Respect Your Data:
Topics in Inference and Modeling in Physics

A Dissertation

Presented to the Faculty of the Graduate School

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Colin Clement

December 2019

ABSTRACT

Respect Your Data:

Topics in Inference and Modeling in Physics

Colin Clement, Ph.D.

Cornell University 2019

We discuss five topics related to inference and modeling in physics: image registration, magnetic image deconvolution, effective models of spin glasses, the two-dimensional Ising model, and a benchmark dataset of the arXiv pre-print service. First, we solve outstanding problems with image registration (which aims to infer the rigid shift relating two or more noisy shifted images), obtaining the information-theoretic limit in the precision of image shift estimation. Then, we use Bayesian inference and develop new physically-motivated priors in order to solve the ill-posed deconvolution problem of reconstructing electric currents from a magnetic images. After that, we apply machine learning and information geometry to study a spin glass model, finding that this model of canonical complexity is sloppy and thus allows for lower-dimensional effective descriptions. Next, we address outstanding questions regarding the corrections to scaling of the two dimensional Ising model by applying Normal Form Theory of dynamical systems to the Renormalization Group (RG) flows and raise important questions about the RG in various statistical ensembles. Finally, we develop tools and practices to cast the entire arXiv pre-print service into a benchmark dataset for studying models on graphs with multi-modal features.

BIOGRAPHICAL SKETCH

Colin Bruce Clement was born in Greeley, Colorado in Sept 1989. When Colin was 9 years old, his parents Julie and Bruce, and sister Maire moved with their two cats and a goldfish to just south of the Twin Cities in Minnesota. Colin attended Southview elementary school, Valley Middle School, and Apple Valley High School. Throughout his childhood he imagined becoming a scientist. It was in high school that this specialization was fostered through multiple Advanced Placement science courses, a mentorship program in aerospace engineering, and Science Olympiad. In high school Colin played the Clarinet in the wind ensemble, and combined the love of music which runs in his family with his interest in science, placing first in the state in Science Olympiad for building a set of hanging chimes which could play only in the key of G.

As high school ended in 2008, Colin was interested in aerospace engineering, inspired by beautiful aircraft, and entered undergraduate studies at the University of Minnesota in Minneapolis, Minnesota. During his first semester he took the Physics courses required for all science and engineering students, and was captured by the fundamental nature of physics. Shortly into his studies he switched to a Physics major, and doubled in

Mathematics. During his studies he continued playing the Clarinet with the U of M wind ensemble, and worked part-time in a low-temperature physics laboratory studying superconductors with Allen Goldman. In the first two summers at the U of M he worked as an intern at Lockheed Martin researching analog optical computers, in the third he went to Cornell University and studied theories of dilute solutions with Benjamin Widom, and in the summer following undergraduate studies he studied a model of spin torque oscillators with an electrical engineering professor.

Colin returned to Ithaca, NY in August 2012, this time as a graduate student in Physics. He began research with James P. Sethna, who inspired him with a computational physics course and an pedagogically experimental Quantum Mechanics course. Colin started working on a model of a particle accelerator, and moved on to applying the Sethna groups sloppines paradigm to spin glasses, the subject of chapter 4. Colin became interested in inference and machine learning, inspired by his truly and ostensibly successful colleagues. During his A-exam, Colin began a collaboration with Katja Nowack, who introduced him to the problem they solved in chapter 3 using Bayesian inference. Colin enjoyed collaborating with electron microscopists, who introduced him to the problem of image registration studied in chapter 2. During this time Colin began studying the Renormalization Group with colleagues Lorien Hayden and Archishman Raju, the fruits of which are discussed in chapter 5. Near the end of his PhD, Colin focused more and more on Machine Learning, and presented the work of chapter 6 at a conference in New Orleans in 2019 with his colleagues.

Colin spent much of his time in Ithaca pursuing interests beyond science. He began playing Bluegrass and Classical Mandolin in 2013, and joined the Cornell University Klezmer ensemble in 2015. He learned to bake sourdough bread from a natural yeast starter, and explored cooking techniques. He joined the Cornell Graduate Student Union recognition campaign in 2016. Though that struggle resulted in a narrow electoral loss, Colin learned history, politics, and philosophy he had never experienced before, and gained many diverse and beautiful friendships. Thus his PhD experience spawned a new chapter of lifelong learning.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my family for putting up with me for so long. I certainly would not have been able to pursue all of this education without the opportunities their enthusiastic support afforded me. Thanks also to my loving partner, Liz Baker, who not only tolerated me through the majority of my PhD, but also spent many hours copy-editing this manuscript.

To my friends throughout my time at Cornell. For the time they spent with me on walks around Beebe Lake, through long lunches and coffee breaks. For the conversations, the favors, and the memories of a truly caring community.

To Katja Nowack who—even as a new professor with a growing group and lab to manage—was so generous with her time and helpful with her feedback that she essentially functioned as my second advisor.

A special thanks, of course, goes to my advisor Jim Sethna who was sympathetic, honest, and generous with his time. Who always gave me credit for the good ideas he came up with in response to my naïve questions. I know many who struggled through graduate school due, in part, to the poor management provided by their advisors. My

experience could not have been more the opposite and I feel truly grateful to Jim for providing the space and time for exploring things at my own pace.

And to the group of exceptional people Jim brought together into his research group. To Alex Alemi and Matt Bierbaum who endured so many of my questions about Linux, `python`, and everything to do with computers. And to Archishman Raju for not only being so fun to work with, but also helping me appreciate the importance of philosophy of history in all work.

Contents

Chapter 1

Introduction

The title of this dissertation—respect your data—is both a succinct summary of what I learned while performing this research and the first advice I would offer to anyone attempting to optimally extract information from data. What I mean by 'respect' is that data should be sacrosanct: never modified, only compared to. Generative models should be built to produce synthetic data as convincingly as possible, including the experimental noise. Since noise is modeled probabilistically, we must use inference to extract the latent information from data and, while Bayesian inference is the best tool I know for extracting information from data, we essentially never have the correct model for real data, so Bayesian inference is not a perfect machine. Therefore, 'respect your data' also means that the human researcher building a model should intelligently interrogate its predictions and compare them directly to the data.

In Physics, modeling serves (at least) two essential purposes. First, interpretable 'toy' models—which are simple but contain the essence of some physical phenomenon—allow physicists to obtain the intuition they are so famed to possess. Second, generative models—some theory capable of producing convincing data— are essential for extracting information from an experiment. In this dissertation, I will explore two toy models: the two-dimensional (2D) Ising model (a model of a sheet of ferromagnetic material), and the 2D Ising spin glass (a model for a sheet of a disordered or glassy material). I will also study two generative models used to infer physical quantities from experiments: inferring the rigid shift relating noisy images for combining and de-noising them and inferring electric surface currents from a magnetic field image. Finally in this dissertation, I lay the groundwork necessary for reproducible machine-learning modeling of the arXiv pre-print service.

1.1 What is inference?

The equations of probability theory are perhaps the only logically consistent set of rules for qualitative reasoning with some degree of common sense. If Bayes' theorem updates probabilites in the face of prior information, we have Bayesian inference. If we add the principal of maximum entropy, we can even obtain thermodynamics and communications theory as special cases [Jaynes, 1988].

What is qualitative reasoning? What does it mean to adhere to common sense? These questions are attempting to develop a *model* for a process the human brain performs regularly with ease. To model reasoning, we must define the *state* of a human mind, which the act of reasoning will modify. The state is a description of all the knowledge and experience of a person, and therefore has many dimensions. Consider for a moment all the conditions you pondered (perhaps unconsciously) when choosing the clothes you are wearing: the temperature outside, at home, at work, precipitation, social expectations, or even morality of materials!

Since we are only beginning to build a model, we will follow Laplace and Jaynes, and consider a one-dimensional state of mind: call it the degree of belief or plausibility (A) of some fact or outcome A, e.g. ‘it will rain today.’ We can define some elementary relations combining dimensions of degrees of belief of (A) and (B):

- (AB): “The plausibility that both A and B are true",
- $(A+B)$: “The plausibility that at least one of A, B are true",
- $(A|B)$: “The plausibility of A conditioned on B being true".

The plausibilities have a kind of ordering, so that $(A) > (B)$ is understood to mean ‘A is more plausible than B’. We can also denote more complex relations like $(A|C) > (B)$ which reads ‘Assuming C is true, then A is more plausible than B.’

How may we calculate (AB), the plausibility of both A and B, using (A) and (B)? We must first recognize that there will always be tertiary propositions C, upon which the plausibilities of A and B *may* depend. Therefore we should seek to calculate $(AB|C)$, the plausibility of both A and B assuming C is true. Jaynes shows us that the only way to relate this plausibility to plausibilities of A and B is

$$(AB|C) = (A|BC)(B|C), \tag{1.1}$$

which is read as ‘the plausibility of A and B (assuming C is true), is equal to the plausibility of A (assuming B and C are true), times the plausibility of B (assuming C is true).’ This expression contains some *common sense*, as the plausibility of A and B increases as either of them become more plausible and vice-versa.

There is a symmetry in this expression, as A and B are merely labels which can be swapped, and so in order that this expression be *consistent*,

$$(AB|C) = (BA|C),$$
$$(A|BC)(B|C) = (B|AC)(A|C) \tag{1.2}$$

must be true. Hiding C for clarity, we can rewrite eqn. 1.2 as

$$(A|B) = \frac{(B|A)(A)}{(B)}. \tag{1.3}$$

If we make a leap to associate these plausibilities with *probabilities*, and interpret B as a precipitating event or information known *prior* to A, this is none other than *Bayes' Theorem*.

Bayes' Theorem may be applied as follows. Say $(B) = 0.5$ (50%) is the probabability that it will be cloudy today at any time in, for example, Ithaca, NY, and $(A) = 0.05$ (5%) is the probability that it will rain at any time today regardless of cloudiness. If $(B|A) = .95$ (95%) is the probability that it is cloudy given that it is raining, then Bayes' theorem tells us $(A|B) = 0.95 \times .05/0.5 = 0.095$ (9.5%) is the updated probability that it is raining, given that it is cloudy. Bayesian inference is the basis for much of machine learning [MacKay, 2003, Murphy, 2012], as it is allows us to quantitatively define 'learning' as an updated state of beliefs in reaction to obtaining new information.

There is a common criticism of Bayesian inference—also a criticism of the theory of subjective probability—that the prior probabilities have arbitrariness associated with choices made or not made by the modeler. Mackay argues that, as one descends into deeper layers of conditional probabilities (wherein the priors have priors), the choice of these 'hyperparameters' is less important [MacKay, 1996]. Jaynes suggests that 'it is only a subjective probability which could possibly be relevant to applications,' recognizing that this inference machinery is merely a tool for us to form reasonable judgements when faced with incomplete information [Jaynes, 1988]. Quantum Mechanics is, in some ways the most precise theory of nature, yet is difficult to interpret without invoking the observer. Therefore, it may not be so difficult for us to reconcile the essential role of the researcher in applying Bayesian inference.

1.2 Generative Modeling

We now write $p(A)$ the probability of proposition A being true, where $p(A) = 0$ means there is no possibility that A is true, and $p(A) = 1$ means A is certainly true. The inference tasks in this dissertation assume that some data d is given, and that there is some latent (to be discovered) experimental conditions or parameters θ which we wish to learn. We wish to *infer* these latent experimental conditions *given* the observation of data: $p(\theta|d)$. However, usually in an experiment we only have a model for $p(d|\theta)$, the probability of observing data d given parameters θ.

In our experiments, we assume random and uncorrelated noise is added by the measurement process, which we interpret as a probability of a specific data instance. Our generative model for data can be expressed as $d = f(\theta) + \eta$, where f is our model of the experiment and η is noise (often normally distributed in the shape of a 'bell curve'). If the probability of a particular noise sample is $p(\eta)$, then we see that $p(d - f(\theta)) \approx p(\eta)$. A generative modeling perspective includes not only the model of the experiment, but the model for the noise as well.

Given this complete description of data generation $p(d|\theta)$, Bayes' Theorem Tells us how to infer the experimental conditions via the posterior distribution

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}. \tag{1.4}$$

$p(\theta)$ is called the prior distribution, and it reflects our preference of experimental conditions before the experiment, or *a priori*. $p(d)$ is called the evidence. Ignored in most inference tasks, it is essential for learning models [MacKay, 2003] as discussed in chapter 2. The prior information $p(\theta)$ will play an essential role in chapter 3, where we appeal to symmetry in order to encode a preference for physically plausible currents. Finally, in chapter 4 we learn the best model and likelihood $p(d|\theta)$ for spin glasses using information geometry.

From the perspective of a Bayesian, $p(\theta|d)$ contains all the information to be discovered from the experiment. The peak of the 'bell curve' of $p(\theta|d)$ is the 'maximum likelihood' solution. The probability also predicts a range of likely possibilities for θ, allowing us to obtain a range of predictions.

Another subtlety of priors is a lack of a strict boundary between the prior probability $p(\theta)$ and the likelihood $p(d|\theta)$. For example, it is common in physics for theories to be translation invariant so that only relative distances matter. In such cases, it is prudent to use a translation invariant model in the likelihood. This is not any different from choosing $p(\theta) = 0$ for any condition which is not translation invariant, though this is usually much less convenient. In deep learning, this idea is so common and important that it has a name: inductive bias.

1.3 Applications of Inference

In chapter 2, our data d are two or more noisy pictures which have been shifted with respect to each other. We wish to learn the true underlying image and the shifts. This fundamental problem is important for state-of-the-art electron microscopy of very sensitive samples, which often requires combining multiple shifted noisy images. Part of the novelty of this chapter is our treatment of the problem as a statistical field theory in which the images fluctuate around the model. We applied perturbation theory to predict the errors in shift inference, explaining a longstanding problem with the large errors of these predictions. Considering the evidence $p(d)$, we followed Bayesian model selection to learn the proper amount of course-graining or image complexity.

In chapter 3, our data d are the scan of a magnetic field taken above some material with electric currents. We wish to infer the latent electric currents. This technique has been used to take pictures of topological insulators, complex oxide domains, and even obtain *local* critical temperatures in superconductors. The relationship between electric current and magnetic fields is similar to taking a picture out of focus. This problem is difficult because there is not enough information remaining in the data, so careful choice of prior information $p(\theta)$ is essential. This chapter derives and compares new priors which consider sample geometry, respect natural symmetries, and can accomodate the lack of current conservation at image boundaries.

In chapter 4, our data d are samples of states of spin glasses. We wish to infer a coarse-grained effective model which has superior interpretability. The challenge in this work is in proposing a good *effective model*, which we develop using sloppiness and information geometry. The spin glass is a toy model for complex systems with many nearly identical low-energy states. This method for finding effective models could be applied to infer more interpretable summaries of maximum entropy models of protein folding.

1.4 The Renormalization Group Idea

The Renormalization Group (RG) is a theory with impressive philosophical implications in Physics. It is the formalization of the idea that many details do not matter to the essential aspects of some physical phenomenon. For example, theories of fluid dynamics do not need specific details of the exact molecules that make up the fluid to make good predictions. RG was developed to explain the striking fact that, for example, no matter what fluid you choose—water, ethanol, liquid Helium—near the boiling point of all liquids, the heat capacities traced an identical *universal* curve. RG motivates the selection of the simplest models for explaining a given physical phenomenon.

RG allows the classification of systems into universality classes, which all collapse onto identical curves. The beauty of this fact means that one need only study one member of that class in order to understand the essential elements of all the other members. Therefore, in order to understand the universal properties of boiling liquids, one need only to study the simplest model for a ferromagnet.

Chapter 5 studies the RG of the two-dimensional Ising model, which is a model for a thin sheet of ferromagnetic material. Using normal form theory from dynamical systems, we predict details of the corrections to the universal scaling particular to a solvable 2D Ising model. We study finite size effects, and the behavior of the RG flow equations under the Legendre transform.

