0 = 10^{-3}$, $D^2_0 = 10^{-4}$ and $N = 60$.

D_0		3×10^{-4}	2×10^{-4}	10^{-4}
$\hat{D}^{E}_{11,60}$		1.4105	1.4395	1.4951
$\hat{D}^{E,a}_{11,50}$	SVD	1.5969	1.6381	1.6854
	SRGAN	1.3111	1.3862	1.5015
relative error	SVD	13.2%	13.8%	12.7%
	SRGAN	**7.0%**	**3.7%**	**0.4%**

Table 5.8: Comparison of $\hat{D}^{E,a}_{11,N}$ for flow (5.4) with $D^1_0 = 5 \times 10^{-3}$, $D^2_0 = 10^{-3}$.

D_0		2×10^{-5}	10^{-5}
$\hat{D}^{E}_{11,60}$		1.6052	1.6301
$\hat{D}^{E,a}_{11,60}$	SVD	1.5107	1.5243
	SRGAN	1.6234	1.7120
relative error	SVD	5.9%	6.5%
	SRGAN	**1.1%**	**5.0%**

Table 5.9: Comparison of $\hat{D}^{E,a}_{11,N}$ for flow (5.4) with $D^1_0 = 10^{-3}$, $D^2_0 = 10^{-4}$.



Singular Vector 2

Singular Vector 2

Singular Vector 4

Singular Vector 4

Singular Vector 6

Singular Vector 6

Figure 5.10: Singular vectors of solution matrix W^2 (left column) and $\mathcal{F}\left(G\left(\mathcal{F}^{-1}\left(W^2\right)\right)\right)$ (right column).

Chapter 6

Quantization of Deep Neural Networks

6.1 BinaryRelax

The training of DNNs with quantized weights can be written as a constrained optimization problem:

$$\min_{x\in\mathbf{R}^n} \; f(x) := \frac{1}{N}\sum_{j=1}^{N} \ell_j(x) \quad \text{subject to} \quad x \in \mathcal{Q}. \tag{6.1}$$

Without loss of generality, we assume the set of quantized weights

$$\mathcal{Q} = \mathbf{R}_+ \times \{\pm q_1, \dots, \pm q_m\}^n$$

throughout the chapter. Hence we only consider the case for simplicity that a single adjustable scaling factor is shared by all weights in the network.

6.1.1 Quantization

For general b-bit quantization,

$$\mathcal{Q} = \bigcup_{i=1}^{p} \mathcal{L}_i$$

is the union of p distinct one-dimensional subspaces $\mathcal{L}_i \subset \mathbf{R}^n$, $i = 1, 2, \ldots, p$, where

$$\mathcal{L}_i = \{s \cdot L_i : s \in \mathbf{R}\}$$

for some $L_i \in \{\pm q_1, \ldots, \pm q_m\}^n \setminus \{\mathbf{0}\} \subset \mathbf{R}^n$ [38]. Figure 6.1 shows an example of $\mathcal{Q} = \mathbf{R}_+ \times \{0, \pm 1\}^2$, i.e. the ternarization of two weights.

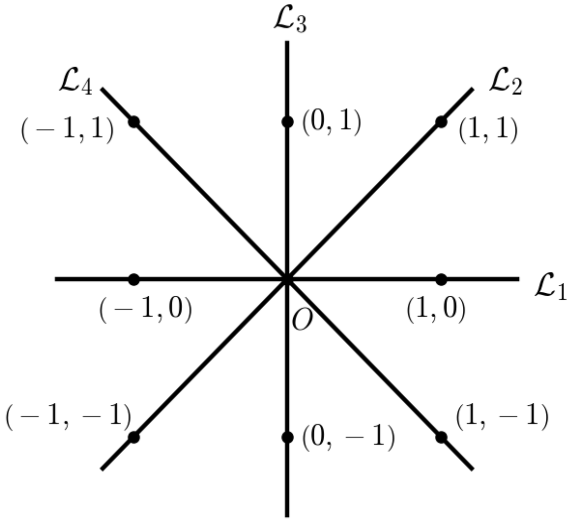


Figure 6.1: Graphic illustration of $\mathcal{Q} = \mathbf{R}_+ \times \{0, \pm 1\}^2$. In this case, $b = n = 2$, $p = 4$.

Given a float weight vector y, its quantization x is the projection of y onto the set $\mathcal{Q}$

$$x = \arg\min_{z \in \mathcal{Q}} \|z - y\|^2 = \text{proj}_{\mathcal{Q}}(y). \tag{6.2}$$

$\mathcal{Q}$ is a non-convex set, so the projection may not be unique. In that case, we just assume x

is one of them. The projection/quantization problem can be reformulated as

$$(s^*,\, Q^*) = \arg\min_{s,Q} \; \|s \cdot Q - y\|^2 \quad \text{subject to} \quad Q \in \{\pm q_1, \dots, \pm q_m\}^n, \tag{6.3}$$

The quantization of y is then given by $\mathrm{proj}_{\mathcal{Q}}(y) = s^* \cdot Q^*$. (6.3) is essentially a constrained K-means clustering problem of 1-D points. The centroids are of the form $\pm(s \cdot q_j)$ with $1 \leq j \leq m$, and they are determined by a single parameter s since q_j's are fixed. For uniform quantization where $q_j = j - 1$, these centroids are equi-spaced. Given s, the assignment of float weights is then governed by Q. Thus the problem (6.3) can be solved by a variant of Lloyd's algorithm [42], which iterates between the assignment step (Q-update) and centroid update step (s-update). In the Q-update of the l-th iteration, fixing the scaling factor s^{l-1}, each Q^l_i is chosen from $\{\pm q_1, \dots, \pm q_m\}$ so that $s^{l-1}Q^l_i$ is the nearest centroid to y_i. In the s-update, a quadratic problem

$$\min_{s \in \mathbf{R}} \; \|s \cdot Q^l - y\|^2$$

is solved by $s^l = \frac{\langle Q^l, y\rangle}{\|Q^l\|^2}$.

The above procedure however, is impractical here, as the quantization is needed in every iteration of training. It has been shown that the closed form (exact) solution of (6.3) can be computed at $\mathcal{O}(n)$ complexity for binarization [59] where $Q \in \{\pm 1\}^n$:

$$s^* = \frac{\|y\|_1}{n}, \; Q^*_i = \begin{cases} 1 & \text{if } y_i \geq 0 \\ -1 & \text{otherwise.} \end{cases} \tag{6.4}$$

In the case of ternarization where $Q \in \{0, \pm 1\}^n$, an $O(n \log n)$ exact formula [76] is

$$t^* = \arg\max_{1 \leq t \leq n} \frac{\|y_{[t]}\|_1^2}{t}, \; s^* = \frac{\|y_{[t^*]}\|_1}{t^*}, \; Q^* = \mathrm{sign}(y_{[t^*]}), \tag{6.5}$$

where $y_{[t]} \in \mathbf{R}^n$ keeps the t largest component in magnitude of y, while zeroing out the others. For quantization with wider bit-width ($b > 2$), accurately solving (6.3) becomes computationally intractable [76]. Empirical formulas have thus been proposed for an approximate quantized solution [39, 76, 80], and they turn out to be sufficient for practical use. For example, a thresholding scheme of $\mathcal{O}(n)$ complexity for ternarization [39] is

$$\delta = \frac{0.7\|y\|_1}{n},\ s^* = \frac{\sum_{i=1}^n |y_i| \cdot 1_{|y_i|\geq\delta}}{\sum_{i=1}^n 1_{|y_i|\geq\delta}},\ Q_i^* = \begin{cases} \text{sign}(y_i) & \text{if } |y_i| \geq \delta, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

For $b > 2$, Yin et al. [76] proposed to just perform one iteration of Lloyd's algorithm with a carefully initialized Q.

In this work, we assume that the quantization $\text{proj}_{\mathcal{Q}}(y)$ can be computed precisely, regardless the choice of q_j's.