Chapter 2

Image Registration and Super Resolution from First Principles

2.1 Introduction

Image registration is the problem of inferring the coordinate transformation between two (or more) noisy and shifted (or distorted) signals or images. This deceptively simple process is fundamental for stereo vision [Lucas et al., 1981], autonomous vehicles [Wolcott and Eustice, 2014], gravitational astronomy [Nicholson and Vecchio, 1998], remote sensing [Inglada et al., 2007, Debella-Gilo and Kääb, 2011], medical imaging [Zöllei et al., 2003, Leventon and Grimson, 1998], microscopy [Savitzky et al., 2018], and nondestructive strain measurement [Kammers and Daly, 2013]. At the cutting edge of microscopy, imaging sensitive biological materials [Bartesaghi et al., 2015, 2014] and metal organic

The work constituting this chapter was done in collaboration with Matthew Bierbaum and James P. Sethna. This chapter is published at `arXiv:1809.05583` and has been submitted to IEEE TIP.

frameworks [Zhang et al., 2018, Zhu et al., 2017] with Transmission Electron Microscopy, requires combining multiple low-dose high-noise images, to obtain a viable signal without destroying the sample. While most techniques for registering and combining images are accurate for low noise, errors significantly larger than theoretical bounds can occur for a signal-to-noise ratio as low as 20 (noise 5% of the signal amplitude); so far a general explanation of this error has been elusive.

Much has been written about the uncertainty of shift estimations by analyzing the information theoretic limit known as the Cramer-Rao bound (CRB) [Robinson and Milanfar, 2004, Yetik and Nehorai, 2006, Pham et al., 2005]. These works observed that no known estimators achieve the CRB for image registration. This sub-optimal performance has been blamed on biased estimators: some claim interpolation errors explain the bias [Rohde et al., 2009, Schreier et al., 2000, Inglada et al., 2007, Bailey et al., 2005] and others claim that the problem is inherently biased [Robinson and Milanfar, 2004]. More works have explored non-perturbative estimations of the uncertainty, which yield larger estimates more consistent with measured error, but also rely on assumptions about the latent image [Uss et al., 2014, Ziv and Zakai, 1969, Xu et al., 2009].

Here we solve these problems by studying the naïve maximum likelihood formulation of image registration. We explore a new derivation of the standard method (comparing one image to match the other) by integrating out the underlying true image. We treat the standard method as a statistical field theory in which two images fluctuate around each other, showing that the shift uncertainty should scale quadratically with image noise ($\sigma_\Delta \propto \sigma^2$), while the naïve CRB is linear ($\sigma_\Delta \propto \sigma$). We also show that bias in image registration is due to the image edges. Our theory makes the novel prediction that coarse-graining images can dramatically improve shift precision, which we confirm numerically. While coarse-graining helps, it requires oversampled images and knowledge

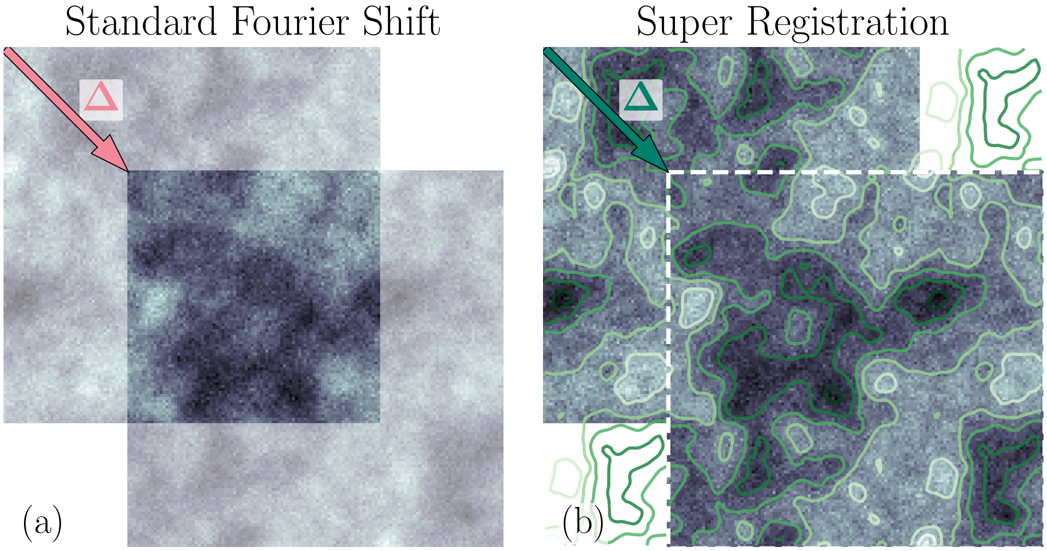


Figure 2.1: Illustrations of image registration techniques. (a) A schematic of the standard method of image registration which measures the shift Δ between noisy data (grayscale images) by shifting one to match the other. (b) A schematic of our proposed method, Super Registration, which infers the shift Δ instead by learning the underlying image I (green contours), and shifting the coordinates until the model image best fits the data (grayscale images).

of the highest frequencies of the underlying image. We overcome this limitation, and reach the true CRB, by shifting a learned model for the underlying image to match the data. We use Bayesian model selection to find the model most supported by the data, effectively learning the amount of necessary coarse-graining. We demonstrate the optimality of our new method—called Super Registration (SR)—with periodic images. We also demonstrate clear improvements in error and removal of bias for general non-periodic images with Chebyshev image models. Finally, we show that particle tracking is 10-20× more precise when performed on images combined with SR. We conclude by discussing the implications of our theory on more general nonlinear registration, and registration of images captured with different imaging modes.

2.2 Theory of image formation

In this work, image registration will be restricted to the task of inferring a rigid shift relating two (or more) discretely sampled noisy images with sub-pixel precision. More general transformations are accommodated by our subsequent arguments through application of the chain rule. Defining some true image (latent, to be discovered) intensity function $I(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^2$, we measure at least two images by sampling discretely:

$$\begin{aligned}\phi_i &= I(\mathbf{x}_i) + \xi_i \\ \psi_i &= I(\mathbf{x}_i + \boldsymbol{\Delta}) + \eta_i,\end{aligned} \tag{2.1}$$

where ϕ_i is the i^{th} pixel of image ϕ and ξ_i η_i are white noise distributed with zero mean and variance σ^2, and $\boldsymbol{\Delta}$ is the shift between the images which we intend to infer.

Equation 2.1 is our model, which we can express as the likelihood $p(\phi, \psi|\boldsymbol{\Delta}, I)$ of measuring ϕ and ψ given Δ and I:

$$p(\phi, \psi|\boldsymbol{\Delta}, I) \propto \exp\left(-\frac{1}{2\sigma^2}\left(||\phi - I||^2 + ||\psi - T_{\boldsymbol{\Delta}} I||^2\right)\right), \tag{2.2}$$

where $||x||^2 = \sum_i x_i^2$. $T_{\boldsymbol{\Delta}}$ represents the operator which translates its argument by $\boldsymbol{\Delta}$, for a continuous image $T_{\boldsymbol{\Delta}} I(\mathbf{x}) = I(\mathbf{x} - \boldsymbol{\Delta})$. We interpret this distribution as our image model fluctuating around data. Note that Eq. 2.2 accommodates multi-image registration by multiplying more products of terms comparing images to the shifted latent image I.

In order to infer $\boldsymbol{\Delta}$ after measuring the images ϕ and ψ we must reverse the conditional probability in Eqn. 2.2 using Bayes' theorem. The posterior (post-measurement) probability $p(\boldsymbol{\Delta}, I|\phi, \psi)$ of $\boldsymbol{\Delta}$ and I is

$$p(\boldsymbol{\Delta}, I|\phi, \psi) = \frac{p(\phi, \psi|\boldsymbol{\Delta}, I)p(\boldsymbol{\Delta}, I)}{p(\phi, \psi)}. \tag{2.3}$$

$p(\mathbf{\Delta}, I)$ is called the prior probability and $p(\phi, \psi)$ is called the evidence because, as we later show, it can be interpreted as the probability of our data given our choice of model. The task of inferring $\mathbf{\Delta}$ is achieved by maximizing this posterior probability. We define the maximum likelihood estimator of $\mathbf{\Delta}$ to be

$$\begin{aligned} \mathbf{\Delta}^\star &= \max\nolimits_{\mathbf{\Delta},I}\ p(\mathbf{\Delta}, I|\phi, \psi), \\ &= \max\nolimits_{\mathbf{\Delta},I}\ p(\phi, \psi|\mathbf{\Delta}, I)p(\mathbf{\Delta}, I), \end{aligned} \tag{2.4}$$

where the second line is possible because the evidence is independent of $\mathbf{\Delta}$ and I.

How accurately should we be able to measure $\mathbf{\Delta}$? If we assume we know the underlying image I, the answer is given by the Cramer-Rao bound (CRB) [Cover and Thomas, 2012]. For any parameter vector θ, the CRB of θ is $\sigma_\theta^2 \geq \theta^T g^{-1} \theta$, where the Information matrix (FIM)

$$g_{\mu\nu} = \left\langle \frac{\partial^2 \log p}{\partial\theta_\mu \partial\theta_\nu} \right\rangle. \tag{2.5}$$

The posterior $p = p(\mathbf{\Delta}, I|\phi, \psi)$ is given by Eqn. 2.3 and θ_μ are the parameters, i.e. $\mathbf{\Delta}$ and I. We can calculate the naïve CRB for image registration, assuming we know the underlying image I, and that $\partial I/\partial x$ and $\partial I/\partial y$ are uncorrelated, the smallest possible variance on the estimation of the x-direction shift Δ_x is

$$\sigma_{\Delta_x}^2 \geq \sigma^2 \Big/ \int \mathrm{d}^2\mathbf{x} \left(\frac{\partial I}{\partial x}\right)^2. \tag{2.6}$$

In other words, if the data are very noisy or if the underlying image has no features, it will be difficult to measure the shifts. Note that the CRB predicts that the shift error will scale linearly with noise ($\sigma_\Delta \propto \sigma$). We reiterate that this is the CRB of the shifts assuming knowledge of the true image I. Since this is an unrealistic assumption for real data, we call Eq. 2.6 and its discrete analog the naïve CRB. For previous derivations and discussions of the naïve CRB for image registration, see [Robinson and Milanfar, 2004, Yetik and Nehorai, 2006]. When discussing the CRB below we use the definition related to Eq. 2.5 and not the intuitive result of Eq. 2.6.

2.2.1 Deriving the standard method of image registration

In an experiment we have no access to the latent image I. We offer a new derivation of the standard method for overcoming this by marginalizing, or integrating out I:

$$p(\mathbf{\Delta}|\phi, \psi) \propto \int \mathrm{d}I \; p(\phi, \psi|\mathbf{\Delta}, I)p(I). \tag{2.7}$$

If we assume that $p(I) \propto 1$, i.e. all images are equally likely, we can perform the integral by first recognizing that $||\psi - T_{\mathbf{\Delta}} I||^2 = ||T_{-\mathbf{\Delta}}\psi - I||^2$ if $T_{\mathbf{\Delta}}$ is a unitary transformation (preserves the L2 norm). Transforming discrete data will require interpolation. Linear, quadratic, cubic, bi-cubic, and other local interpolation schemes previously studied for this problem [Rohde et al., 2009, Schreier et al., 2000, Inglada et al., 2007, Bailey et al., 2005] are not unitary—neatly explaining some of their observed bias. In this work we will consider only unitary interpolation by using Fourier shifting, however our ultimate solution will obviate this discussion by directly employing Eq. 2.2. Now the posterior $p(\mathbf{\Delta}|\phi, \psi)$ is a product of integrals of the form

$$\int \mathrm{d}x e^{-\frac{1}{2\sigma^2}\left((x-a)^2+(x-b)^2\right)} \propto \exp\left(-\frac{(a-b)^2}{4\sigma^2}\right). \tag{2.8}$$

Applying this to each pixel in the data we arrive at the marginal likelihood

$$p(\mathbf{\Delta}|\phi, \psi) \propto \exp\left(-\frac{1}{4\sigma^2}||\psi - T_{-\mathbf{\Delta}}\phi||^2\right). \tag{2.9}$$

We have derived the standard least-squares similarity measure (it is usually written down intuitively), in which one simply shifts one image until it most closely matches the other. This process is illustrated by Fig.2.1(a), which shows a pair of synthetic data which will serve as I in our numerical studies of periodic registration. It was calculated by sampling a 64×64 image from a power law in Fourier space

$$P(|I(k)|) \sim k^{-1.8} e^{-\frac{1}{2}\left(\frac{k}{k_c}\right)^2}, \tag{2.10}$$

damped by a Gaussian with scale $k_c = k_{\text{Nyquist}}/3$ to ensure a smooth cutoff approaching the Nyquist limit, preventing aliasing.

Notice that if T_Δ is not unitary that this objective is different depending on whether you shift one measured image or the other. Note also that in general image registration this inverse transformation may not exist; in such cases this method will fail. The literature features multiple implementations of Eq. 2.9 using Fourier interpolation by either shifting the data [Jacovitti and Scarano, 1993] or upsampling by padding in Fourier space and finding the maximum cross-correlation [Guizar-Sicairos et al., 2008]. The latter method can only be as accurate as the factor of upsampling, e.g. quadrupling (in $2D$) the number of Fourier modes allows evaluating shifts of half a pixel. While sophisticated extrapolations have been used to overcome the arbitrary choice of how much to upscale, we will exactly shift the data and optimize Eqn. 2.12 directly. Writing the $2D$ Fourier transform operator as $\mathcal{F}$, we implement $T_{\mathbf{\Delta}}\phi$ as:

$$T_{\mathbf{\Delta}}\phi = \mathcal{F}^{-1} e^{-i\mathbf{k}\cdot\mathbf{\Delta}} \mathcal{F}\phi \tag{2.11}$$

Another important result of our theory is the $4\sigma^2 = (2\sigma)^2$ in the denominator of Eq. 2.9: this likelihood function is for data with twice the variance of our original problem, which is consistent with taking the difference of two noisy signals. Some of the reported discrepancy ($\sqrt{2} \sim 40\%$) between the CRB and observed error [Robinson and Milanfar, 2004, Aguerrebere et al., 2016, Yetik and Nehorai, 2006] can be explained by the absence of this factor. Those studying multi-image registration have also neglected this modification of the noise fluctuations in their estimating of shift precision [Aguerrebere et al., 2016]. We have obtained by integrating out the latent image I a distribution which depends only on our data ϕ and ψ and the unknown shift $\mathbf{\Delta}$. We can now define $\mathbf{\Delta}_m^\star$, the marginal maximum likelihood (ML) solution, which we will now refer to as the standard Fourier shift (FS) method:

$$\mathbf{\Delta}_m^\star = \max_{\mathbf{\Delta}} p(\mathbf{\Delta}|\phi, \psi) = \min_{\mathbf{\Delta}} ||\psi - T_{-\mathbf{\Delta}}\phi||^2. \tag{2.12}$$

This new derivation of the standard method of image registration highlights and clarifies some important limitations. Only unitary (L2-preserving) interpolation for shifting images will lead to unbiased shift estimation, otherwise we are simply optimizing a corrupted likelihood. Second, comparing the squared error between shifted images is only correct if the noise in the images is Gaussian. If we were studying images with Poisson-distributed noise, for instance, the likelihood in Eqn. 2.2 should be a Poisson distribution. The standard method is often successfully employed for non-Gaussian noise. We do not doubt its efficacy, but instead claim that the standard method cannot be optimal in this case because it violates the implicit assumptions of Gaussian noise.