6.1.2 Relaxed Quantization

Moreau [50] introduced what is now called the Moreau envelope and the proximity operator (proximal mapping) that generalizes the projection. Let $g : \mathbf{R}^n \to (-\infty, \infty]$ be a lower semi-continuous extended-real-valued function. For $t > 0$, the Moreau envelope function is

$$g_t(x) := \inf_{z\in\mathbf{R}^n} g(z) + \frac{1}{2t}\|z - x\|^2.$$

In general, g_t is everywhere finite and locally Lipschitz continuous. Moreover, g_t converges pointwise to g as $t \to 0^+$. Moreau envelope is closely related to the inviscid Hamilton-Jacobi equation [14]

$$u_t + \frac{1}{2} |\,\nabla_x u\,|^2 = 0,\ \ u(x, 0) = g(x),$$

where $u(x,t) = g_t(x)$ is the unique viscosity solution of the above initial-value problem via the Hopf-Lax formula

$$u(x,t) = \inf_z \left\{ g(z) + tH^* \left(\frac{z-x}{t} \right) \right\}$$

with the Hamiltonian $H(t,x,v) = \frac{1}{2}\|v\|^2$ and its Fenchel conjugate $H^* = H$. The proximal mapping of g is defined by

$$\operatorname{prox}_g(x) := \arg\min_{z \in \mathbf{R}^n} \; g(z) + \frac{1}{2}\|z-x\|^2.$$

It is frequently used in optimization algorithms associated with non-smooth optimization problems such as total variation denoising [27].

In particular, if $g = \chi_{\mathcal{A}}$ is the indicator function of a close set $\mathcal{A} \subset \mathbf{R}^n$, where

$$\chi_{\mathcal{A}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{A}, \\ \infty & \text{otherwise,} \end{cases}$$

the Moreau envelope is well defined for $t > 0$

$$g_t(x) = \inf_z \; \chi_{\mathcal{A}}(z) + \frac{1}{2t}\|z-x\|^2 = \inf_{z \in \mathcal{A}} \; \frac{1}{2t}\|z-x\|^2 = \frac{1}{2t}\mathrm{dist}(x,\mathcal{A})^2,$$

and the proximal mapping $\operatorname{prox}_g(x)$ is reduced to the projection $\operatorname{proj}_{\mathcal{A}}(x)$.

Consider an alternative form of DNNs quantization problem (6.1)

$$\min_{x \in \mathbf{R}^n} \; f(x) + \chi_{\mathcal{Q}}(x), \tag{6.7}$$

When both the objective function $f(x)$ and the set $\mathcal{Q}$ are non-convex, the discontinuity of $\chi_{\mathcal{Q}}$ poses an extra challenge in minimization since a continuous gradient descent update can be made stagnant when projected discontinuously. The Moreau envelope of $\chi_{\mathcal{Q}}$ is $\frac{1}{2t}\text{dist}(x, \mathcal{Q})^2$, which is continuously differentiable almost everywhere, except at points that have at least two nearest line subspaces, i.e. there exist two different ways to quantize x. We use $\frac{1}{2t}\text{dist}(x, \mathcal{Q})^2$ as the approximant of the discontinuous $\chi_{\mathcal{Q}}(z)$ and minimize the relaxed training error

$$\min_{x\in\mathbf{R}^n} \; f(x) + \frac{\lambda}{2}\text{dist}(x, \mathcal{Q})^2, \tag{6.8}$$

where $\lambda = t^{-1} > 0$ is the regularization parameter. When $\lambda \to \infty$, $\frac{\lambda}{2}\text{dist}(x, \mathcal{Q})^2$ converges pointwise to $\chi_{\mathcal{Q}}(x)$, and the global minimum of (6.8) converges to that of (6.7).

Proposition 6.1. *Suppose $f(x)$ is continuous. Let $f^*_{\mathcal{Q}} = \min_{x\in\mathcal{Q}} f(x)$ be the global minimum of (6.7) and x^*_λ be the global minimizer of relaxed quantization problem* (6.8). *Then*

$$\text{dist}(x^*_\lambda, \mathcal{Q}) \to 0 \text{ and } f(x^*_\lambda) \to f^*_{\mathcal{Q}}, \text{ as } \lambda \to \infty.$$

Proof. Since x^*_λ is the global minimizer of (6.8),

$$f^*_{\mathcal{Q}} \geq f(x^*_\lambda) + \frac{\lambda}{2}\text{dist}(x^*_\lambda, \mathcal{Q})^2 \geq f^* + \frac{\lambda}{2}\text{dist}(x^*_\lambda, \mathcal{Q})^2,$$

where $f^* = \min\limits_{x\in\mathbf{R}^n} f(x) > -\infty$, so

$$\text{dist}(x^*_\lambda, \mathcal{Q}) \leq \sqrt{\frac{2(f^*_{\mathcal{Q}} - f^*)}{\lambda}} \to 0, \text{ as } \lambda \to \infty.$$

Denote $x^*_{\lambda,\mathcal{Q}} = \text{proj}_{\mathcal{Q}}(x^*_\lambda)$, then $\|x^*_{\lambda,\mathcal{Q}} - x^*_\lambda\| \to 0$ as $\lambda \to \infty$. Since $f^*_{\mathcal{Q}}$ is the minimum in $\mathcal{Q}$,

$$f(x^*_\lambda) + \frac{\lambda}{2}\text{dist}(x^*_\lambda, \mathcal{Q})^2 \leq f^*_{\mathcal{Q}} \leq f(x^*_{\lambda,\mathcal{Q}}) \to f(x^*_\lambda), \text{ as } \lambda \to \infty.$$

Therefore, $\lim\limits_{\lambda\to\infty} f(x^*_\lambda) = f^*_{\mathcal{Q}}$. □

6.1.3 Algorithm

Inspired by the hybrid gradient update proposed in [19], we write a two-line solver for the minimization problem (6.8):

$$\begin{cases} y^{k+1} = y^k - \gamma_k \nabla f_k(x^k) \\ x^{k+1} = \arg\min_{x \in \mathbf{R}^n} \dfrac{1}{2}\,\|x - y^{k+1}\|^2 + \dfrac{\lambda}{2}\mathrm{dist}(x, \mathcal{Q})^2. \end{cases} \tag{6.9}$$

The algorithm constructs two sequences: an auxiliary sequence of float weights $\{y^k\}$ and a sequence of *nearly* quantized weights $\{x^k\}$. The mismatch of discontinuous projection and continuous gradient descent is resolved by the relaxed quantization step in (6.9), which calls for computing the proximal mapping of the function $\dfrac{\lambda}{2}\mathrm{dist}(x, \mathcal{Q})^2$. This can be done via the following closed-form formula.

Proposition 6.2. *Let* $\mathrm{proj}_{\mathcal{Q}}(y^{k+1}) = \arg\min_{x\in\mathcal{Q}} \|x - y^{k+1}\|^2$ *be the quantization of* y^{k+1}*, then the solution to relaxed quantization subproblem in* (6.9) *is*

$$x^{k+1} = \frac{\lambda\,\mathrm{proj}_{\mathcal{Q}}(y^{k+1}) + y^{k+1}}{\lambda + 1}. \tag{6.10}$$

Proof. Problem (6.9) is equivalent to

$$\min_x \min_{z\in\mathcal{Q}} \ \frac{1}{2}\|x - y^k\|^2 + \frac{\lambda}{2}\|z - x\|^2 = \min_{z\in\mathcal{Q}} \min_x \ \frac{1}{2}\|x - y^k\|^2 + \frac{\lambda}{2}\|z - x\|^2.$$

With fixed $z \in \mathcal{Q}$, the inner problem is minimized at $x = \frac{\lambda z + y^k}{\lambda+1}$, then

$$\begin{aligned} z^* =& \arg\min_{z\in\mathcal{Q}} \frac{1}{2}\left\|\frac{\lambda z + y^k}{\lambda+1} - y^k\right\|^2 + \frac{\lambda}{2}\left\|z - \frac{\lambda z + y^k}{\lambda + 1}\right\|^2 \\ =& \arg\min_{z\in\mathcal{Q}} \|z - y^k\|^2 = \mathrm{proj}_{\mathcal{Q}}(y^k). \end{aligned}$$

Therefore, $x^k = \dfrac{\lambda\,\mathrm{proj}_{\mathcal{Q}}(y^k) + y^k}{\lambda + 1}$ is the optimal solution. □

We still need the exact quantization $\text{proj}_{\mathcal{Q}}(y^{k+1})$ to perform relaxed quantization. The update x^{k+1} is essentially a linear interpolation between y^{k+1} and its quantization $\text{proj}_{\mathcal{Q}}(y^{k+1})$, and λ controls the weighted average. x^{k+1} is thus not quantized because $x^{k+1} \notin \mathcal{Q}$, but x^{k+1} approaches $\mathcal{Q}$ as λ increases. Hereby we adopt a continuation strategy and let λ grow slowly. Specifically, we inflate λ after a certain number of epochs by a factor $\rho > 1$. Intuitively, the relaxation with continuation will help skip over some bad local minima of (6.7) located in $\mathcal{Q}$, because they are not local minima of the relaxed formulation in general.