2.3 Statistical properties of the standard method

It is well documented in the literature that the errors in shift inference via FS are much larger than the naïve CRB. Figure 2.2 shows the noise-averaged error (pink dots) of inferring the shifts as measured using the standard Fourier shift method in Eq. 2.12. The measured error grows quadratically with the Gaussian additive noise σ, dwarfing The naïve CRB (shaded pink region). The follow section will derive a theory (black dotted) to predict this quadratic error growth.

Say we measure the fields ψ_i and ϕ_i, then the log-marginal posterior is (up to a constant) proportional to

$$\mathcal{L} = \frac{1}{2}\sum_i (\psi_i - T_{-\Delta}\phi_i)^2 = \frac{1}{2}\sum_k |\widetilde{\psi}_k - e^{ik\Delta}\widetilde{\phi}_k|^2, \tag{2.13}$$

where $\widetilde{\phi}_k$ and $\widetilde{\psi}_k$ are the Fourier transforms of our data. Our measurements fluctuate around the true latent image I according to

$$\begin{aligned} p(\psi) &\propto \exp\left(-\frac{1}{2\sigma^2}||\psi - I(x)||^2\right), \\ p(\phi) &\propto \exp\left(-\frac{1}{2\sigma^2}||\phi - I(x-\Delta_0)||^2\right), \end{aligned} \tag{2.14}$$

where Δ_0 is the latent shift and σ^2 is the variance of the noise. Near the true shift Δ_0 we can expand the marginal likelihood as

$$\mathcal{L}(\Delta) = \mathcal{L}(\Delta_0) + (\Delta - \Delta_0)\frac{\partial \mathcal{L}}{\partial \Delta} + \frac{1}{2}(\Delta - \Delta_0)^2\frac{\partial^2 \mathcal{L}}{\partial \Delta^2} + \dots, \tag{2.15}$$

which is approximately minimized by

$$\Delta - \Delta_0 = -\frac{\partial \mathcal{L}}{\partial \Delta}\Big/\frac{\partial^2 \mathcal{L}}{\partial \Delta^2} = -i\frac{\sum_k k\ \widetilde{\psi}_k e^{-ik\Delta_0}\widetilde{\phi}_{-k}}{\sum_k k^2\ \widetilde{\psi}_k e^{-ik\Delta_0}\widetilde{\phi}_{-k}}. \tag{2.16}$$

We can calculate the error of the standard method by averaging Eqn. 2.16 and its square over the distributions in Eqn. 2.14.

2.3.1 Bias of the standard method (1D)

Writing Eqn. 2.16 as A/B we can Taylor expand about $A = \langle A\rangle$ and $B = \langle B\rangle$, then average over the noise to find

$$\left\langle \frac{A}{B} \right\rangle = \frac{\langle A\rangle}{\langle B\rangle}\left(1 + \frac{\mathrm{var}(B)}{\langle B\rangle^2}\right) - \frac{\mathrm{cov}(A,B)}{\langle B\rangle^2} + \dots, \tag{2.17}$$

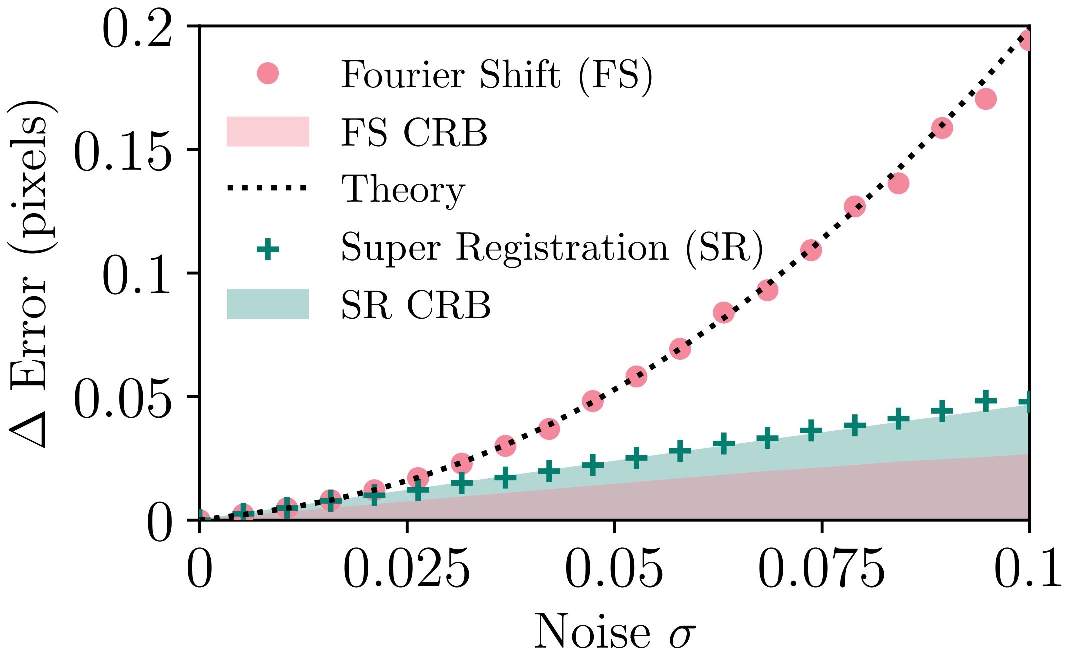


Figure 2.2: Comparing the noise-averaged errors of the inferred shift Δ measured by the standard Fourier Shift method and Super Registration in the case of aligning synthetic periodic images. For each noise level, we generate an ensemble of 1000 64×64 images statistically similar to Fig. 2.1 ($I(k) \sim k^{-1.8}$), measuring the average error for both methods, along with the minimum expected error, CRB. The error of the standard method (pink dots) grows quadratically with noise, whereas the naive CRB (pink shaded region) predicts a linear relationship. Our theory (black dashed line) accurately describes the quadratic dependence in the error, matching numerical experiments. Super Registration (green pluses) demonstrates much lower error, recovers the linear relationship between error and noise, and reaches its CRB (green shaded region).

where $\langle\cdot\rangle$ denotes integration over the distributions of Eqn. 2.14. Notice that

$$\langle A\rangle = \left\langle \sum_k k\widetilde{\phi}_k e^{-ik\Delta_0}\widetilde{\psi}_{-k} \right\rangle = \cancelto{0}{\sum_k kI_kI_{-k}}, \tag{2.18}$$

which is zero because the summand is odd in k. Therefore the average bias for periodic images is to lowest order

$$\left\langle \frac{A}{B} \right\rangle = -\frac{\langle AB\rangle}{\langle B\rangle^2}. \tag{2.19}$$

In general for non-periodic images $\langle A\rangle \neq 0$. Examining the continuum limit of $\langle A\rangle$ in real space, we find

$$\begin{aligned} \langle A\rangle &= \int \mathrm{d}x\ I\frac{\partial I}{\partial x} \\ &= \frac{1}{2}\int \mathrm{d}x\frac{\partial}{\partial x}I^2 = \frac{1}{2}\left(I(x_N)^2 - I(x_0)^2\right), \end{aligned} \tag{2.20}$$

where x_N and x_0 are the endpoints of the domain; $\langle A\rangle$ is a total derivative depending only on the edges of the image. Therefore we hypothesize that the bias of the standard FS method of image registration shown in Fig. 2.5 will be dominated by the edges of the data. Ziv and Zakai in 1969 [Ziv and Zakai, 1969] and others [Robinson and Milanfar, 2004, Nicholson and Vecchio, 1998], share this speculation, however, whereas they argued that impingement of shift fluctuations onto the limits of the domain caused bias, our theory suggests that structures of the edges of images themselves cause bias.

Evaluating the remaining moments of Eq. 2.19 we find

$$\langle B\rangle = \sum_k k^2 I_k I_{-k}, \tag{2.21}$$

which is the roughness of the latent image I, found in the denominator of the naïve CRB in Eqn. 2.6. The last correlation for the average bias is

$$\langle AB\rangle = \sum_{kk'} kk'^2\ e^{-i(k+k')\Delta_0}\langle\widetilde{\psi}_k\widetilde{\psi}_{k'}\rangle\langle\widetilde{\phi}_{-k}\widetilde{\phi}_{-k'}\rangle, \tag{2.22}$$

which can be evaluated using the moments

$$\langle\widetilde{\psi}_k\rangle = I_k, \qquad \langle\widetilde{\phi}_k\rangle = e^{-ik\Delta_0} I_k, \tag{2.23}$$

$$\langle\widetilde{\psi}_k\widetilde{\psi}_k\rangle = I_k I_k, \qquad \langle\widetilde{\phi}_k\widetilde{\phi}_k\rangle = e^{-ik2\Delta_0} I_k I_k, \tag{2.24}$$

$$\langle\widetilde{\psi}_k\widetilde{\psi}_{-k}\rangle = I_k I_{-k} + \sigma^2, \qquad \langle\widetilde{\phi}_k\widetilde{\phi}_{-k}\rangle = I_k I_{-k} + \sigma^2. \tag{2.25}$$

Considering the sum in Eqn. 2.22 in three cases $k' = -k$, $k' = k$ and $k' \neq \pm k$ we can apply the moments to find

$$\begin{aligned}\langle AB\rangle = \sum_k \Big(k^3 \left((I_k I_{-k} + \sigma^2)^2 + (I_k I_{-k})^2\right) + \\ k \sum_{k' \neq \pm k} k'^2 (I_k I_{-k})^2\Big) = 0,\end{aligned} \tag{2.26}$$

from which we conclude the entire correlation function vanishes due to each term of the summand being odd in k. Further, numerical evidence and inspection of higher order terms in the expansion of Eq. 2.17 support the conclusion that for periodic images the standard Fourier shift method of image registration is unbiased.

2.3.2 Variance of the standard method (1D)

Turning our attention to the variance or expected error of the bias given by Eq. 2.16; an expansion and average of $(A/B)^2$ (simplifying for $\langle A\rangle = 0$) yields to lowest order

$$\mathrm{var}\left(\frac{A}{B}\right) = \frac{\langle A^2\rangle}{\langle B\rangle^2}. \tag{2.27}$$

Equation 2.21 gives us $\langle B\rangle$, so we need only to compute the correlation function $\langle A^2\rangle$:

$$\begin{aligned}\langle A^2\rangle &= -\sum_k\sum_{k'} kk' e^{-i(k+k')\Delta_0}\langle\widetilde{\psi}_k\widetilde{\psi}_{k'}\rangle\langle\widetilde{\phi}_{-k}\widetilde{\phi}_{-k'}\rangle\\ &= -\cancelto{0}{\sum_k\sum_{k'\neq k} kk'|I_k|^2|I_{k'}|^2}\\ &\quad +\sum_k k^2\Big((I_kI_{-k}+\sigma^2)^2-(I_kI_{-k})^2\Big), \end{aligned} \tag{2.28}$$

where as before we have decomposed the sum into terms for which $k'\neq k$, $k'=-k$ and $k'=k$. We find that the variance of the bias (which is also the variance of the estimated shifts since we have shown $\langle\Delta\rangle=\Delta_0$) is approximately

$$\sigma_\Delta^2=\Big\langle(\Delta-\Delta_0)^2\Big\rangle=2\frac{\sigma^2}{D^2}+\frac{L\pi^2}{3}\frac{\sigma^4}{D^4}, \tag{2.29}$$

where $D^2=\sum_k k^2I_kI_{-k}$ is the roughness of the image. We used the fact that $\sum_k k^2=(2+L^2)\pi^2/3L\approx L\pi^2/3$ for a one-dimensional signal with L points. The lowest order term in Eq. 2.29 is twice the naïve CRB shown in Eq. 2.6, consistent with the fact that the marginal posterior in Eq. 2.9 has twice the variance of the noise. We have shown that the standard Fourier shift method cannot achieve the naïve CRB. Notice that the variance grows beyond the CRB at a rate proportional to σ^4 and the image size L, so that error grows quadratically with noise. This extra factor of the image volume means that sampling a band-limited (sampled below the Nyquist limit) image at a higher rate—increasing the resolution without increasing information content—can actually decrease the registration precision for the standard Fourier shift method. We discuss and verify this observation following an extension of this theory to two-dimensions.

2.3.3 Variance of the standard method in two dimensions

Generalizing our expansion of the marginal likelihood we find

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Delta}) =&\mathcal{L}(\boldsymbol{\Delta}_0) + (\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)^T \nabla \mathcal{L} \\ &+ \frac{1}{2}(\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)^T \; \nabla^2 \mathcal{L} \; (\boldsymbol{\Delta} - \boldsymbol{\Delta}_0) + \ldots,\end{aligned} \tag{2.30}$$

from which we conclude that the two-dimensional analogue of Eq. 2.16 is

$$\boldsymbol{\Delta} - \boldsymbol{\Delta}_0 = -\left(\nabla^2 \mathcal{L}\right)^{-1} \nabla \mathcal{L}. \tag{2.31}$$

If the off-diagonal terms of the Hessian $\nabla^2 \mathcal{L}$ are small compared to the diagonal terms (the image is approximately isotropic), the two dimensions decouple into an application of Eq. 2.29 for each dimension. This is generally a good approximation except for contrived data. In this case we find the precision of two-dimensional image registration is approximately

$$\left\langle (\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)^2 \right\rangle = \begin{pmatrix} 2\frac{\sigma^2}{D_x^2} + \frac{N\pi^2}{3}\frac{\sigma^4}{D_x^4} \\ \\ 2\frac{\sigma^2}{D_y^2} + \frac{N\pi^2}{3}\frac{\sigma^4}{D_y^4} \end{pmatrix}, \tag{2.32}$$

where N is the number of pixels in the one of the measured images, and $D_x = \sum_{\mathbf{k}} k_x^2 I_{\mathbf{k}} I_{-\mathbf{k}}$ and $D_y = \sum_{\mathbf{k}} k_y^2 I_{\mathbf{k}} I_{-\mathbf{k}}$ are the horizontal and vertical image roughness. Eq. 2.32 is used in Fig. 2.2 (black dotted) where we see excellent agreement with the numerically measured error (pink dots). The excellent agreement—in spite of ignoring the cross terms—can be explained by expanding Eq. 2.31 for small values of the off-diagonal terms: the lowest order correction averages to zero.

Our analysis has shown that the error of shift estimates of the standard Fourier shift method grow much faster than the CRB. Why do the errors scale quadratically with noise? Mackay found that in general and especially for ill-posed problems (like distinguishing noise from signal), integrating over parameters can yield distributions with stretched and skewed peaks, biasing the maximum and leading to large errors [MacKay, 1996]. We integrated over all possible images in order to derive the standard FS registration method. Did this choice sabotage our effort to achieve the ultimate precision? For exponential functions (like a Gaussian or our likelihoods above), there is a deep relationship between optimization and integration through Laplace's method or the method of steepest descent [De Bruijn, 1970]. By integrating over all possible images, we essentially maximized $\log p(\phi, \psi | I, \boldsymbol{\Delta})$ over I—estimating the latent image—and used that estimate for predicting the shift. This estimate is, however, unreliable as it makes no distinction between the signal and the noise. The high frequency modes of the data, dominated by noise and ironically most discriminating for shift localization, cause the fluctuation of our inferred shifts to be much larger than the CRB. This is illuminated by the following section which considers the process of coarse-graining or binning image data.