Proposition 6.3. *Suppose $f(x)$ is differentiable. Any point $x^* \in \mathcal{Q}$ is not a local minimizer of the relaxed quantization problem* (6.8) *unless $\nabla f(x^*) = \mathbf{0}$.*

Proof. Proof by contradiction. Assume $x^* \in \mathcal{Q}$ is a local minimizer of problem (6.8) and $\nabla f(x^*) \neq \mathbf{0}$, then for any point x in the neighborhood of x^*,

$$f(x^*) \leq f(x) + \frac{\lambda}{2}\text{dist}(x, \mathcal{Q})^2 \leq f(x) + \frac{\lambda}{2}\|x - x^*\|^2.$$

Set $x = x^* - \beta\nabla f(x^*)$ with a small $\beta > 0$. The above inequality is reduced to

$$f(x^*) \leq f(x^* - \beta\nabla f(x^*)) + \frac{\lambda\beta^2}{2}\|\nabla f(x^*)\|^2. \tag{6.11}$$

On the other hand, by Taylor's expansion,

$$f(x^* - \beta\nabla f(x^*)) = f(x^*) - \beta\|\nabla f(x^*)\|^2 + o(\beta). \tag{6.12}$$

(6.11) and (6.12) imply

$$\|\nabla f(x^*)\|^2 \leq \frac{\lambda\beta}{2}\|\nabla f(x^*)\|^2 + o(1),$$

which leads to a contradiction as we let $\beta \to 0$. □

In order to obtain quantized weights in the end, we turn off the relaxation mode and enforce quantization. The BinaryRelax algorithm is summarized in Algorithm 3.

Algorithm 3 BinaryRelax.

Input: number of epochs for training, batch size, schedule of learning rate $\{\gamma_k\}$, growth factor $\rho > 1$.

```
for i = 1, 2,..., nb-epoch do
  Randomly shuffle the data and partition into batches.
  for j = 1, 2, ..., nb-batch do
    $y^{k+1} = y^k - \gamma_k \nabla f_k(x^k)$
    if $i \leq T$ then
      $x^{k+1} = \frac{\lambda_k \mathrm{proj}_{\mathcal{Q}}(y^{k+1}) + y^{k+1}}{\lambda_k + 1}$      // Phase I
      if increase $\lambda$ then
        $\lambda_{k+1} = \rho\lambda_k$
      else
        $\lambda_{k+1} = \lambda_k$
      end if
    else
      $x^{k+1} = \mathrm{proj}_{\mathcal{Q}}(y^{k+1})$     // Phase II
    end if
    $k = k + 1$
  end for
end for
```

Remark 6.1. *For BinaryRelax, we replace a discrete quantization constraint with a continuous regularizer. The similar idea of relaxing the discrete sparsity constraint $\|x\|_0 \leq s$ into a continuous and possibly non-convex sparse regularizer has been long known in the contexts of statistics and compressed sensing [67, 25, 9]. For example, compressed sensing solvers for minimizing the convex ℓ_1 norm [27] or non-convex sparse proxies, such as $\ell_{1/2}$ (with smoothing) [13] and ℓ_{1-2} [73], often empirically outperform those directly tackling the nonzero counting metric ℓ_0. Similar to the quantization set $\mathcal{Q}$, the sparsity constraint set $\{x \in \mathbf{R}^n : \|x\|_0 \leq s\}$ is also a finite union of low-dimensional subspaces in $\mathbf{R}^n$,*

$$\{x \in \mathbf{R}^n : \|x\|_0 \leq s\} = \bigcup_{\mathcal{S}\subset\{1,\ldots,n\},\, |\mathcal{S}|=s} \{x \in \mathbf{R}^n : \mathrm{supp}(x) \subseteq \mathcal{S}\}.$$

Remark 6.2. *BinaryRelax resembles the linearized Bregman algorithm proposed by Yin et* *al. [77, 78] for solving the basis pursuit problem [16, 9]*

$$\min_{x \in \mathbf{R}^n} \|x\|_1 \quad \textit{subject to} \quad Ax = b,$$

by iterating

$$\begin{cases} y^{k+1} = y^k - \tau_k A^\top (Ax^k - b) \\ x^{k+1} = \delta \cdot \mathrm{shrink}(y^{k+1}, \mu) \end{cases}$$

where δ, μ, $\tau_k > 0$ are algorithmic parameters. In linearized Bregman, $A^\top(Ax - b)$ is the gradient of sum of squares error $\frac{1}{2}\|Ax - b\|^2$, and $\mathrm{shrink}(y, \mu)$ *is the proximal mapping of ℓ_1 norm (soft-thresholding operator [23]):*

$$\mathrm{shrink}(y, \mu) := \arg\min_{u} \frac{1}{2\mu}\|u - y\|^2 + \|u\|_1.$$

Linearized Bregman also iterates between hybrid gradient step and proximal mapping, but is not exactly the same as BinaryRelax since there is a scaling by δ in the proximal step.

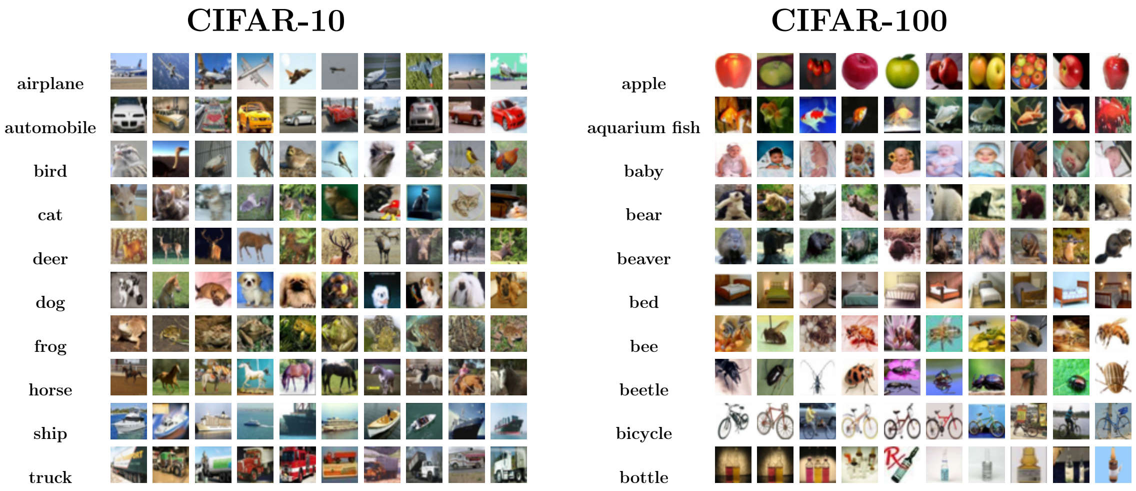
6.2 Experimental Results

We tested BinaryRelax on benchmark CIFAR [34] and ImageNet [21] color image datasets. The two baselines are the BinaryConnect framework combined with the exact binarization formula (6.4) (BWN) [59] and the heuristic ternarization scheme (6.6) (TWN) [39]. We used the same quantization formulas for BinaryRelax in the relaxed quantization update (6.10). Both algorithms were initialized with the weights of a pre-trained float model.

The relaxation parameter is initialized with $\lambda_0 = 1$. We split into roughly $4/5$ and $1/5$ of the training epochs for Phase I and Phase II. To guarantee the smooth transitioning to Phase II from Phase I, a proper growth factor $\rho > 1$ is chosen so that $\lambda \in (100, 300)$ at the moment Phase I ends. Small λ is likely to results in noticeable drop in accuracy when Phase II starts.

6.2.1 CIFAR datasets

The CIFAR-10 dataset consists of 60,000 32×32 color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. CIFAR-100 dataset is like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 test images per class. Figure 6.2 shows some sample images from CIFAR datasets. In the experiments, we used the testing images for validation. We coded up the BinaryRelax in PyTorch [55] platform. The experiments were carried out on two desktops with Nvidia graphics cards GTX 1080 Ti and Titan X.



CIFAR-10

CIFAR-100

Figure 6.2: Sample images from CIFAR datasets.

We ran 300 epochs. The initial learning rate $\gamma_0 = 0.1$ with decay by a factor of 0.1 at epochs $\{120, 220\}$. Phase II starts at epoch 240. λ increases by a factor of $\rho = 1.02$ after every epoch. In addition, we used batch size = 128, ℓ_2 weight decay $= 10^{-4}$, batch normalization [32], and momentum = 0.95.

We tested the algorithms on the popular VGG [65] and ResNet[28] architectures, and the validation accuracy for CIFAR-10 and CIFAR-100 is summarized in Table 6.1 and Table 6.2. ResNet-18 and ResNet-34 tested here were originally constructed for the more challenging ImageNet classification [21] and then adapted for CIFAR datasets. They have wider channels in the convolutional layers and are much larger than the other ResNets. For example, ResNet-18 has $\sim$ 11 million parameters, whereas ResNet-110 has only $\sim$ 1.7 million. This explains their higher accuracies. All quantized networks were initialized from their full-precision counterparts whose validation accuracies are listed in the second column.

Figure 6.3 shows the validation accuracies for CIFAR-100 tests with VGG-16 and ResNet-34 during the training process. The initial learning rate $\gamma_0 = 0.1$ and decays by a factor of 0.1 at epoch 120 and 220. The initial regularization parameter $\lambda_0 = 1$ and grows by a factor of $\rho = 1.02$ after each epoch until epoch 240 where Phase II starts. For the VGG-16 tests, we notice the decay of the validation accuracies of BinaryRelax occurs in Phase I training. This is due to the increase of the parameter λ, which makes the regularization more and more stringent. With approximately the same training cost, our relaxed quantization approach consistently outperforms the hard quantization counterpart in validation accuracies. As seen from the tables and figure, the advantage of relaxed quantization is particularly clear when it comes to the large nets ResNet-18 and ResNet-34, where we have more complex landscapes with spurious local minima. In this case, our accuracies of binarized networks even surpass that of TWN. The relaxation indeed helps skip over bad local minima during the training.

CIFAR-10	Float	Binary		Ternary	
		BWN	Ours	TWN	Ours
VGG-11	91.93	88.70	**89.28**	90.48	**91.01**
VGG-16	93.59	91.60	**91.98**	92.75	**93.20**
ResNet-20	92.68	87.44	**87.82**	88.65	**90.07**
ResNet-32	93.40	89.49	**90.65**	90.94	**92.04**
ResNet-18	95.49	92.72	**94.19**	93.55	**94.98**
ResNet-34	95.70	93.25	**94.66**	94.05	**95.07**

Table 6.1: CIFAR-10 validation accuracy.

CIFAR-100	Float	Binary		Ternary	
		BWN	Ours	TWN	Ours
VGG-11	70.43	62.35	**63.82**	64.16	**65.87**
VGG-16	73.55	69.03	**70.14**	71.41	**72.10**
ResNet-56	70.86	66.73	**67.65**	68.26	**69.83**
ResNet-110	73.21	68.67	**69.85**	68.95	**72.32**
ResNet-18	76.32	72.31	**74.04**	73.15	**75.24**
ResNet-34	77.23	72.92	**75.62**	74.43	**76.16**

Table 6.2: CIFAR-100 validation accuracy.

6.2.2 ImageNet

ImageNet (ILSVRC12) dataset [21] is a benchmark for large-scale image classification task, which has 1.2 million images for training and $50,000$ for validation of 1,000 categories. We quantize ResNet-18 at bit-widths 1 (binary) and 2 (ternary). The experiments were carried out on a machine with 8 Nvidia GeForce GTX 1080 Ti GPUs.

We initialized BinaryRelax with the pre-trained full-precision (32-bit) models available from the PyTorch torchvision package [55]. We trained in total 70 epochs, with phase II starting at epoch 55. The initial learning rate $\gamma_0 = 0.1$ and decays by a factor of 0.1 at epochs $\{30, 40, 50\}$. Relaxation parameter λ starts at 1 and increases by a growth factor of $\rho = 1.045$ after each half (1/2) epoch. In all these experiments, we used momentum$= 0.9$ and weight decay $= 10^{-4}$. The comparison results with BWN and TWN are listed in Table 6.3.

Network	Bit-width	Method	Top-1	Top-5
ResNet-18	32 (float)		69.6	89.0
	1 (binary)	BWN	60.8	83.0
		Ours	**63.2**	**85.1**
	2 (ternary)	TWN	61.8	84.2
		Ours	**66.5**	**87.3**

Table 6.3: ImageNet validation accuracy.

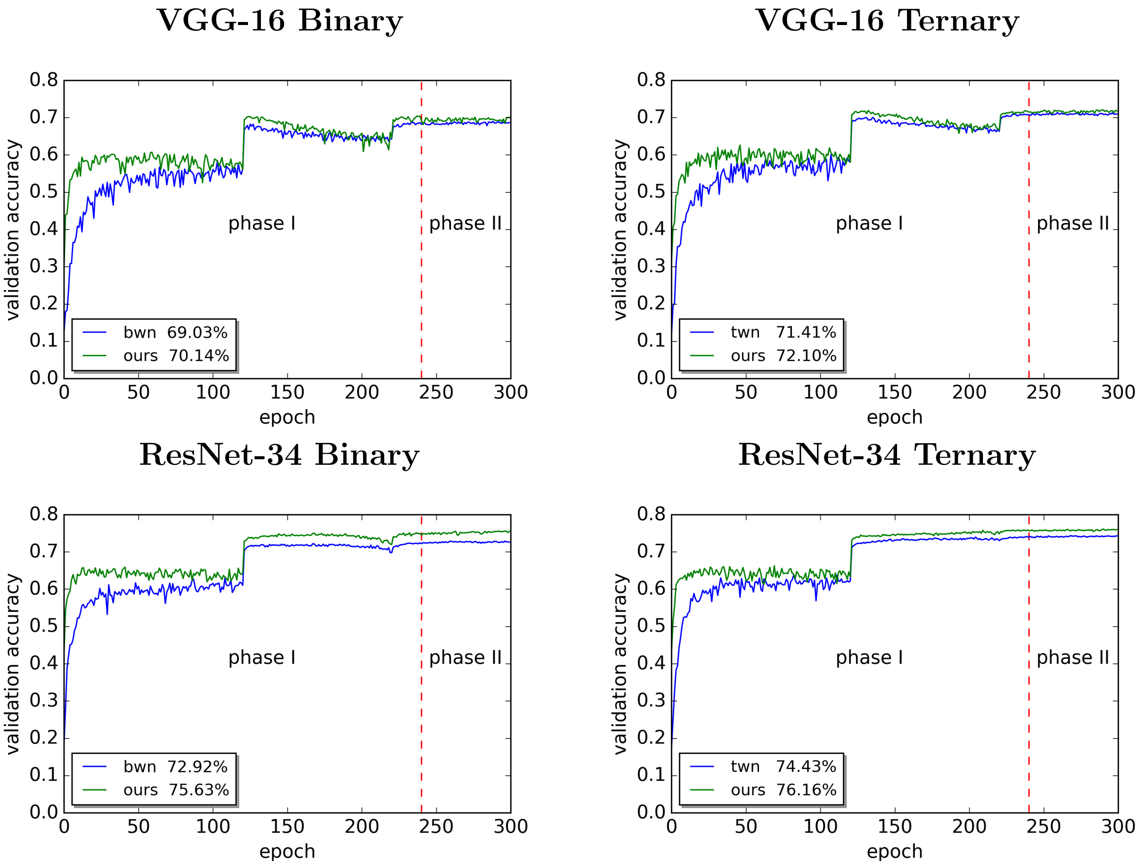


Figure 6.3: Validation accuracy curves for CIFAR-100 using VGG-16 and ResNet-34.