2.3.4 Coarse Graining Data can Improve Precision

Our theory for the variance of the shift predicts that $\sigma_{\Delta}^2 = 2\frac{\sigma^2}{D^2}\left(1 + \frac{N\pi}{6}\frac{\sigma^2}{D^2}\right)$. The factor of the image volume N in the correction term inspired us to consider reducing N without changing σ or D^2. Coarse-graining the data by some linear factor a—shown schematically in Fig 2.3(a)—should not change the CRB assuming the latent image I is smooth on that length scale (or, equivalently, assuming that the data is sampled at least a-times the Nyquist frequency). Assuming that each pixel of the data has noise of variance σ^2, the variance of noise for each $a \times a$ block should be $a^2\sigma^2$ (variances of uncorrelated noise add). The denominator of the naïve CRB $D^2 = \sum_k k^2 |I_k|^2$ is subtler: the amplitude of each pixel increases by a factor of a^2 ($I_k \to a^2 I_k$), and the block sum only removed Fourier modes with zero amplitude by our assumption above, so $D^2 \to a^4 D^2$. Finally the coarse-grained image will have its coordinates expanded by a, so that the variance should be rescaled by a^2. Therefore coarsening should modify our variance prediction of the Fourier shift method accordingly:

$$\begin{aligned}\sigma_\Delta^2 &= a^2 \cdot 2\frac{a^2\sigma^2}{a^4D^2}\left(1 + \frac{\pi N/a^2}{6}\frac{a^2\sigma^2}{a^4D^2}\right) \\ &= 2\frac{\sigma^2}{D^2}\left(1 + \frac{\pi N}{6a^4}\frac{\sigma^2}{D^2}\right). \end{aligned} \tag{2.33}$$

Our theory predicts that coarse-graining over-sampled images can improve shift inference by reducing the correction term, but that the method can at best yield a variance equal to twice the naive CRB. This result may explain improvements in registration precision from re-binning image intensities observed in other works [Hutton and Braun, 2003, Pekin et al., 2017]. Figure 2.3(b) confirms the predicted relationship, where the black dots indicate the variance of a $N = 1024^2$ image which was oversampled by a factor of 20. Each lighter colored dot series is the variance after coarsening by some factor a, and the solid lines are given by Eq. 2.33. We see excellent agreement with our theory, and a convergence of the variances onto the $2\sigma^2/D^2$ line. Note that the original image ($a = 1$) variances differ from our theory for large noise: perhaps the limits of large images and large noise are where our approximations in truncating the Taylor expansion in Eq. 2.27 breaks down.

Coarsening smooth images only throws away information which is dominated by noise. When we use the coarsened images in the standard FS method, we implicitly estimate the underlying image but with less noisy modes, and will get a more reliable estimate. In a real experiment without knowledge of the true length scale of the image, we will not know the optimal coarsening length scale. In the following section we propose our generative model which will use Bayesian model selection to infer the image complexity supported by the data.

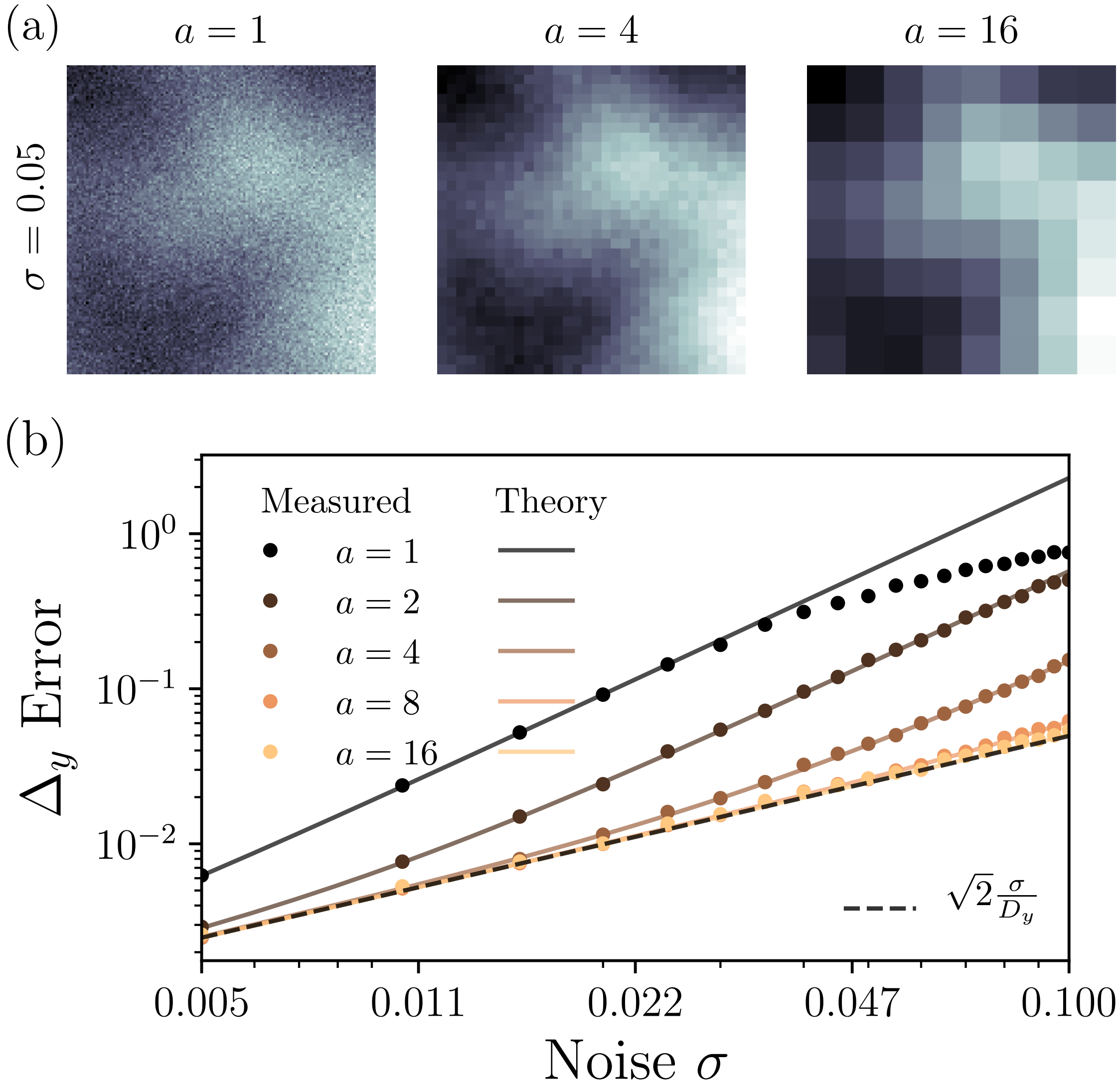


Figure 2.3: (a) An oversampled 1024^2 image (the image varies on a scale 20× smaller than the Nyquist frequency limit) with 5% additive white Gaussian noise then coarse-grained by summing over $a \times a$ blocks. Shown are $a = 1$, $a = 4$, and $a = 16$, representing a drastic reduction in image size while not removing any information which localizes the shifts between images. (b) The error in inferred shifts (dots) for the standard Fourier shift method applied to the image after coarsening by 1, 2, 4, 8, and 16 blocks. The original image was chosen to be smooth enough so that coarsening by a factor of 16 would not violate the Nyquist sampling theorem. The solid lines are the prediction of our theory, and the dotted line is $\sqrt{2}$ times the naïve CRB, $\sqrt{2}\sigma/D_y$.

2.4 Super registration

How can we achieve the ultimate precision for image registration as predicted by the CRB? We have seen that the standard FS method of image registration which directly compares two images has a variance in its shift prediction of the form $\sigma_{\Delta}^2 = 2\sigma_{\mathrm{CRB}}^2(1 + N\pi\sigma_{\mathrm{CRB}}^2/6)$, where the CRB is $\sigma_{\mathrm{CRB}}^2 = \sigma^2/\sum_k k^2 I_k I_{-k}$. We are still studying periodic images, so it is natural to consider removing noise with a filter like the optimal Wiener filter. This manifests by modifying our log-marginal likelihood in Eq.2.13 with the rule $\widetilde{\psi}_k \to A_k\widetilde{\psi}_k$ and $\widetilde{\phi}_k \to A_k\widetilde{\phi}_k$, for some filter function A_k. This modification simply changes $\sigma_{\mathrm{CRB}}^2 \to \sigma^2/\sum_k k^2 A_k I_k A_{-k} I_{-k}$, and since $A_k A_{-k} \leq 1$ (a filter only reduces power), this can only increase σ_{CRB}^2 and thus reduce our precision.

Faced with this fact we abandon the standard method of image registration and return to first principles by studying the likelihood defined by the image formation model in Eq. 2.2. Instead of shifting the data, we will model the image and shift that, as shown schematically in Fig. 2.1. This method will result in a de-noised and, depending on the data, a super-resolution estimate of the latent image. Inspired by the inextricable relationship between registration and super-resolution that we have discovered, we call our new method Super Registration (SR). Our success depends on using all that Bayesian inference has to offer, and so we proceed with a discussion of evidence-based model selection.

2.4.1 Bayesian inference and model selection

Following Mackay's discussion on integration versus optimization in inference with hyperparameters we will choose a model space and from this select the best model by comparing the model evidence, $p(\phi, \psi)$. The evidence is simply the normalization constant of the posterior Eq. 2.3; its utility for selecting the best model can be exposed by a seemingly erudite increase in notational complexity which makes manifest more of the assumptions in our model. Consider a model of image formation for the case of periodic image registration, expressed as the likelihood of measuring two images $p(\phi, \psi|\mathbf{\Delta}, I)$. Now that we are optimizing over I instead of integrating, we must choose some parameterization $I \in \mathcal{H}$ where $\mathcal{H}$ is some space of image models, e.g. a Fourier series or sums of polynomials. This choice must be reflected in the conditionals of our probabilities, so that the likelihood of measuring ϕ and ψ must now be written $p(\phi, \psi|\mathbf{\Delta}, I, \mathcal{H}_\lambda)$, where $\mathcal{H}_\lambda$ represents a specific choice of image model.

Proceeding with the inference task at hand by writing again (with our new notation) the result of Bayes' theorem shown in Eq. 2.3 we see that the posterior now reads

$$p(\mathbf{\Delta}, I|\phi, \psi, \mathcal{H}_\lambda) = \frac{p(\phi, \psi|\mathbf{\Delta}, I, \mathcal{H}_\lambda)p(\mathbf{\Delta}, I|\mathcal{H}_\lambda)}{p(\phi, \psi|\mathcal{H}_\lambda)}. \tag{2.34}$$

The solution to our problem still lies in studying this posterior distribution, but we now must also infer the best model $\mathcal{H}_\lambda$. We again apply Bayes' theorem, finding the probability that our model is true given our measured images

$$p(\mathcal{H}_\lambda|\phi, \psi) \propto p(\phi, \psi|\mathcal{H}_\lambda)p(\mathcal{H}_\lambda). \tag{2.35}$$

We have explicitly ignored the normalization constant $p(\phi, \psi)$ [1]. Assuming we have no prior preference for some models over others, $p(\mathcal{H}_\lambda) \sim 1$, so inferring which model is most likely given the data is equivalent to maximizing $p(\phi, \psi|\mathcal{H}_\lambda)$, which is the normalization of Eq. 2.34.

Therefore Bayesian inference for image registration consists of the following steps given some data ϕ and ψ.

[1] $p(\phi, \psi) = \sum_i p(\phi, \psi|\mathcal{H}_i)p(\mathcal{H}_i)$. This constant changes when we consider more models, which naturally must happen when we obtain more data, but does not influence the preference of one model over another.

1. Choose some model $\mathcal{H}_\lambda$ and evaluate Eqn. 2.34, the posterior $p(\boldsymbol{\Delta}, I|\phi, \psi, \mathcal{H}_\lambda)$.
2. Summarize the posterior by calculating the position and widths of the maximum likelihood $\boldsymbol{\Delta}$ and I.
3. Evaluate Eqn. 2.35, the model evidence $p(\phi, \psi|\mathcal{H}_\lambda)$, by estimating the normalization of the posterior.
4. Repeat steps 1-3 with some subset of the model space $\mathcal{H}$.
5. Choose the model $\mathcal{H}_\lambda$ with the largest evidence and examine its concomitant posterior distribution.

The final (unlisted) step is to examine and decide whether the residuals and the maximum likelihood image and shifts are reasonable.

This recursive process of acknowledging all the context and condition of our model and inverting them with Bayes' theorem can go on forever. We could for instance consider a probability over the parameters θ of our model $\mathcal{H}_\lambda(\theta)$, adding another integration or optimization to the steps above. Fortunately, the deeper these model assumptions go, the less these decisions affect the outcome of our inference [MacKay, 1996]. Bayesian inference does not exclude the experience of the researcher; we will terminate the inference recursion with our own judgement.

2.4.2 Super Registration for periodic images

Returning to our periodic image registration problem, let us pursue the inference steps above in a concrete example. The natural model space for periodic images consists of Fourier series, indexed by the maximum frequency allowed. Given two images ϕ and ψ, the probability of measuring these images given some latent image I and shift $\mathbf{\Delta}$ is

$$\log p(\phi, \psi | \mathbf{\Delta}, I, \mathcal{H}_\lambda) = -\frac{1}{2\sigma^2} \sum_{k=0}^{\lambda} |\phi_k - I_k|^2 + \\ |\psi_k - e^{-i\mathbf{k}\cdot\mathbf{\Delta}} I_k|^2 \\ - \log Z_L, \tag{2.36}$$

where λ indexes the complexity of the model and ϕ_k, ψ_k, I_k are the components of the Fourier transforms of our image model, and $\mathcal{Z}_L$ is the normalization. Assuming a constant prior on shifts and images, the maximum likelihood of the shifts and image is the solution of

$$\mathbf{\Delta}_{\mathrm{ML}}, I_{\mathrm{ML}} = \min_{\mathbf{\Delta},I} \sum_{k=0}^{\lambda} |\phi_k - I_k|^2 + |\psi_k - e^{-i\mathbf{k}\cdot\mathbf{\Delta}} I_k|^2. \tag{2.37}$$

Equation 2.37 is in the standard form of a nonlinear least square problem which we solve by alternating linear least squares for I_k and using Levenberg-Marquardt for $\mathbf{\Delta}$. For a given image model $\mathcal{H}_\lambda$ we can find the most likely shift and image by evaluating Eq. 2.37, calculate the covariance, and compute the evidence. Assuming flat priors on $\mathbf{\Delta}$, I_k and $\mathcal{H}_\lambda$ the evidence is the integral of our likelihood over $\mathbf{\Delta}$ and I:

$$\mathcal{Z}_L = \int \mathrm{d}I_k \mathrm{d}\mathbf{\Delta}\ p(\phi, \psi | \mathbf{\Delta}, I, \mathcal{H}_\lambda). \tag{2.38}$$

$\mathcal{Z}_L$ can be computed by applying Laplace's method of integration using the Jacobian of the least squares problem.