6.3 Convergence Analysis

Consider Phase II of BinaryRelax (i.e. BinaryConnect):

$$\begin{cases} y^{k+1} = y^k - \gamma_k \nabla f_k(x^k) \\ x^{k+1} = \mathrm{proj}_{\mathcal{Q}}(y^{k+1}). \end{cases} \tag{6.13}$$

Although the convergence of BinaryConnect at a small learning rate is observed empirically, the only convergence results, to our knowledge, were proved in [40], in terms of the objective value of float ergodic averages $\left\{ f\left(\frac{\sum_{i=1}^k y^i}{k} \right) \right\}$ under convexity assumption. Moreover, the quantization set $\mathcal{Q}$ considered in [40] is quite different. They constrain the quantized weights to $\{0, \pm\Delta, \pm 2\Delta, \dots\}$, where $\Delta > 0$ is some fixed resolution, hence $\mathcal{Q}$ is an unbounded Δ-lattice in $\mathbb{R}^n$. In section 6.1.1 $\mathcal{Q}$ takes the form of $\cup_{i=1}^{p} \mathcal{L}_i$ with each $\mathcal{L}_i$ being a line passing through the origin. This assumption on $\mathcal{Q}$ generalizes the binary and ternary cases in the existing literature such as [59, 39]. Without assuming the convexity of f, we will show the sequence $\{x^k\}$ generated by the iteration (6.13) subsequentially converges in expectation to an approximate critical point. To establish the convergence, we need to exploit the property of the set $\mathcal{Q}$ being the union of line subspaces by introducing several technical lemma.

6.3.1 Preliminaries

We have the following basic assumptions.

(i) $f(x)$ is bounded from below. Without loss of generality, we assume $f(x) \geq 0$.

(ii) $f(x)$ is L-Lipschitz differentiable, i.e. for any $x, y \in \mathbf{R}^n$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

(iii) $\mathbf{E}[\|\nabla f(x^k) - \nabla f_k(x^k)\|^2] \leq \sigma^2$ for all $k \in \mathbf{N}$, where the expectation is taken over the stochasticity of the algorithm (i.e. random selection of f_k).

Our proof relies on the following technical lemma that exploit the structure of set $\mathcal{Q}$.

Lemma 6.1 (Approximate orthogonality). *Let* $\{y^k\}, \{x^k\}$ *be defined in* (6.13). *There exists* $\alpha_k \geq 0$*, such that*

$$\alpha_k \|x^{k+1} - x^k\|^2 + \|y^k - x^k\|^2 = \|y^k - x^{k+1}\|^2.$$

Proof. Since $x^k, x^{k+1} \in \mathcal{Q}$ and $x^k = \operatorname{proj}_{\mathcal{Q}}(y^k)$, $\|y^k - x^k\|^2 \leq \|y^k - x^{k+1}\|^2$, i.e. $\alpha_k \geq 0$. □

Proposition 6.4. *Let $\theta_{\min}$ be the smallest angle formed by any two line subspaces in $\mathcal{Q}$. If $\|x^{k+1} - x^k\| < \|x^k\| \sin\theta_{\min}$, then $\alpha_k = 1$ in Lemma 6.1. Moreover, α_k may have to be* 0 *only when $\|y^k - x^k\| = \|y^k - x^{k+1}\|$ and $\nabla f_k(x^k) \perp \mathcal{L}_i$ with $\mathcal{L}_i$ containing x^{k+1}.*

Proof. Since the only intersection of the line subspaces is the origin, the distance between x^k and any other line is at least $\|x^k\| \sin\theta_{\min}$. If $\|x^{k+1} - x^k\| < \|x^k\| \sin\theta_{\min}$, then x^k and x^{k+1} must lie in the same line, and therefore $\alpha_k = 1$. On the other hand, if α_k can only be 0, then it must hold that $\|y^k - x^k\| = \|y^k - x^{k+1}\|$ and $x^k \neq x^{k+1}$, meaning that x^{k+1} is a different projection of y^k onto $\mathcal{Q}$. Moreover, since the projection of $y^{k+1} = y^k - \gamma_k \nabla f_k(x^k)$ onto $\mathcal{Q}$ is also x^{k+1}. Suppose $x^{k+1} \in \mathcal{L}_i \subset \mathcal{Q}$, then $\nabla f_k(x^k) \perp \mathcal{L}_i$. □

Lemma 6.2 (Alternative update)**.** *Let $\{x^k\}$ be defined in* (6.13)*. Suppose $x^{k+1} \in \mathcal{L}_i \subset \mathcal{Q}$ with $\mathcal{L}_i$ being some line subspace and define $\tilde{x}^k := \operatorname{proj}_{\mathcal{L}_i}(y^k)$, then*

$$x^{k+1} = \arg\min_{x \in \mathcal{L}_i} \|x - (\tilde{x}^k - \gamma_k \nabla f_k(x^k))\|^2.$$

Moreover, x^{k+1} is a local minimizer of the following problem

$$\min_{x \in \mathcal{Q}} \|x - (\tilde{x}^k - \gamma_k \nabla f_k(x^k))\|^2. \tag{6.14}$$

Proof. By the assumption,

$$x^{k+1} = \operatorname{proj}_{\mathcal{L}_i}(y^k - \gamma_k \nabla f_k(x^k)) = \operatorname{proj}_{\mathcal{L}_i}(\tilde{x}^k - \gamma_k \nabla f_k(x^k) + y^k - \tilde{x}^k).$$

Since $y^k - \tilde{x}^k \perp \mathcal{L}_i$ (see Figure 6.4),

$$x^{k+1} = \operatorname{proj}_{\mathcal{L}_i}(\tilde{x}^k - \gamma_k \nabla f_k(x^k)),$$

so x^{k+1} is the closest point to $\tilde{x}^k - \gamma_k \nabla f_k(x^k)$ on $\mathcal{L}_i$. If $\tilde{x}^k - \gamma_k \nabla f_k(x^k) = \mathbf{0}$, then $x^{k+1} = \mathbf{0}$ is the global minimizer of (6.14). Otherwise, $x^{k+1} \neq \mathbf{0}$. Since the line subspaces that constitute $\mathcal{Q}$ only intersect at the origin, there exists a neighborhood $\mathcal{N}$ of x^{k+1} such that $\mathcal{N} \cap \mathcal{Q} \subset \mathcal{L}_i$. Therefore, x^{k+1} is a local minimizer of problem (6.14). □

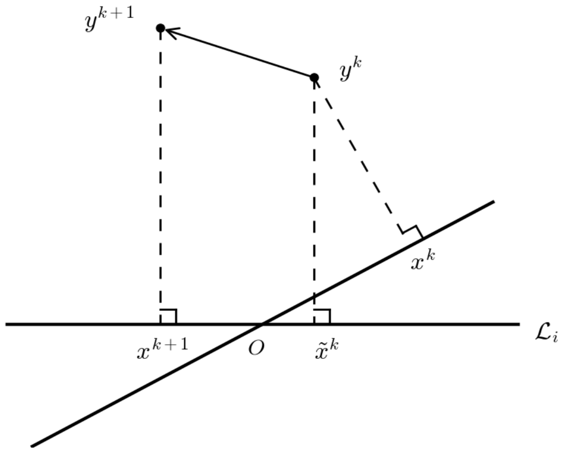


Figure 6.4: Illustration of Lemma 6.2. $y^{k+1} = y^k - \gamma_k \nabla f_k(x^k)$.

Lemma 6.3. *Let α_k and $\tilde{x}^k$ be defined in Lemma 6.1 and 6.2, resp., it holds that*

$$\|x^{k+1} - \tilde{x}^k\|^2 \leq \alpha_k \|x^{k+1} - x^k\|^2.$$

Proof. By the facts $x^k = \text{proj}_{\mathcal{Q}}(y^k)$, $\tilde{x}^k = \text{proj}_{\mathcal{L}_i}(y^k) \in \mathcal{Q}$, $x^{k+1} \in \mathcal{L}_i$ and Lemma 6.1,

$$\begin{aligned} &\|x^{k+1} - \tilde{x}^k\|^2 = \|y^k - x^{k+1}\|^2 - \|y^k - \tilde{x}^k\|^2 \\ \leq &\|y^k - x^{k+1}\|^2 - \|y^k - x^k\|^2 = \alpha_k \|x^{k+1} - x^k\|^2. \end{aligned}$$

□

Lemma 6.4 (Descent lemma [4])**.** *For any x, y, it holds that*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2.$$

Definition 6.1 (Subdifferential [49, 60])**.** *Let $h : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Define* $\mathrm{dom}(h) := \{x \in \mathbf{R}^n : h(x) < +\infty\}$*. For a given* $x \in \mathrm{dom}(h)$*, the Fréchet subdifferential of h at x, written as $\hat{\partial} h(x)$, is the set of all vectors $u \in \mathbf{R}^n$ which satisfy*

$$\liminf_{y \neq x \; y \to x} \frac{h(y) - h(x) - \langle u, y - x \rangle}{\|y - x\|} \geq 0.$$

When $x \notin \mathrm{dom}(h)$, we set $\hat{\partial} h(x) = \emptyset$. The (limiting) subdifferential, or simply the subdifferential, of h at $x \in \mathbb{R}^n$, written as $\partial h(x)$, is defined through the following closure process

$$\partial h(x) := \{u \in \mathbb{R}^n : \exists\, x^k \to x, \; h(x^k) \to h(x) \text{ and } u^k \in \hat{\partial} h(x^k) \to u \text{ as } k \to \infty\}.$$

6.3.2 Main results

Theorem 6.1. *Let $\{x^k\}$ be the sequence generated by* (6.13)*. Suppose there exist $\underline{\alpha}, \bar{\alpha}, \gamma > 0$ such that $\underline{\alpha} \leq \alpha_k \leq \bar{\alpha}$ and $\gamma_{k+1} \leq \gamma_k \leq \gamma < \dfrac{\bar{\alpha}}{2L}$ for all $k \in \mathbf{N}$, then*

$$\lim_{k\to\infty} \mathbf{E}\left[\|x^{k+1} - x^k\|^2\right] = 0,$$

if $\displaystyle\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. If further $\displaystyle\sum_{k=0}^{\infty} \gamma_k = \infty$, we have

$$\liminf_{k\to\infty} \mathbf{E}[\mathrm{dist}(\mathbf{0}, \partial h(x^k))^2] \leq 3\sigma^2 \left(\frac{4\bar{\alpha}}{\underline{\alpha}^2} + 1\right),$$

where $h = f + \chi_Q$ is the overall objective function.

Proof. By Lemma 6.4,

$$\begin{aligned} &f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ =&f(x^k) + \langle \nabla f_k(x^k), x^{k+1} - x^k \rangle + \langle \nabla f(x^k) - \nabla f_k(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2. \end{aligned} \tag{6.15}$$

The cross terms need care. Rewrite the update $x^{k+1} = \mathrm{proj}_{\mathcal{Q}}(y^k - \gamma_k \nabla f_k(x^k))$ as

$$x^{k+1} = \arg\min_{x \in \mathcal{Q}} \langle \nabla f_k(x^k), x \rangle + \frac{1}{2\gamma_k}\|x - y^k\|^2.$$

Since $x^k \in \mathcal{Q}$,

$$\langle \nabla f_k(x^k), x^{k+1} \rangle + \frac{1}{2\gamma_k}\|x^{k+1} - y^k\|^2 \leq \langle \nabla f_k(x^k), x^k \rangle + \frac{1}{2\gamma_k}\|x^k - y^k\|^2.$$

By Lemma 6.1,

$$\langle \nabla f_k(x^k), x^{k+1} - x^k \rangle \leq \frac{1}{2\gamma_k}(\|x^k - y^k\|^2 - \|x^{k+1} - y^k\|^2) \leq -\frac{\underline{\alpha}}{2\gamma_k}\|x^{k+1} - x^k\|^2. \tag{6.16}$$

By Young's inequality,

$$\langle \nabla f(x^k) - \nabla f_k(x^k), x^{k+1} - x^k \rangle \leq \frac{\gamma_k}{\underline{\alpha}}\|\nabla f(x^k) - \nabla f_k(x^k)\|^2 + \frac{\underline{\alpha}}{4\gamma_k}\|x^{k+1} - x^k\|^2. \tag{6.17}$$

Combining (6.15), (6.16) and (6.17) and taking the expectation gives

$$\mathbf{E}[f(x^{k+1})] \leq \mathbf{E}[f(x^k)] - \frac{\underline{\alpha} - 2\gamma_k L}{4\gamma_k}\mathbf{E}[\|x^{k+1} - x^k\|^2] + \frac{\gamma_k \sigma^2}{\underline{\alpha}}. \tag{6.18}$$

Multiplying (6.18) by γ_k and using $\alpha_k \geq \underline{\alpha} > 0$, $\gamma_{k+1} \leq \gamma_k \leq \gamma < \dfrac{\underline{\alpha}}{2L}$ and $f \geq 0$, we obtain

$$\gamma_{k+1}\mathbf{E}[f(x^{k+1})] \leq \gamma_k \mathbf{E}[f(x^{k+1})] \leq \gamma_k \mathbf{E}[f(x^k)] - (\underline{\alpha} - 2\gamma L)\mathbf{E}[\|x^{k+1} - x^k\|^2] + \frac{\gamma_k^2 \sigma^2}{\underline{\alpha}}.$$

Rearranging terms in the above inequality and taking the sum over k yields

$$(\underline{\alpha} - 2\gamma L)\sum_{k=0}^{\infty} \mathbf{E}[\|x^{k+1} - x^k\|^2] \leq \gamma f(x^0) - \lim_{k\to\infty} \gamma_k \mathbf{E}[f(x^k)] + \frac{\sigma^2}{\underline{\alpha}} \sum_{k=0}^{\infty} \gamma_k^2 < \infty.$$

Therefore, $\lim\limits_{k\to\infty} \mathbf{E}[\|x^{k+1} - x^k\|^2] = 0$.

By Lemma 6.2, the first-order optimality condition of (6.14) holds at x^{k+1}, so

$$\mathbf{0} \in \nabla f_k(x^k) + \frac{x^{k+1} - \tilde{x}^k}{\gamma_k} + \partial \chi_{\mathcal{Q}}(x^{k+1}),$$

which implies

$$-\frac{x^{k+1} - \tilde{x}^k}{\gamma_k} - \nabla f_k(x^k) + \nabla f(x^{k+1}) \in \nabla f(x^{k+1}) + \partial \chi_{\mathcal{Q}}(x^{k+1}) = \partial h(x^{k+1}).$$

By Lemma 6.3 and the assumption that f is L-Lipschitz differentiable,

$$\begin{aligned}
&\mathbf{E}[\operatorname{dist}(\mathbf{0}, \partial h(x^{k+1}))^2] \\
\leq&\mathbf{E}\left[\left\|-\frac{x^{k+1} - \tilde{x}^k}{\gamma_k} - \nabla f_k(x^k) + \nabla f(x^{k+1})\right\|^2\right] \\
\leq&3\left(\mathbf{E}\left[\frac{\|x^{k+1} - \tilde{x}^k\|^2}{\gamma_k^2}\right] + \mathbf{E}[\|\nabla f_k(x^k) - \nabla f(x^k)\|^2] + \mathbf{E}[\|\nabla f(x^k) - \nabla f(x^{k+1})\|^2]\right) \\
\leq&3\left(\bar{\alpha}\mathbf{E}\left[\frac{\|x^{k+1} - x^k\|^2}{\gamma_k^2}\right] + \sigma^2 + L^2\mathbf{E}[\|x^{k+1} - x^k\|^2]\right). \qquad (6.19)
\end{aligned}$$

It follows from (6.18) that

$$\gamma_k\left((\underline{\alpha} - 2\gamma_k L)\mathbf{E}\left[\frac{\|x^{k+1} - x^k\|^2}{4\gamma_k^2}\right] - \frac{\sigma^2}{\underline{\alpha}}\right) \leq \mathbf{E}[f(x^k) - f(x^{k+1})].$$