Figure 2.4 shows the result of step 4 of our algorithm for the periodic data used in all numerical experiments so far (shown in Fig. 2.1), where have used every possible Fourier cutoff. We have inverted the evidence to guide the eye, so that the minimum of the black curve is the most likely model. For this true image and noise level the most likely model is $\lambda = 15$ (15×15 sinusoids). The smallest observed error (green crosses) in shift inference

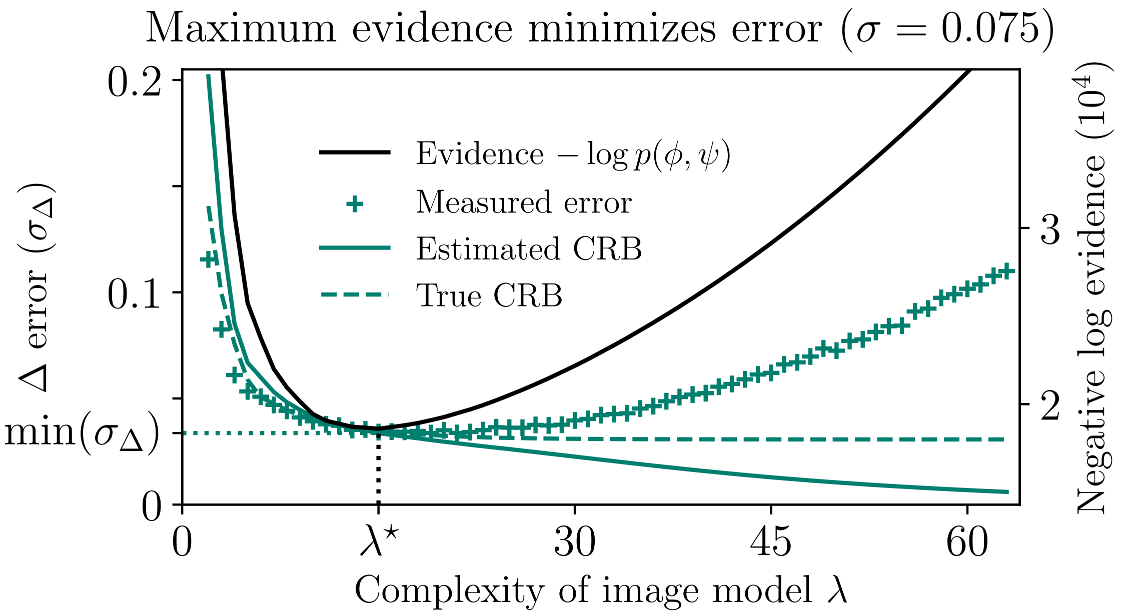


Figure 2.4: Using 1000 pairs of 64×64 images with additive Gaussian noise and $I(k) \sim k^{-1.8}$, we computed the model evidence $p(\mathcal{H}_\lambda|\phi, \psi)$ (black curve) for all Fourier cutoffs indexed by λ, showing that when the evidence is maximized the actual shift error (green crosses) is minimized. Further, this error is nearly indistinguishable from the CRB (green dashed). Finally, the naïve estimate of the CRB (solid green) is computed from the curvature of the posterior using Eqn 2.5 the Fisher Information. During a real experiment only the evidence (black curve) and the naïve curvature estimate of the CRB (solid green) are available, but when the evidence is maximized all estimate of the error match.

is also precisely at $\lambda = 15$, and is consistent with the CRB (green dashed). The most likely model provides the most precise inference of the shifts. The maximum evidence solution has been interpreted to embody Occam's Razor that the simplest explanation is most likely [Balasubramanian, 1997]. Therefore evidence-based model selection can systematically infer the number of degrees of freedom as supported by the data, avoiding over-fitting and larger errors than the CRB.

The solid green line of Fig. 2.4 is the CRB estimated by evaluating the second derivative of the log-likelihood; notice that this erroneously continues to decrease with increasing complexity. In a real experiment we only have access to the evidence (solid black line) and this curvature estimate of the CRB (solid green line). The maximum evidence model is also where all of our estimates of the shift error, motivating further the utility of the evidence-based choice of model complexity. Finally note that when the complexity is chosen to be 64 (or all Fourier modes are used) the measured error $\sigma_\Delta \approx 0.1$. In Fig. 2.1, when the noise is $\sigma = 0.075$, the same as in the evidence experiment above, the observed error of the standard FS method is also $\sigma_\Delta \approx 0.1$. Therefore we see numerical correspondence between integration over the underlying image and optimization without selecting model complexity by considering the evidence.

2.4.3 General non-periodic Super Registration

Following the clarity of studying image registration in the periodic case, we turn our attention to general non-periodic images. Here there is no clearly natural model; images are extremely complicated. While there are exciting candidates in the form of deep convolutional neural networks, these objects cannot (currently) be evaluated at arbitrary points in space; they have no notion of continuous locality [Ulyanov et al., 2017]. In general the researcher's knowledge about the physical objects being imaged should inspire the model space. A very specific and successful example is the Parameter Extraction by Modeling Images (PERI), which modeled almost every aspect of a confocal microscope, extracting enough information from a light microscope to infer the parameters of the van der Waals interaction [Bierbaum et al., 2017]. Lacking such specific inspiration therefore we chose sums of Chebyshev polynomials, in part because of their excellent approximation properties [Press et al., 2007].

We generated non-periodic data from the same distribution in Eq. 2.10, sampled twice as large (128×128), shifted one by $\mathbf{\Delta}_0$, cropped out a 64×64 region, and added noise.

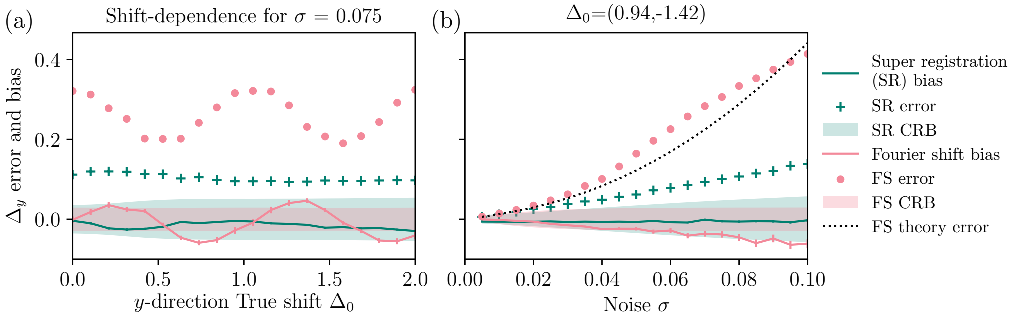


Figure 2.5: Comparing the error and bias of the standard Fourier shift (FS) method and Super Registration (SR) for non-periodic data. The synthetic data were generated by the model $I(k) \sim k^{-1.8}$, twice as large as necessary, Fourier shifted and then cropped to produce non-periodic images. Errors and biases were measured from 1500 64×64 noise samples. (a) The Δ_y biases, errors, and CRBs for the standard FS (pink) and SR (green) are shown as a function of the true real shift Δ_0. The standard method suffers from errors (pink dots) and bias (pink line) that are periodic in Δ_0. Super Registration shows almost zero bias (green line) and no periodic structure in the error (green crosses). Similarly to the periodic case, SR is much closer to its CRB (green shaded) than the standard FS method is to its CRB (pink shaded). (b) The biases, errors and CRBs for FS and SR methods as a function of noise for a fixed random shift $\Delta_0 = (0.94, -1.42)$. The standard FS method has super-linear error (pink dots) growth with noise, and a monotonic bias (pink line) large than its CRB (pink shaded). Super Registration has linear error (green cross) growth about twice its CRB (green shaded), and a bias (green line) consistent with zero.

Figure 2.5 show results for the error (pink dots and green crosses) and bias (pink and green lines) using these synthetic data, as a function of both noise σ (Fig. 2.5(a)) and true shift $\boldsymbol{\Delta}_0$ (Fig. 2.5(b)). Pink denotes the standard FS method and green denotes Super Registration. Figure 2.5(a) shows that the standard FS method has an oscillating bias which is zero at whole and half-pixels, and has an oscillating error which is largest at whole pixel shifts and smallest at half pixel shifts. The pink shaded region is the CRB of the FS method. Figure 2.5(b) shows super-linear error (pink) growth for FS, compared with our theory from Eq. 2.32 (black dotted), and a bias (pink line) deviating slowly but consistently from zero.

Figure 2.5(a) shows that Super Registration has nearly a constant bias (green line) and error (green crosses) as a function of true shift $\mathbf{\Delta}_0$, and bias smaller its CRB (green shaded). The error is much smaller than the standard FS method, and is one-third the error of the FS method when $\sigma = 0.1$ (10% noise). Finally we see in Fig. 2.5(b) that the error of SR grows linearly with noise. While SR here does not reach the CRB, it scales the same as the CRB. A better image model should result in errors more consistent with the CRB. Because we generated data by randomly sampling in Fourier space, shifting, then cropping, our Chebyshev polynomials cannot perfectly represent that signal. This is an important reminder that the CRB depends on the chosen model. Since the CRB is defined as the inverse of the Fisher Information in Eq. 2.5, the CRB is model-dependent, and thus the standard FS method and SR have different bounds.

How would Super Registration perform on data which has non-Gaussian noise? We cannot guarantee optimal precision in this case, because our model assumes the noise is Gaussian. SR would provide reliable results, however, in the same way that the FS standard method provides reliable results in this case. We can claim this because optimization (SR) and integration (FS) are the same—following the method of steepest descent or Laplace's method of integration—so that a fully complex image model (one degree of freedom for each pixel) would be statistically the same as shifting one image to match the other. The evidence maximization procedure, however, is not guaranteed to be effective, as we know the model assumes the incorrect noise distribution.

For many experimental images, Super Registration offers only a marginal improvement in the image quality as measured by eye. For a small shift error $\mathbf{\Delta} - \mathbf{\Delta}_0$ the image intensity reconstruction error is $\Delta I \approx (\mathbf{\Delta} - \mathbf{\Delta}_0) \cdot \vec{\nabla} I$. For smooth, highly sampled images visual changes will be small. Most experiments do not operate in the regime where they are not sampling at a high enough rate to see the structure of their sample. Although the reconstructions for many experiments will not vary dramatically visually, we show that the shift errors can dramatically interfere with the information extracted from the reconstructions. When inferring parameters from data such as object sizes, positions, and orientations, correlation functions, and local contrast, the precision of these quantities will be limited by the quality of the registration. To emphasize the scale of these errors, in the next section we demonstrate a dramatic improvement in particle position inference from correctly registered images.

2.5 Particle tracking errors

A very common task in image processing is tracking particle positions. High precision, especially in atomic-scale TEM and STEM, is important for understanding real-space structure. For example, charge density waves cause atoms to deviate from their lattice by tiny amounts, and can be studied by carefully measuring the positions of the atoms in real space [El Baggari et al., 2018]. For High-angle Annular Dark Field (HAADF) STEM, the image of an atom is well-approximated by a 2D Gaussian [Yankovich et al., 2014]. In TEM and STEM, noise is often Poisson-distributed. Both SR and the standard method assume image noise is Gaussian, and achieving optimalty for Poisson noise will require modeling the noise correctly by modifying the likelihood in Eqn. 2.2. Assuming Gaussian noise, then, we created synthetic data of a pair of Gaussian particles, shown in Fig. 2.6(a) with 10% additive noise. Simulating drift in a realistic STEM experiment, we created 8 copies of the two particle images, randomly shifted. For each noise level we sampled 1000 noise instances, with each reconstructing the underlying with both FS and our Chebyshev-polynomial based Super Registration.

Figure. 2.6(b) shows the error of inferring the position of the larger particle using both the FS reconstruction (pink line) and SR reconstructions (green line). For $\sigma = 0.3$ or 30% noise we see that the precisions of particle position are 10x better using SR than FS. Further, the SR method, not even using the correct model (a sum of Gaussian particles), is only about twice the CRB for particle position inference (black dotted). Finally, we show the result when using shifts inferred by the same data coarsened by $a = 3$, which was chosen to have the lowest error without being biased. In summary we see that even though small shift errors do not have a dramatic effect on the reconstructed image as measured by eye, there are drastic effects on the precision of extractable information from the reconstructions.

2.5.1 Computational complexity

The standard Fourier shift method requires a Fourier transform of one of the images for each iteration of the optimization, ultimately scaling in time as $\mathcal{O}(N \log N)$, where N is the number of pixels in one image. Super registration requires estimating the underlying image, and thus requires $O(NM)$ where M is the number of polynomials used in the image model. SR requires trying multiple values of M, and multiple models, to find the greatest evidence. For two 128×128 images, FS takes less than a second on a modern computer. SR requires an hour or more to try multiple values of M, but only a few minutes to find the shifts and image for a given M. For multi-image registration, optimal FS requires comparing all pairs of images, and so scales as $\mathcal{O}(L^2 N \log N)$ for L images, whereas SR scales as $\mathcal{O}(LNM)$, as it compares the data only to the model. Memory requirements depend on the algorithm used. In this work we used Levenberg-Marquardt nonlinear least-squares optimization, which requires $\mathcal{O}(LNM)$ memory to store the Jacobian, and so images larger than 128×128 are impractical.

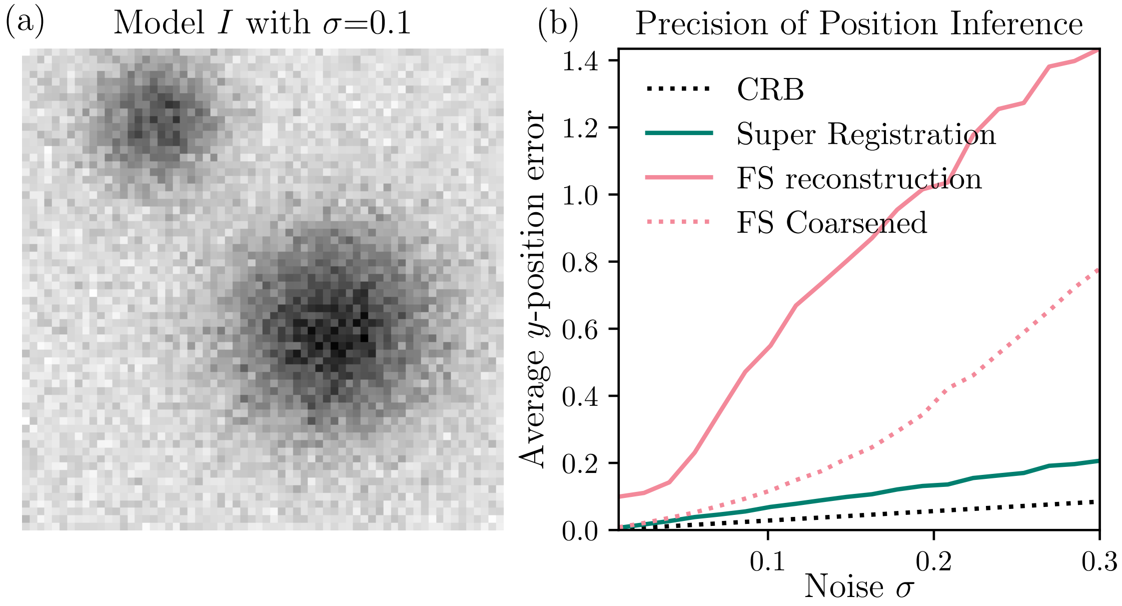


Figure 2.6: (a) A model image of two Gaussian particles with 10% Gaussian additive noise. Eight of these images with random sub-pixel relative shifts were generated, and 1000 noise samples were drawn. For each noise sample, the underlying image was reconstructed either by the standard Fourier shift (FS) reconstruction or with Super Registration. With each reconstruction we fit the Gaussian models which generated the data, inferring the most likely particle position and width. (b) The average error of inferring the y-position of the larger particle from images reconstructed with the standard FS method (pink line), a coarse-grained image (pink dotted), and Super Registration (green line).

There are several open opportunities for improving the performance of Super Registration. Memory consumption and computational time can be improved by using Variational Inference and Stochastic Gradient Descent, which scales with $\mathcal{O}(LN)$ in memory, and will be the subject of future work. A local image model (where each image parameter only modifies a small area of the image), such as radial basis functions, would scale even better than the Fast Fourier Transform, as $\mathcal{O}(N)$. Finally, GPUs are designed to perform optimal image calculations, and SR could achieve at least 10× (by naïve FLOP counts) the performance as compared to a CPU.

2.6 Conclusion

Through a statistical theory of image formation, we have derived the standard method of image registration, which shifts one image to match another. Our theory predicts that shift errors for the standard FS method grow quadratically with noise, much faster than the linear relationship of the CRB. Our explanation for the deviation between the naïve CRB and the standard method comes from a deep relationship between integration and optimization. The resulting formula is useful for designing experiments which require image registration and must be performed using the standard method. Our analysis leads to the surprising fact that coarse-graining the data can improve the shift errors.

We develop a new method of image registration, which models the underlying image, shifts that to match the data, and follows Bayesian inference to select the image model for which there is the most evidence. Our theory reveals an inextricable relationship between image registration and super-resolution—that ultimate shift precision is predicated on selecting a probable model. Therefore we named our new method Super Registration. We showed for periodic images that a Fourier series image model achieves errors consistent with the CRB. We demonstrated superior bias and expected error performance for general non-periodic images, and discussed the shortcomings of our general model. Finally, we showed that, despite marginal improvements in image quality as measured by eye, particle tracking experiments can be 10× more precise when using Super Registration reconstructions.