Summing the above inequality over k yields

$$\sum_{k=0}^{\infty} \gamma_k \left((\underline{\alpha} - 2\gamma_k L)\mathbf{E}\left[\frac{\|x^{k+1} - x^k\|^2}{4\gamma_k^2} \right] - \frac{\sigma^2}{\underline{\alpha}} \right) \leq f(x^0) < \infty.$$

Since $\gamma_k > 0$ and $\displaystyle\sum_{k=1}^{\infty} \gamma_k = \infty$,

$$\liminf_{k \to \infty} (\underline{\alpha} - 2\gamma_k L)\mathbf{E}\left[\frac{\|x^{k+1} - x^k\|^2}{4\gamma_k^2} \right] - \frac{\sigma^2}{\underline{\alpha}} \leq 0,$$

thus

$$\liminf_{k \to \infty} \ \mathbf{E}\left[\frac{\|x^{k+1} - x^k\|^2}{\gamma_k^2} \right] \leq \lim_{k \to \infty} \frac{4\sigma^2}{\underline{\alpha}(\underline{\alpha} - 2\gamma_k L)} = \frac{4\sigma^2}{\underline{\alpha}^2}.$$

It follows from (6.19) that

$$\liminf_{k \to \infty} \ \mathbf{E}[\mathrm{dist}(\mathbf{0}, \partial h(x^k))^2] \leq 3\sigma^2 \left(\frac{4\bar{\alpha}}{\underline{\alpha}^2} + 1 \right),$$

□

Chapter 7

Conclusions

We have studied the residual diffusion arising from the singularly-perturbed advection diffusion equations. Due to the periodic property of the model, we were able to apply spectral method to solving for the singluar solutions. By constructing the Poincaré map of the truncated system, we found the numerical periodic solution and computed the effective diffusivity in the frequency domain. Numerical results indicate the existence of residual diffusivity in time periodic cellular flow as well as the non monotone dependence on the chaotic terms. However, as the cellular diffusivity D_0 becomes smaller, much more Fourier modes are needed to approximate the singular solution, which result in more computational cost. Hence a challenge is to find effective ways to solve the singular problem. We constructed orthogonal adaptive basis functions based on learning from the fully resolved spectral method at sampled small D_0 to aid the low cost computation of residual diffusivity. Even though solutions develop large gradients and demand a large number of Fourier modes to resolve, the adaptive basis functions maintain accuracy of residual diffusivity at much smaller number of basis functions, uniform in the limit of zero molecular diffusivity. We have tested the straight forward SVD of solution matrix and it achieves satisfactory performance in certain regimes. We have also applied deep neural network, namely SRGAN, to the adaptive basis learning in order to capture the structural characteristics of singular solutions. The SRGAN assisted learning actually gives better predictions of residual diffusivity than straight forward SVD for some D_0. Moreover, there is a lot of freedom in exploring the network model so to refine the training for adaptive basis and increase the robustness of the learning.

We have investigated the residual diffusion phenomenon in random walk as well. We found that residual diffusivity occurs in ERWS models in one and two dimensions with an inclusion of small probability of symmetric random walk steps. A wedge like sub-diffusive parameter region in the (r, γ) plane transitions into a diffusive region with residual diffusivity in the sense that the enhanced diffusivity strictly exceeds the un-perturbed diffusivity in the limit of vanishing symmetric random walks. It would be of great interest to identify other discrete stochastic models for residual diffusivity so that the region where this occurs remains distinct from the un-perturbed diffusivity region in the limit of vanishing diffusive perturbations. A recent work along this line is a study on perturbed senile reinforced random walk models where the enhanced diffusivity is near residual diffusivity within poly-logarithmic factors [22]. For the nonlinear case of the curvature dependent flame propagation, we studied the effective burning velocity. We have proved that the turbulent flame speed is decreasing with respect to the Markstein number for shear flows in the G-equation model. In the proof, we have established several novel and rather sophisticated inequalities arising from the nonlinear structure of the equation. We also found the "physical fluctuations" when the Markstein number goes to zero, and the analytical formula of the limiting solution.

In the end, we have done dimensionality reduction for general DNNs. From optimization point of view, we proposed BinaryRelax, a novel relaxation approach for training quantized neural networks. Our algorithm iterates between a hybrid gradient step for updating the float weights and a weighted average of the computed float weights and their quantizations. We increase slowly the parameter that controls the average to drive the weights to the quantized state. In order to get the purely quantized weights, exact quantization replaces the weighted average in the second phase of training. Extensive experiments show that with about the same training cost, BinaryRelax is consistently better than its BinaryConnect counterpart in terms of validation accuracy. It has clearer advantage on larger networks, which yield more complex landscape of the training loss with spurious local minima. In addition, our convergence analysis shows BinaryRelax is provably convergent in expectation under an approximate orthogonality condition. It is natural to ask if we can compress DNNs even more, which motivates the construction of fully quantized DNNs that have both weights and activation functions quantized. Our most recent work has proposed a blended method to train DNNs with full quantization [74].

Bibliography

[1] N. Anantharaman. On the zero-temperature or vanishing viscosity limit for certain markov processes arising from lagrangian dynamics. *Journal of the European Mathematical Society*, 6(2):207–276, 2004.

[2] N. Anantharaman, R. Iturriaga, P. Padilla, and H. Sánchez-Morgado. Physical solutions of the hamilton-jacobi equation. *Discrete and Continuous Dynamical Systems Series B*, 5(3):513–528, 2005.

[3] A. Bensoussan, G. C. Papanicolaou, and P.-L. Lions. *Asymptotic analysis for periodic structures.* AMS Chelsea Publishing, 2011.

[4] D. P. Bertsekas. *Nonlinear programming.* Athena scientific Belmont, 1999.

[5] L. Biferale, A. Cristini, M. Vergassola, and A. Vulpiani. Eddy diffusivities in scalar transport. *Physics Fluids*, 7(11):2725–2734, 1995.

[6] L. Caffarelli and R. Monneau. Counter-example in three dimension and homogenization of geometric motions in two dimension. *Archive for Rational Mechanics and Analysis*, 212(2):503–574, 2014.

[7] Z. Cai, X. He, J. Sun, and N. Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. *arXiv preprint arXiv:1702.00953*, 2017.

[8] R. Camassa and S. Wiggins. Chaotic advection in a rayleigh-bénard flow. *Physical Review A*, 43(2):774–797, 1990.

[9] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal rconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52:489–509, 2006.

[10] P. Cardaliaguet, J. Nolen, and P. E. Souganidis. Homogenization and enhancement for the g-equation in periodic media. *Archive for Rational Mechanics and Analysis*, 199(2):527–561, 2011.

[11] M. Carreira-Perpinán. Model compression as constrained optimization, with application to neural nets. part i: General framework. *arXiv preprint arXiv:1707.01209*, 2017.

[12] M. Carreira-Perpinán and Y. Idelbayev. Model compression as constrained optimization, with application to neural nets. part ii: quantization. *arXiv preprint arXiv:1707.04319*, 2017.

[13] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *ICASSP*, pages 3869–3872, 2008.

[14] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *arXiv preprint arXiv:1704.04932*, 2017.

[15] S. Chaudhuri, F. Wu, and C. K. Law. Scaling of turbulent flame speed for expanding flames with markstein diffusion considerations. *Physical Review E*, 88, 2013.

[16] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[17] S. Childress and A. D. Gilbert. Stretch, twist, fold: the fast dynamo. *Lecture Notes in Physics Monographs*, 37, 1995.

[18] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1:13–37, 2016.

[19] M. Courbariaux, Y. Bengio, and J. David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, pages 3123—-3131, 2015.

[20] J. C. Cressoni, G. M. Viswanathan, and M. A. A. d. Silva. Exact solution of an anisotropic 2d random walk model with strong memory correlations. *Journal of Physics A: Mathematical and Theoretical*, 46(50), 2013.

[21] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[22] T. Dinh and J. Xin. Enhanced diffusivity in perturbed senile reinforced random walk models. *arXiv:1807.03744*, 2018.