Our results can be extended to more general transformations: by application of the chain rule each term in our calculation of the average bias and variance will be modified by partial derivatives. It is reasonable to assume that the same problems—nonzero bias and errors which are much larger than the CRB—will persist for transformations like affine skews, rotations, and non-rigid registrations. Super Registration can accommodate all of these problems by constructing the forward transformation instead of reconstructing the inverse transformation.

Finally, medical imaging consists of lining up images of the same tissue from different modes like X-ray and Magnetic Resonance Imaging (MRI) [Zöllei et al., 2003, Leventon and Grimson, 1998]. The Super Registration method involves constructing a generative model for the data, and this perspective reminds us that contrast and features in X-ray and MRI will be different because they respond to different underlying tissue structures. Bias and large errors for this problem have been observed and attributed to this fact [Tyler, 2018]. Therefore some underlying model of tissue component densities and a model of image formation (Super Registration) will be critical for accurately and precisely registering these images.

Image registration is a very important and fundamental problem in medical imaging, remote sensing, self-driving automobiles, non-destructive stress measurement, microscopy, and more. Our theoretical study of the fundamental problem of rigid shift registration in the presence of noise answers long-standing questions on the precision and accuracy of shift inference, elucidates an inextricable link between registration and super-resolution, and inspires a solution to these problems with wide applicability.

Acknowledgements

Thanks to Ismail El Baggari, S.B. Kachuck, K.P. O'Keeffe and D.B. Liarte for useful conversations in the preparation of this manuscript. This work was supported by the NSF Center for Bright Beams, award #1549132.

Chapter 3

Reconstruction of current densities from magnetic images by Bayesian Inference

3.1 Introduction

Two-dimensional (2D) materials host a variety of electronic transport phenomena, many of which are associated with a non-trivial spatial structure of the current density in the material. A non-invasive, local way to image a two-dimensional current density is to image the magnetic field produced by the current and infer the current density. To date, numerous magnetic imaging techniques have been used to image current densities including scanning SQUID [Kalisky et al., 2013, Nowack et al., 2013, Vasyukov et al.,

The work constituting this chapter was done in collaboration with Katja C. Nowack and James P. Sethna

2013], scanning Hall probe [Dinner et al., 2007], magneto-optics [Pashitski et al., 1997], and nitrogen-vacancy (NV) centers in diamond [Chang et al., 2017, Tetienne et al., 2017, Ku et al., 2019].

Most magnetic imaging techniques probe a single magnetic field component in a plane at a constant height above the sample. The relation between the current density and the measured magnetic image is defined through two convolutions: (1) the Biot-Savart law relates the magnetic field to the current density and (2) a convolution of the magnetic field with the point spread function of the magnetic sensor. The current reconstruction problem is therefore a linear deconvolution problem. To obtain the local current density from a magnetic image, the two convolutions have to be inverted. If the current density only varies in two dimensions, this inversion is in principle possible because current conservation relates the two in-plane components. In practice, the inversion is a non-trivial task because the problem is ill-posed: (1) experimental images contain noise and (2) the finite scan height and point spread function lead to a loss of information. Noise with high spatial frequencies dominates the reconstructed image. There are many solutions that predict the data *including* noise perfectly. Most of these solutions are not physical, And so, a criterion for what constitutes a physically sensible solution—often called regularization—is required.

A detailed overview of existing methods for current reconstruction are given by Meltzer et al. [Meltzer et al., 2017]. The most intuitive method is to invert the convolutions directly in Fourier space [Roth et al., 1989], filtering high spatial frequencies that otherwise cause instability. However, the shape and cutoff frequencies of the applied filters limit the resolution of the reconstructed image in a sub-optimal and uncontrolled way. Iterative conjugate gradient methods have also been employed [Wijngaarden et al., 1998], but—though they are more stable to noise—the regularization is not well controlled.

Feldmann [Feldmann, 2004] and Meltzer et al. [Meltzer et al., 2017] have reported reconstruction procedures using Tikhonov regularization penalizing the Laplacian of the current dipole field, combined with a cross-validation-based choice of the regularization strength. Tikhonov regularization is an attractive method because it is analytically tractable in Fourier space, allowing for computationally efficient solutions and theoretically motivated methods of choosing the regularization strength.

In the wider image reconstruction literature, a variety of regularization penalties have been developed that are not analytically tractable. For example, total variation penalizes oscillations in a solution, but does not penalize sharpness like Gaussian or Tikhonov regularization. An additional complication when reconstructing current densities is that typically some current leaves and enters the imaged field of view. At the points along the image boundary where this happens, the current density is not conserved. This violates the assumption of conserved current, without which the problem is under constrained. Meltzer et al. [Meltzer et al., 2017] have implemented mirror boundary conditions (similar to image charges) for accommodating currents which enter or leave the image. While the mirror boundary conditions suppress ringing at the edges—which otherwise results in an artifact—it is not faithful to the sample geometry unless the sample is mirror-symmetric.

Here we describe a new procedure to reconstruct current density from magnetic images that is amenable to a wider class of priors than previous work, and can accommodate currents crossing the image boundaries. Formulating the reconstruction problem in a probabilistic framework suitable for Bayesian inference offers unprecedented flexibility to make use of prior information about the current density, including the sample geometry. This knowledge is imposed through different so-called priors, which assure physically sensible current densities and which, are equivalent to choosing a regularization. Previous methods have penalized the Laplacian of the current dipole field, from which the current density is computed, or the components of the currents themselves. We discuss a set of requirements which any sensible prior should obey. From this we show that a prior based on the Frobenius of the Hessian is better motivated than the commonly used Laplacian prior. This new prior improves existing Tikhonov regularization, and supports our new methods. This generative approach could be extended to any imaging problems in Physics, where a detailed theory of image formation is known.

The priors we explicitly discuss include (1) a Gaussian prior (Tikhonov regularization) penalizing the Laplacian and the Frobenius of the Hessian, (2) a total variation prior, which penalizes strong fluctuations in the current density, but does not necessarily smooth sharp edges, and (3) a finite support prior (which allows the user to specify areas in the field of view where no current flows). In addition, taking advantage of the generative approach, we show that we can accommodate currents crossing the image boundary through modeling current densities flowing outside the field of view based on the sample geometry. This reconstruction problem ultimately leads to a convex optimization problem which we solve using the Alternating Difference Method of Multipliers (ADMM)[Boyd et al., 2011].

The paper is organized as follows. In Sec. 3.2, we define the forward problem, describe how we use Bayesian inference, and explain a generative approach for current reconstruction. In Sec. 3.3 we propose a set of requirements any prior should obey, derive a new prior which satisfies them, and compare it to a previously studied prior. We explore the Gaussian priors and develop new total variation and finite support priors. In Sec. 3.3.3, we discuss how to optimize the strength of each prior. In Sec. 3.4, we describe how we account for currents flowing outside the field of view, and benchmark our method using numerical results throughout. Details of the inversion algorithm, the implementation of finite support priors, and the external current models are presented in

Sec. 3.6.2 and Sec. 3.6.3. The code is in an open-source `python` module called `pysquid`, publicly available in a github repository[1].

3.2 Bayesian inference formulation of the current reconstruction problem

3.2.1 Forward problem

First, we describe the forward problem: calculating the magnetic image resulting from a given two-dimensional current density, assuming the magnetic sensor probes only the perpendicular component of the magnetic field at a fixed height above the current. We also assume the sensitive area is small compared to the scan height so that the point spread function can be ignored, but all the methods we present here can be generalized to include a PSF. Experimental determination of the PSF for a SQUID imaging system will be presented in future work. It is also straightforward to extend our reconstruction procedure to different magnetic field components, and allow for a finite thickness of the current carrying sheet, uniform in the perpendicular direction.

The perpendicular or $z-$component of the magnetic field produced by the two-dimensional current at a position $\boldsymbol{r} = (x, y, z)$ above the sample is given by the Biot-Savart law: $B_z(\boldsymbol{r}) = K(\boldsymbol{r}) * j(\boldsymbol{r})$ where $K(\boldsymbol{r}) = \frac{\mu_0}{4\pi}\frac{3z^2-r^2}{r^5}$ with $r = |\boldsymbol{r}|$, μ_0 the vacuum permeability. The symbol $*$ denotes a convolution $f(\boldsymbol{r}) * h(\boldsymbol{r}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(\boldsymbol{r}-\boldsymbol{s})j(\boldsymbol{s})dx'dy'$ where $\boldsymbol{s} = (x', y', 0)$ because the current density is constrained to the $x - y$ plane. Assuming no current sources and drains are present in the field of view, the $x-$ and $y-$ components of the current density $j_{x,y}$ obey current conservation: $\partial_x j_x + \partial_y j_y = 0$. We explicitly

[1] `https://github.com/colinclement/pysquid`

enforce current conservation through introducing a single scalar field $g(\mathbf{s})$ which only depends on two dimensions. From the scalar field, we calculate the current density as $\boldsymbol{j}(\boldsymbol{s}) = \nabla \times g(\boldsymbol{s})\hat{\boldsymbol{z}}$. After employing a number of vector identities, we can write the Biot-Savart law as a function of g [Wijngaarden et al., 1996]:

$$B_z(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \int\int g(\boldsymbol{s}) \frac{3(\hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{n}})^2 - 1}{|\boldsymbol{r} - \boldsymbol{s}|^3} dx' dy', \tag{3.1}$$

where $\hat{\boldsymbol{n}} = (\boldsymbol{r} - \boldsymbol{s})/|\boldsymbol{r} - \boldsymbol{s}|$. The kernel convoluted with g in Eqn. 3.1 is recognizable as the magnetic field of a point dipole oriented along the $z-$ direction. g can be therefore viewed as a decomposition of a 2D current density into circulations of current, which is why we call g the *current dipole field.* Our model then takes the form of a magnetic image vector $\boldsymbol{\phi}$ with pixel values $\phi(\boldsymbol{r_i}) = B_z(\boldsymbol{r_i})$ sampled in a discrete rectangular grid $\{\boldsymbol{r}_i\}$, given some current dipole field g, which we will also assume is discretely sampled on the same coordinates, but separated from the imaging plane by some distance.

3.2.2 Inverse Problem and Bayesian Inference

The Biot-Savart law is linear, so that the relationship between a discretely sampled magnetic image $\boldsymbol{\phi} \in \mathcal{R}^N$ and a discrete current dipole field $\boldsymbol{g} \in \mathcal{R}^M$ can be written $\boldsymbol{\phi} = M\boldsymbol{g}$ for some suitable linear operator $M \in \mathcal{R}^{N \times M}$. Assuming that the current is thin and circulates in a rectangle around the pixels, the matrix elements are calculated as a function of height above the sample in the appendix (see eqn. 3.23). While the matrix representation of M is impractical to store for any reasonable image sizes, the matrix-vector products $M\boldsymbol{g}$ can be efficiently computed using a Fast Fourier Transform (FFT). In general, magnetic sensors are sensitive to magnetic fields integrated or averaged over an area and characterized by a point spread function (PSF). The PSF can be incorporated into M and will be the subject of future work for SQUID imaging in particular.

We assume that the experimental noise is independent and identically distributed for each pixel, so that our model for a measured magnetic image is

$$\boldsymbol{\phi} = M\boldsymbol{g} + \boldsymbol{\eta} \tag{3.2}$$

where for each pixel $p(\eta_i) \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise with variance σ^2. Written this way, we interpret the noise as causing the data to fluctuate around the model with characteristic distance σ. We can therefore define the likelihood $p(\boldsymbol{\phi}|\boldsymbol{g})$ of measuring $\boldsymbol{\phi}$ given $\boldsymbol{g}$:

$$p(\boldsymbol{\phi}|\boldsymbol{g}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}||M\boldsymbol{g} - \boldsymbol{\phi}||^2\right), \tag{3.3}$$

where N is the number of pixels in the image $\boldsymbol{\phi}$, and $||\cdot||^2$ is the Euclidean L_2 norm. This likelihood $p(\boldsymbol{\phi}|\boldsymbol{g})$ is our model of the data. The model allows us to infer the current dipole field from the experimental data with better inference possible for lower experimental noise.

Our goal is to learn $\boldsymbol{g}$ after having measured $\boldsymbol{\phi}$. We therefore need $p(\boldsymbol{g}|\boldsymbol{\phi})$ (called the posterior (post-measurement) probability). Bayes' Theorem tells us how to reverse the conditional probability:

$$p(\boldsymbol{g}|\boldsymbol{\phi}) = \frac{p(\boldsymbol{\phi}|\boldsymbol{g})p(\boldsymbol{g})}{p(\boldsymbol{\phi})}, \tag{3.4}$$

where $p(\boldsymbol{g})$ is the prior probability, encoding a criterion for preferable and physically sensible solutions. $p(\boldsymbol{\phi})$ is called the evidence—which is the normalization of the posterior—and is useful for quantitatively justifying the selection of one model over another [MacKay, 1992]. The maximum likelihood solution to the reconstruction problem is then the most likely $\boldsymbol{g}$ given $\boldsymbol{\phi}$:

$$\boldsymbol{g}^\star = \max_{\boldsymbol{g}}\ p(\boldsymbol{g}|\boldsymbol{\phi}) = \max_{\boldsymbol{g}}\ p(\boldsymbol{\phi}|\boldsymbol{g})p(\boldsymbol{g}). \tag{3.5}$$

At this stage of inference the evidence $p(\boldsymbol{\phi})$ can be left out as it does not change the peak of the distribution (it is independent of $\boldsymbol{g}$, being the integral over all $\boldsymbol{g}$). Full treatment of Bayesian inference, including optimal model selection is described in detail by Mackay [MacKay, 1992].

We can now define the maximum-likelihood solution for current inference by combining eqns. 3.5 and 3.3 to find

$$\boldsymbol{g}_\lambda = \min_{\boldsymbol{g}} \frac{1}{2}||M\boldsymbol{g} - \boldsymbol{\phi}||^2 + (\lambda\sigma)^2 \ \ell(\boldsymbol{g}), \tag{3.6}$$

where we assume that the prior can be written $p(\boldsymbol{g}) \propto \exp\left(-\lambda^2 \ell(\boldsymbol{g})\right)$ for some non-negative penalty function ℓ and real number λ. Inference of currents is now cast as minimizing the negative log-posterior, or minimizing the distance between our model $M\boldsymbol{g}$ and the data $\boldsymbol{\phi}$, constrained by a regularization $l(\boldsymbol{g})$.

It is instructive to demonstrate the necessity of some nontrivial prior $l(\boldsymbol{g})$. If $p(\boldsymbol{g}) \sim 1$ (all solutions are considered equally preferable), then the maximum likelihood solution of eqn. 3.6 is

$$\boldsymbol{g} = (M^T M)^{-1} M^T \boldsymbol{\phi}, \tag{3.7}$$

which is the pseudoinverse (the 'closest' inverse to the singular M), which is calculated only from the eigenvectors of M with non-zero eigenvalue. M has at least one zero eigenvalue, as adding any constant to $\boldsymbol{g}$ does not change $\boldsymbol{\phi}$. The pseudoinverse can gracefully ignore this symmetry, but since the Biot-Savart law is a long-range (power law) blurring kernel, M has exponentially small eigenvalues corresponding to high frequencies. Additive Gaussian noise has support in all frequencies; deconvolution is highly unstable as the pseudoinverse amplifies any amount of noise.

The pseudoinverse can yield solutions which fits the data perfectly, however, it will fit the noise (over-fitting) as well as the data. There is a huge space of solutions $\boldsymbol{g}$ that can overfit the data like this, and most of them oscillate rapidly throughout the image.

The role of a non-trivial prior $p(\boldsymbol{g})$ is to restrict this space by using physical principles to specify which current solutions are more likely.

3.3 Understanding Priors

A natural starting place for understanding priors is by assigning a Gaussian penalty of a linearly transformed part of $\boldsymbol{g}$:

$$p(\boldsymbol{g}) \propto e^{-\lambda^2 \ell(\boldsymbol{g})} = \exp\left(-\lambda^2 ||\Gamma \boldsymbol{g}||^2\right), \tag{3.8}$$

where Γ is a linear operator. The maximum likelihood solution to eqn. 3.6 is equivalent to Tikhonov regularization discussed in the context of current reconstruction in refs. [Feldmann, 2004, Meltzer et al., 2017] and the optimal Wiener filter for some choice of Γ [Press et al., 1989]. Gaussian priors are the conjugate prior to a Gaussian likelihood, with the explicit solution given by:

$$\boldsymbol{g}_\lambda = (M^T M + (\sigma\lambda)^2 \Gamma^T \Gamma)^{-1} M^T \boldsymbol{\phi}. \tag{3.9}$$

In this form, we can see that the role of Γ is to override the exponentially small eigenvalues of $M^T M$, regularizing the instability of the pseudoinverse in eqn. 3.7. Written in this form, we see that σ sets the scale for the regularization strength λ. We will discuss in detail how to choose λ in section 3.3.3.

What should determine the operator Γ? The simplest choice is the identity $\Gamma = \mathbb{I}$. In this case, the prior favors a small-magnitude solution, which is not often physically motivated. If Γ corresponds to derivatives, the prior prefers smooth solutions. The Laplacian $\Gamma = D_x^2 + D_y^2$—with second derivative operators D_x^2 in the x-direction and D_y^2 in the y-direction—is a common choice for image reconstruction problems and has been discussed in the context of current reconstruction [Feldmann, 2004, Meltzer et al., 2017].

The Laplacian is translation invariant, prefers small accumulated curvature, and allows eqn. 3.9 to be solved directly in Fourier space [Feldmann, 2004, Meltzer et al., 2017].

We can physically interpret the Gaussian Laplacian (GL) prior by (1) recognizing that $\boldsymbol{j} = \nabla \times g\hat{\boldsymbol{z}} = \partial_y g\hat{\boldsymbol{x}} - \partial_x g\hat{\boldsymbol{y}}$ and (2) writing the log of eqn. 3.8 in the continuum limit for clarity of notation as:

$$\begin{aligned}\ell_{\mathrm{GL}}(g) &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r}\ (\partial_x^2 g + \partial_y^2 g)^2 \\ &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r}\ |\nabla \times \boldsymbol{j}|^2,\end{aligned} \tag{3.10}$$

where the second line assumes that current only varies in the $x - y$ plane. Therefore the Laplacian prior prefers solutions with small accumulated circulation of current. The Laplacian prior is certainly more physical than $\Gamma = \mathbb{I}$, but it is not clear why we should penalize only the circulation of current.

This leads us to seek other choices for the Gaussian prior in the context of current reconstruction. In general, we would prefer that a prior obeys physically-motivated symmetries and that the prior be a functional, that is, an integral of some scalar quantity defined locally in the plane of currents. Such a scalar quantity should have the following properties:

1. invariance under current inversions $g \to -g$,
2. invariance under rotations and reflections,
3. penalize all variations in currents, i.e. first derivatives of $\boldsymbol{j}$ and thus second derivatives of g.

The Laplacian prior $\Gamma = \nabla^2$ satisfies almost all of these; it is the integral of $(\nabla^2 g)^2$; (1) the quadratic makes it invariant to $g \to -g$, (2) the Laplacian is rotation and reflection

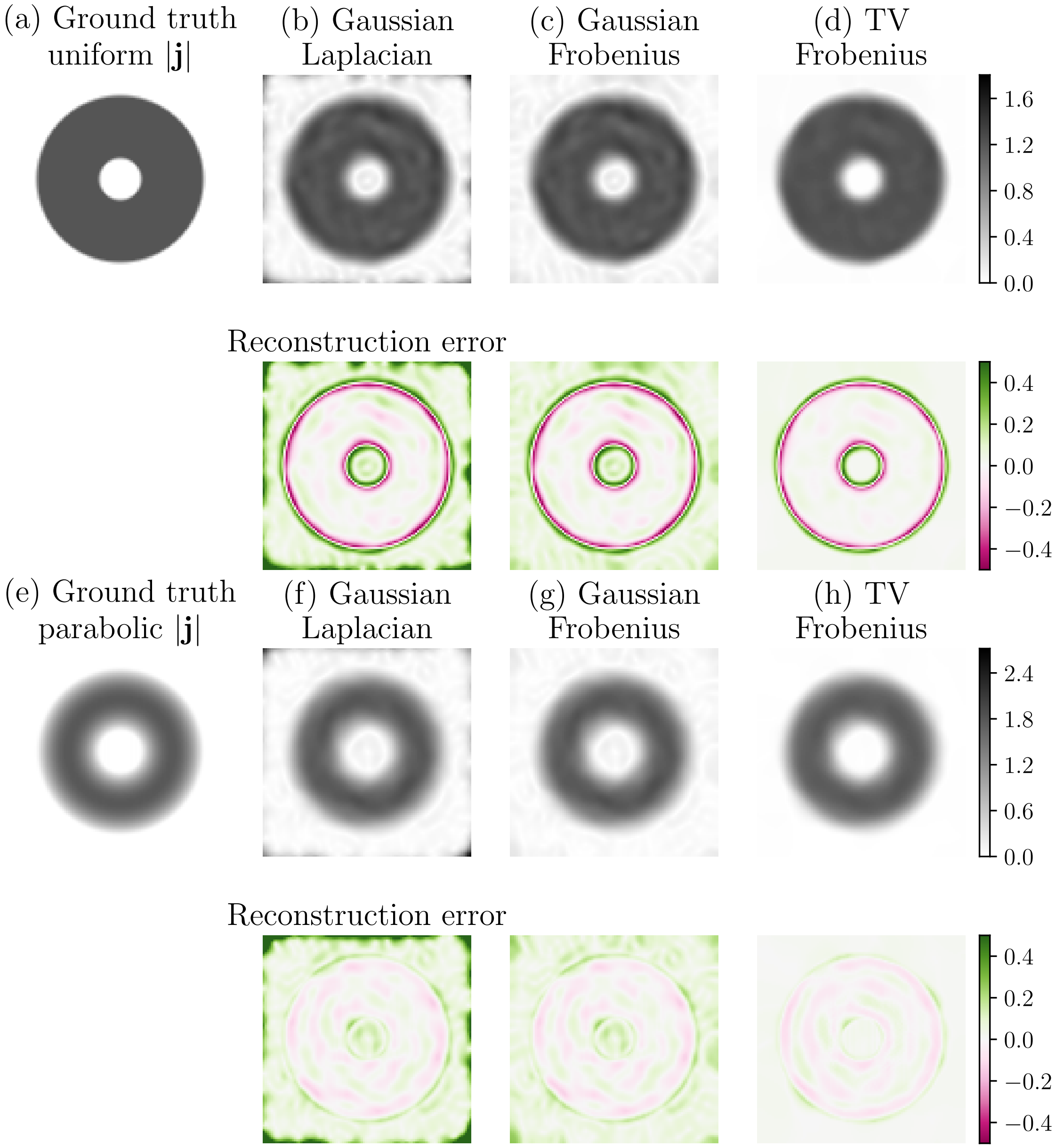


Figure 3.1: Ground truth 100×100 current density $|\boldsymbol{j}|$, with uniform profile (a) and parabolic profile (e). Synthetic ϕ was computed with a height above the plane of 4 pixel widths and 5% noise was added. Reconstructions for the uniform annulus in (b-d) and the parabolic annulus in (f-h) with Gaussian prior penalizing the Laplacian, the Gaussian prior penalizing the Frobenius Hessian, and the total variation of the Frobenius Hessian. The regularization strength for each was chosen by our Bayesian discrepancy principle. The data $\boldsymbol{\phi}$ was re-scaled to have unit peak-to-peak range, and $\lambda = 1.4$ was used for the TVF prior reconstruction, and $\lambda = 2.$ was used for the GL and GF priors.

invariant (as we will soon show), (3) following eqn. 3.10 we see that not all the possible variations of the current are penalized.

The combination of the inversion symmetry and only allowing second derivative of g implies that all physically motivated functionals must contain two powers of elements of the second derivative matrix. Our prior must depend on the Hessian $H_{\alpha\beta} = \partial_\alpha \partial_\beta g$, where $\alpha = x, y$ and $\beta = x, y$. To construct a functional that is in addition rotation invariant, we must contract the indices (following Einstein summation conventions), and there are only two ways to do this [Cvitanović, 2008] with two powers of g: $H_{\alpha\alpha} H_{\beta\beta} = (\mathrm{Tr} H)^2$ and $H_{\alpha\beta} H_{\alpha\beta} = \mathrm{Tr} H^T H$. The former is the Laplacian we have studied above; the latter is the square of Frobenius norm of the Hessian. This new Gaussian Frobenius (GF) prior satisfies our first two criterion by construction. Following a similar calculation to eqn. 3.10,

$$\begin{aligned} \ell_{\mathrm{GF}}(g) &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r} \quad H_{\alpha\beta} H_{\alpha\beta} \\ &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r} \quad (\partial_x j_y)^2 + (\partial_y j_x)^2 + \\ & \qquad\qquad\qquad (\partial_x j_x)^2 + (\partial_y j_y)^2, \end{aligned} \tag{3.11}$$

we find that $\ell_{\mathrm{GF}}(g)$ penalizes all variations in currents and satisfies our third and final criteria.

The superiority of this new prior is demonstrated in fig. 3.1. First, for comparison, it explores simple synthetic data examples in the form of two annuli: one with currents with uniform current density (a) and one with parabolic current density going to zero at the edges (e). Synthetic data was produced by creating a kernel M from a height about the plane of 4 pixel-widths, and noise of $\sigma = 0.05$ was added.

The second column of fig. 3.1 shows reconstructions using the Gaussian Laplacian prior in eqn. 3.10 on the uniform annulus data (b) and the parabolic annulus data (f).

Both reconstructions show large magnitudes of spurious currents at the edges of the image–likely due to the Laplacian not penalizing all current components from varying too much. The third column (c) and (g) show reconstructions using the new Frobenius prior in eqn. 3.11, with improved edge reconstructions. The regularization strength λ was set using our Bayesian discrepancy principle (described in section 3.3.3); the priors were computed using centered finite difference derivatives (described in section 3.6.2). The fourth column of 3.1 (d) and (h) shows reconstructions with our new total variation of the Frobenius norm prior: with more even reconstructions where current is zero and errors largely concentrated at the edges of the annuli.

We solved the maximum likelihood problem of eqn. 3.5 with the Gaussian Laplace prior in eqn. 3.10 and the Gaussian Frobenius prior in eqn. 3.11 by iteratively solving the appropriate regularization psuedoinverse in eqn. 3.9. The construction of appropriate Γ operators is discussed in sec. 3.6.2 of the Appendix. The computational complexity of one iteration of the solution of eqn. 3.9 is $\mathcal{O}(N \log N)$ for N pixels in the data $\boldsymbol{\phi}$ via an FFT. The iterative method scales the same way, but will take a number of steps which depends on the condition number of the operator (which depends on Γ, σ, and λ).

3.3.1 Total Variation Priors

As we discussed in the introduction, the current reconstruction literature has so far only considered analytically tractable, Gaussian priors. These are attractive because the resulting reconstruction problem can be solved with a couple FFTs and there exists calculations for motivating the choice of regularization strength λ. Unfortunately, Gaussian priors in particular suffer from ringing, especially at sharp boundaries due to the Gibbs phenomenon [Gottlieb and Shu, 1997]. For example, in figure 3.1(b) and (c), the Gaussian prior always allows unnecessary variations of current inside the uniform annulus.

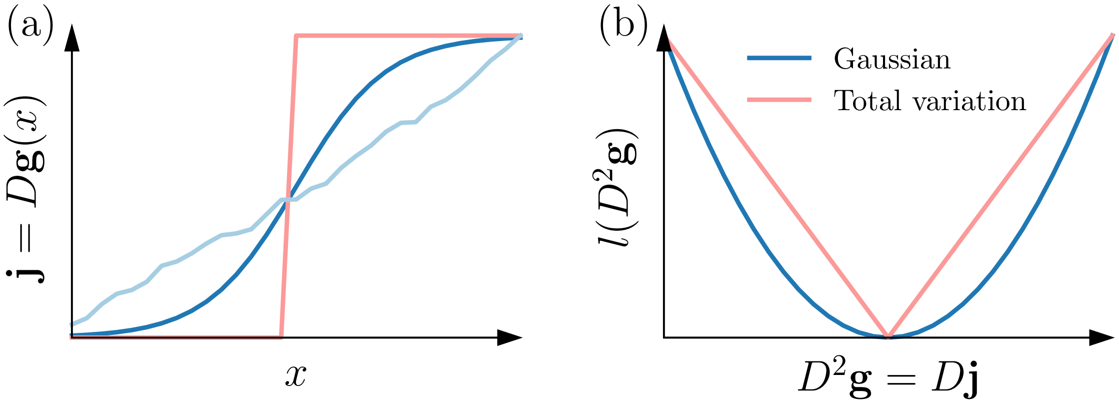


Figure 3.2: (a) Three possible current $\boldsymbol{j}$ profiles, all of which are equally probable under a total variation prior. (b) Illustration comparing Gaussian and total variation priors. Gaussian tolerates very small variations in current, while total variation much more heavily penalizes any non-zero amount. The result is total variation prefers reconstructions with regions of constant current, and suppresses ringing more effectively than Gaussian priors.

Therefore a prior which penalizes oscillations without penalizing real sharp changes in the solution is desirable. In image reconstruction, this is achieved via the total variation (TV) prior, summing over the absolute values of the derivative [Vogel and Oman, 1998, Osher et al., 2005]. Since we want to penalize derivatives of $\boldsymbol{j}$, we again penalize second derivatives of g:

$$\begin{aligned}\ell_{\mathrm{TV}}(g) &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r}\ |\partial_x^2 g| + |\partial_y^2 g| \\ &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r}\ |\partial_x j_y| + |\partial_y j_x|. \end{aligned} \tag{3.12}$$

In order to obtain some intuition about the original total variation ℓ_{TV} (in eqn 3.12), we must ignore (for a moment) its manifest failure to be rotation invariant. Considering a one dimensional image, perhaps a slice of our two dimensional image, fig. 3.2(a) shows three hypothetical variations in currents. The TV prior considers all three of these—regardless of their smoothness—equally probable. Therefore, the TV prior will suppress oscillations in the solution and, unlike Gaussian priors, remain agnostic to the sharpness of the transition. Figure 3.2(b) shows the value of both the Gaussian and TV priors as functions of the second derivatives of $\boldsymbol{g}$. However, the Gaussian prior is much more permissive of small variations of $\boldsymbol{j}$. On the other hand, the absolute value of the TV causes any finite amount of variations $\boldsymbol{j}$ to be penalized, thus preferring solutions of $\boldsymbol{g}$ with regions of constant $\boldsymbol{j}$ and avoiding penalization of sharp edges.

Section 3.3 shows that any good prior for current reconstruction must satisfy principles, thus any prior must be a function of the Frobenius norm of the Hessian of g: $\mathrm{Tr} H^T H$. We consider Gaussian priors of quantity integrated over the image. Inspired by the properties of the TV prior, we introduce our TV Frobenius (TVF) prior:

$$\begin{aligned}\ell_{\mathrm{TVF}}(g) &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r} \sqrt{H_{\alpha\beta} H_{\alpha\beta}} \\ &= -\lambda^2 \int \mathrm{d}^2\boldsymbol{r} \sqrt{(\partial_x^2 g)^2 + (\partial_y g)^2 + 2(\partial_x \partial_y g)^2}, \end{aligned} \tag{3.13}$$

where the square root of a sum of squares gives us a rotation invariant absolute value (like in the original TV prior).