[23] D. Donoho. De-noising by soft-thresholding. *IEEE Trans. Info. Theory*, 41:613–627, 1995.

[24] L. C. Evans. Towards a quantum analog of weak kam theory. *Communications in Mathematical Physics*, 244(2):311–334, 2004.

[25] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Stat. Assoc.*, 96:1348–1360, 2001.

[26] A. Fannjiang and G. C. Papanicolaou. Convection enhanced diffusion for periodic flows. *SIAM Journal on Applied Mathematics*, 54(2):333–408, 1994.

[27] T. Goldstein and S. Osher. The split bregman method for ℓ_1-regularized problems. *SIAM J. Imaging Sci.*, 2:323–343, 2009.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[29] S. Heinze. Diffusion-advection in cellular flows with large peclet numbers. *Archive for Rational Mechanics and Analysis*, 168(4):329–342, 2003.

[30] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge University Press, 1998.

[31] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks: Training neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016.

[32] S. Ioffe and C. Szegedy. Normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[33] H. Jauslin, H. Kreiss, and J. Moser. On the forced burgers equation with periodic boundary conditions. *Proceedings of Symposia in Pure Mathematics*, 65:133–153, 1999.

[34] A. Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

[35] N. Kumar, U. Harbola, and K. Lindenberg. Memory-induced anomalous dynamics: Emergence of diffusion, subdiffusion, and superdiffusion from a single random walk model. *Physical Review E*, 82(2 Pt 1):021101, 2010.

[36] G. Lacey, G. W. Taylor, and S. Areibi. Stochastic layer-wise precision in deep neural networks. *arXiv preprint arXiv:1807.00942*, 2018.

[37] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, pages 105–114, 07 2017.

[38] C. Leng, H. Li, S. Zhu, and R. Jin. Extremely low bit neural network: Squeeze the last bit out with admm. *arXiv preprint arXiv:1707.09870*, 2017.

[39] F. Li, B. Zhang, and B. Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[40] H. Li, S. De, Z. Xu, C. Studer, H. Samet, and T. Goldstein. Training quantized nets: A deeper understanding. In *NIPS*, pages 5813–5823, 2017.

[41] Z. Li, X. Wang, X. Lv, and T. Yang. Sep-nets: Small and effective pattern networks. *arXiv preprint arXiv:1706.03912*, 2017.

[42] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Info. Theory*, 28:129–137, 1982.

[43] J. L. Lumley. *Coherent structures in turbulence*. Academic Press, 1981.

[44] J. Lyu, J. Xin, and Y. Yu. Computing residual diffusivity by adaptive basis learning via spectral method. *Numerical Mathematics: Theory, Methods & Applications*, 10(2):351–372, 2017.

[45] J. Lyu, J. Xin, and Y. Yu. Curvature effect in shear flow: slowdown of turbulent flame speeds with markstein number. *Communications in Mathematical Physics*, 359(2):515–533, 2018.

[46] J. Lyu, J. Xin, and Y. Yu. Residual diffusivity in elephant random walk models with stops. *Communications in Mathematical Sciences*, to appear (arXiv:1705.02711, 2017).

[47] A. J. Majda and P. R. Kramer. Simplified models for turbulent diffusion: theory, numerical modelling, and physical phenomena. *Physics Reports*, 314(4-5):237–574, 1999.

[48] G. H. Markstein. Experimental and theoretical studies of flame front stability. *Journal of the Aeronautical Sciences*, 18:199–209, 1951.

[49] B. Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*. Springer Science & Business Media, 2006.

[50] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[51] N. B. Murphy, E. Cherkaev, J. Xin, J. Zhu, and K. M. Golden. Spectral analysis and computation of effective diffusivity in space-time periodic incompressible flows. *Annals of Mathematical Sciences and Applications*, 2(1):3–66, 2017.

[52] A. Nasrollahi, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[53] A. Novikov and L. Ryzhik. Boundary layers and kpp fronts in a cellular flows. *Archive for Rational Mechanics and Analysis*, 184(1):23–48, 2007.

[54] E. Park, J. Ahn, and S. Yoo. Weighted-entropy-based quantization for deep neural networks. In *CVPR*, pages 5456–5464, 2017.

[55] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. *Tech Report*, 2017.

[56] N. Peters. *Turbulent Combustion*. Cambridge University Press, Cambridge, 2000.

[57] A. Quarteroni and G. Rozza. *Reduced order methods for modeling and computational reduction*, volume 9. Springer, 2014.

[58] K. Radford and T. B. Moeslund. Super-resolution: A comprehensive survey. *Machine Vision and Applications*, 25:1423–1468, 2014.

[59] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. *arXiv preprint arXiv:1603.05279*, 2016.

[60] R. Rockafellar and R. Wets. *Variational analysis*. Springer Science & Business Media, 2009.

[61] P. D. Ronney. Some open issues in premixed turbulent combustion. *Buckmaster J., Takeno T. (eds) Modeling in Combustion Science. Lecture Notes in Physics*, 449:3–22, 1995.

[62] L. Rosasco, S. Villa, and B. C. Vũ. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

[63] G. Schuetz and S. Trimper. Elephants can always remember: exact long-range memory effects in a non-markovian random walk. *Physical Review E*, 70(4 Pt 2):045101, 2004.

[64] J. A. Sethian. Curvature and the evolution of fronts. *Communications in Mathematical Physics*, 101(4):487–499, 1985.

[65] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[66] G. I. Taylor. Diffusion by continuous movements. *Proceedings of the London Mathematical Society*, 2:196–211, 1922.

[67] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.*, 58:267–288, 1996.

[68] P. F. C. Tilles, S. V. Petrovskii, and P. L. Natti. A random walk discription of individual animal movement accounting for periods of rest. *Royal Society Open Science*, 3(11):160566, 2016.

[69] J. Xin and Y. Yu. Periodic homogenization of inviscid g-equation for incompressible flows. *Communications in Mathematical Sciences*, 8(4):1067–1078, 2010.

[70] J. Xin and Y. Yu. Sharp asymptotic growth laws of turbulent flame speeds in cellular flows by inviscid hamilton-jacobi models. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 30(6):1049–1068, 2013.

[71] J. Xin and Y. Yu. Front quenching in g-equation model induced by straining of cellular flow. *Archive for Rational Mechanics and Analysis*, 214:1–34, 2014.

[72] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. *CVPR*, pages 1–8, 2007.

[73] P. Yin, Y. Lou, Q. He, and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM J. Sci. Comput.*, 37:A536–A563, 2015.

[74] P. Yin, S. Zhang, J. Lyu, S. Osher, Y. Qi, and J. Xin. Blended coarse gradient descent for full quantization of deep neural networks. *arXiv:1808.05240*, 2018.

[75] P. Yin, S. Zhang, J. Lyu, S. Osher, Y. Qi, and J. Xin. Binaryrelax: a relaxation approach for training deep neural networks with quantized weights. *SIAM Journal on Imaging Sciences*, to appear (arXiv:1801.06313, 2017).

[76] P. Yin, S. Zhang, Y. Qi, and J. Xin. Quantization and training of low bit-width convolutional neural networks for object detection. *J. Comput. Math., to appear*, 2018.

[77] W. Yin. Analysis and generalizations of the linearized bregman method. *SIAM J. Imaging Sci.*, 3(4):856–877, 2010.

[78] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143—-168, 2010.

[79] Y. Yoshida, R. Oiwa, and T. Kawahara. Ternary sparse xnor-net for fpga implementation. In *International Symposium on Next Generation Electronics (ISNE)*, pages 1–2. IEEE, 2018.

[80] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.

[81] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv: 1606.06160*, 2016.

[82] C. Zhu, S. Han, H. Miao, and W. Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

[83] A. Zlatoš. Sharp asymptotics for kpp pulsating front speed-up and diffusion enhancement by flows. *Archive for Rational Mechanics and Analysis*, 195:441–453, 2010.

[84] W. W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21:327–340, 2012.

[85] P. Zu, L. Chen, and J. Xin. A computational study of residual kpp front speeds in time-periodic cellular flows in the small diffusion limit. *Physica D*, 311-312:37–44, 2015.