Figure 3.1(d) and (h) show the result of the TV Frobenius prior on the reconstructions of the uniform and parabolic current annuli respectively. Notice how, in both reconstructions, the background which should be empty of currents is uniformly empty. Since TV penalizes any variation—no matter how small—the optimization problem prioritizes flat (zero) current when zero current is as effective at explaining the data. In fig. 3.1(d) ringing is more penalized, allowing for a more uniform interior of the annulus. Whereas the interior of the uniform annulus is more faithfully reconstructed with a TVF prior (d) than with the Gaussian prior (c), in the case of a parabolic annulus, a TVF prior (g) and Gaussian prior (h) are comparable.

3.3.2 Finite support

The total variation prior prefers current solutions which have contiguous regions of constant current with minimal 'total variation.' Another natural prior—which can accommodate sample geometry—is the finite support prior which uses knowledge of the sample to impose zero current in regions of the image plane. Since many devices are lithographically defined, it is often known where the sample should have current and where it should not. Recall that we are solving for the current dipole field $\boldsymbol{g}$, such that $\boldsymbol{j} = \boldsymbol{\nabla} \times \boldsymbol{g}$. To impose zero current we must also impose that regions with no internal current correspond to regions of constant current dipole field.

In order to impose that certain regions of $\boldsymbol{g}$ be constant, we must assume that we are given an image mask $\boldsymbol{m}$ of the same shape as $\boldsymbol{g}$ (which is 0 where current is unrestricted and 1 where currents are not allowed). Note that since current is imagined to flow 'around' our pixels, current can flow only around the periphery of contiguous regions pixels in the mask with value 1: there is only one current loop for each contiguous region of 1's. From $\boldsymbol{m}$, we then identify each of the regions of contiguous of 0's with their own free parameter. Therefore we optimize $\tilde{\boldsymbol{g}}$–which is the vector of all free current dipole field parameters, equal in number to the number of 1's in $\boldsymbol{m}$ plus the number of contiguous regions of 0's in $\boldsymbol{m}$.

There is a linear operator F such that $\boldsymbol{g} = F\tilde{\boldsymbol{g}}$, where $F \in \mathbb{R}^{N\times P}$ for N pixels in the image plane and P free current dipole field parameters. Concretely,

$$F_{jk} = \begin{cases} 1 & \text{if free parameter } \tilde{g}_k \text{ sets } g_j \\ 0 & \text{else.} \end{cases} \tag{3.14}$$

In the case $N = P$, all pixels are free parameters, and F is (up to permutations) the identity matrix. With this formulation, $\boldsymbol{g} = F\tilde{\boldsymbol{g}}$. We can impose that regions of the image have zero internal circulating currents by simply replacing $\boldsymbol{g} \to F\tilde{\boldsymbol{g}}$ in the log-posterior of eqn. 3.5. The finite support prior is quite powerful because it directly reduces the number of degrees of freedom and highly constrains the solution space. Since the imposition can be implemented with yet another linear operator, it is straightforward to include it with both Gaussian priors and our TVF prior (see sec. 3.6.3 for details).

Figure 3.3 shows the result of adding the finite support prior to the uniform and parabolic annuli models studied in fig. 3.1. By setting values outside and inside of the annulus to 1 and inside the annulus to 0, we form the mask before following up with one step of binary erosion in order to make it slightly imperfect. Figure 3.3(b) shows the reconstruction of the uniform annulus using finite support added to the Gaussian Frobenius prior: we see smooth edges and some remaining ringing inside the annulus. Figure 3.3(c) shows another reconstruction of the uniform annulus using finite support added to the TV Frobenius prior: we see a very uniform interior current density and slightly sharper edges than in fig. 3.1(d). As for the parabolic annulus with finite support priors, fig. 3.3 shows that Gaussian (e) and TVF (f) yield very similar results. Figure. 3.3 demonstrates again that the best priors really depend on the data at hand. For a uniform annulus of current, the TVF prior is clearly more effective than GF, whereas for smoother parabolic currents, GF and TVF perform similarly.

3.3.3 Choosing the strength of the prior

There are many interesting and effective methods for choosing the strength of the prior, including Bayesian evidence maximization [MacKay, 1992], the l-curve method [Hansen, 1992], cross-validation [Golub et al., 1979], and the discrepancy principle [Galatsanos and Katsaggelos, 1992]. None of the existing methods works well for every chosen of prior. In fact, most methods require analytically tractable (Gaussian) priors. Since the TVF and

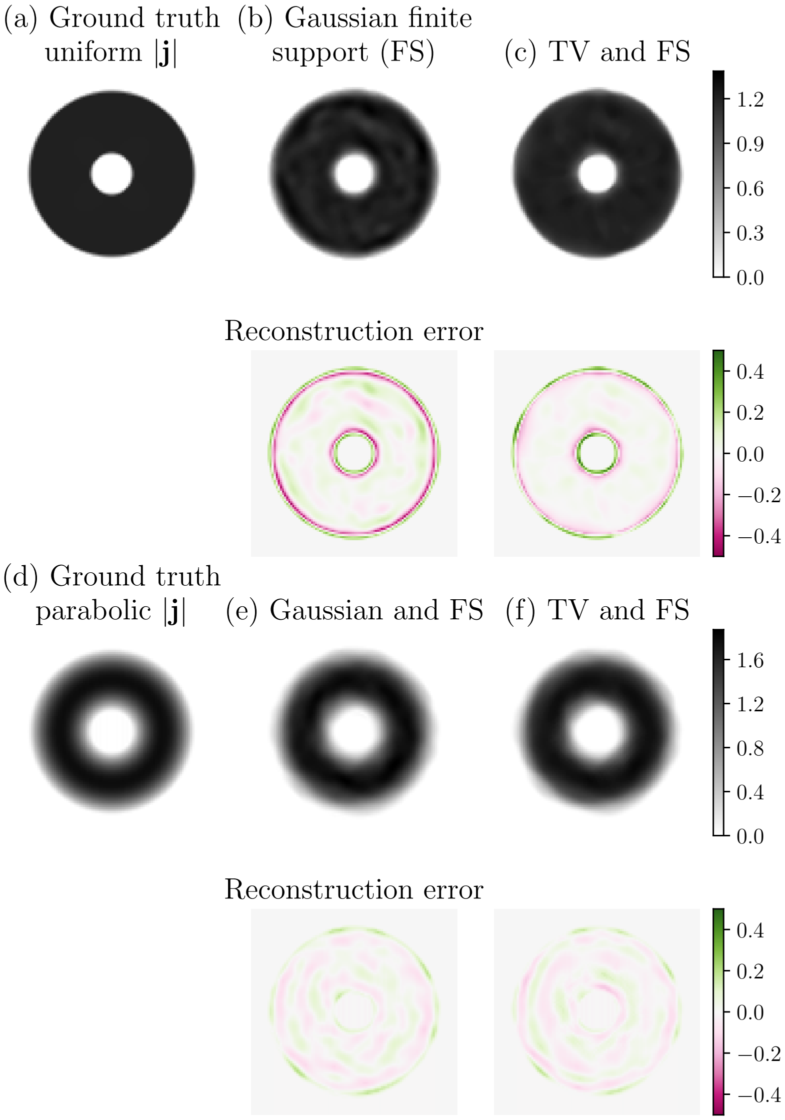


Figure 3.3: Ground truth current density $|\boldsymbol{j}|$, with uniform profile (a) and parabolic profile (d). Synthetic ϕ was computed with a height above the plane of 4 pixel widths and 5% noise was added. Reconstructions for the uniform annulus in (b,c) and the parabolic annulus in (e,f) with Gaussian prior penalizing the Frobenius Hessian, and the total variation of the Frobenius Hessian. The regularization strength for each was chosen by our Bayesian discrepancy principle. The data $\boldsymbol{\phi}$ was re-scaled to have unit peak-to-peak range, and $\lambda = 1.4$ was used for the TVF prior reconstruction, and $\lambda = 2.$ was used for the GF prior.

finite support priors preclude analytic tractability, a more general method for setting the prior strength is needed. Here, we describe a Bayesian modification of the discrepancy principle.

Assuming that the model or likelihood is determined by $\boldsymbol{\phi} = M\boldsymbol{g} + \boldsymbol{\eta}$ such that $p(\eta) \sim \mathcal{N}(0, \sigma^2)$, the inferred $\boldsymbol{g}_\lambda$ for some λ should have the property that the reconstruction error $\text{std}(||M\boldsymbol{g}_\lambda - \boldsymbol{\phi}||^2) = \sigma$. Since the deconvolution problem is ill-posed, and the reconstruction error without regularization can be made arbitrarily small, the discrepancy principle suggests that the strength of the prior should be increased until the reconstruction error has the same spectrum as the noise (set by σ). Figure 3.4 (top) shows the standard deviation of this error using the GF prior as a function of λ for the uniform annulus data of fig. 3.1(a), where a black cross marks the point that satisfies the discrepancy principle.

Since convolution smears together finite regions of current to produce a magnetic image, there are actually fewer degrees of freedom than the number of pixels. Thus many have found that following the discrepancy principle leads to over-smoothed data [Galatsanos and Katsaggelos, 1992]; the image of residuals (in fig. 3.4 marked by a black cross) shows a faint ring in real space. This spatial structure in the residuals (easier to see in Fourier space, above) is a violating of our modeling assumptions: the difference between the model and data should be identically and independently distributed noise. The black hexagon shows this effect exaggerated with a much larger regularization than is necessary. The discrepancy principle can be modified by finding $\gamma < 1$ such that $\text{std}(||M\boldsymbol{g}_\lambda - \boldsymbol{\phi}||^2) = \gamma\sigma$, where γN is the effective number of degrees of freedom. For Gaussian priors the effective number of parameters γN can be estimated, but for general priors, we follow Bayesian inference and assign less posterior probability to $\boldsymbol{g}$-fields resulting in residuals which are not independent and identically distributed.

In practice, we simply find the regularization that satisfies the discrepancy principle,

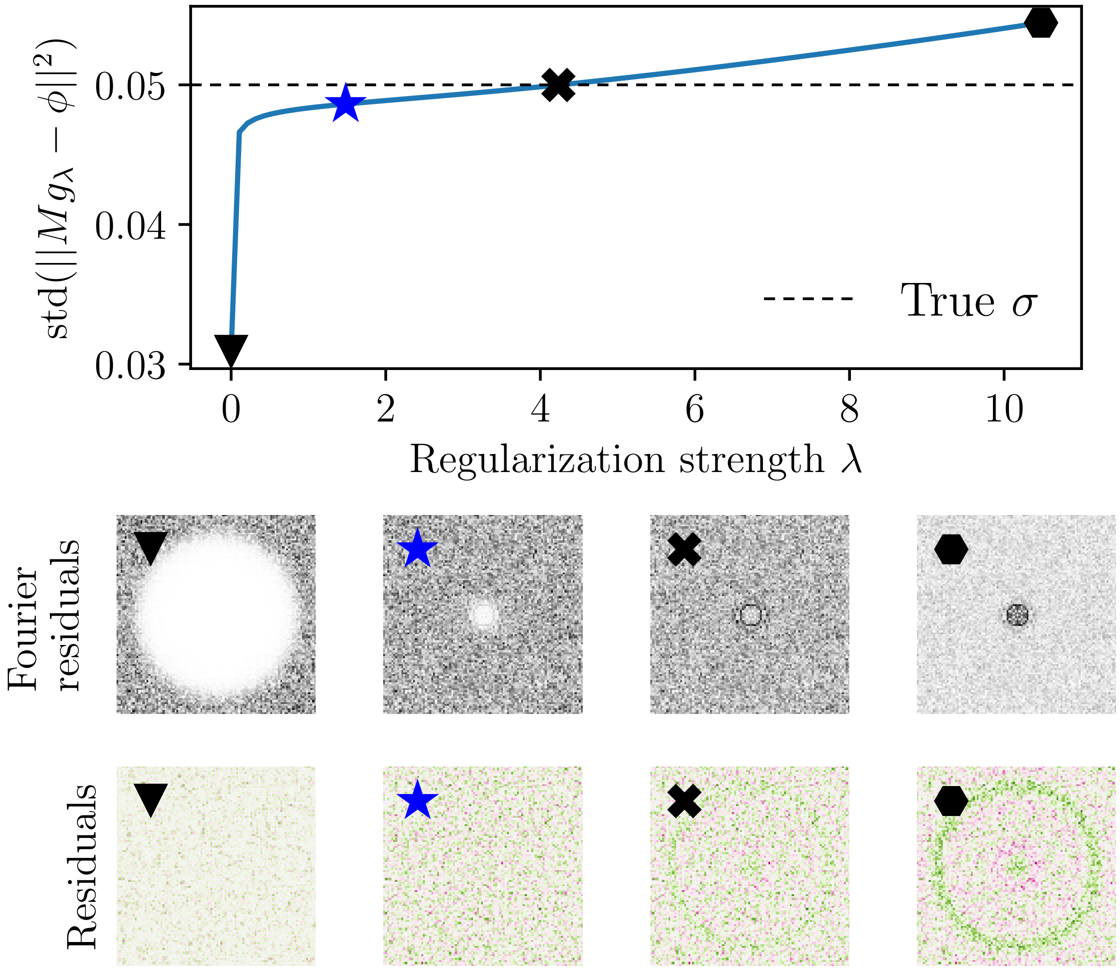


Figure 3.4: Illustrating our Bayesian discrepancy principle, choosing the largest regularization strength of the GF prior for which the residuals are spatially invariant, as shown in the upper-left inset image. The four inset images are the Fourier amplitudes of the residuals superimposed above the real-space residuals, for the regularization strength indicated by each arrow. The model error $\mathrm{std}(||M\boldsymbol{g}_\lambda - \phi||^2$ is plotted as a function of the strength of the prior, using synthetic data from the uniform-profiled annulus with height above the plane of 4 pixels and noise level $\sigma = 0.05$, indicated by the horizontal line.

then reduce it until the residuals have minimal spatial structure in either real or Fourier space (as indicated by the blue star in fig. 3.4). The Fourier space residuals should ideally be uniform noise, but we see some degree of fitting the noise (shown by the light spot in the center). This is likely because we used the GF prior; we have shown that the TVF prior is better suited to data with sharp edges. The near-zero λ case (indicated by the black triangle of fig. 3.4) shows an exaggeration of this over-fitting of noise.

3.4 Currents outside the field of view

Of the key model assumptions we have followed so far, many current distributions of experimental interest violate one in particular: current conservation. We have been working exclusively with the current dipole field $\boldsymbol{g}$ for which $\boldsymbol{j} = \nabla \times \boldsymbol{g}$, so our model exclusively requires currents to circulate in the field of view. One way of accommodating data in which current crosses the field of view is to solve the problem assuming mirror boundary conditions [Meltzer et al., 2017]. While mirror-symmetry is analytically attractive, few (if any) real-world samples can be expected to be mirror-symmetric, and assuming so without reason goes against our generative modeling principal.

We approach a solution by building a model of leads feeding the sample. For most experiments the lithographic design of the imaged device is known, and so just as finite support can be useful, we can assume our leads have uniform resistivity, and build a loop which enters and leaves the field of view, canceling as much as possible the currents incident on the edges. Figure 3.5 shows a van der Pauw geometry, useful for measuring transport properties of samples (see the zoomed inset). The gray region corresponds to the sample—with current flowing from left to right with some current dipole field $\boldsymbol{g}$—and a resulting image $\boldsymbol{\phi}$. We then build an external model (pink), with a corresponding

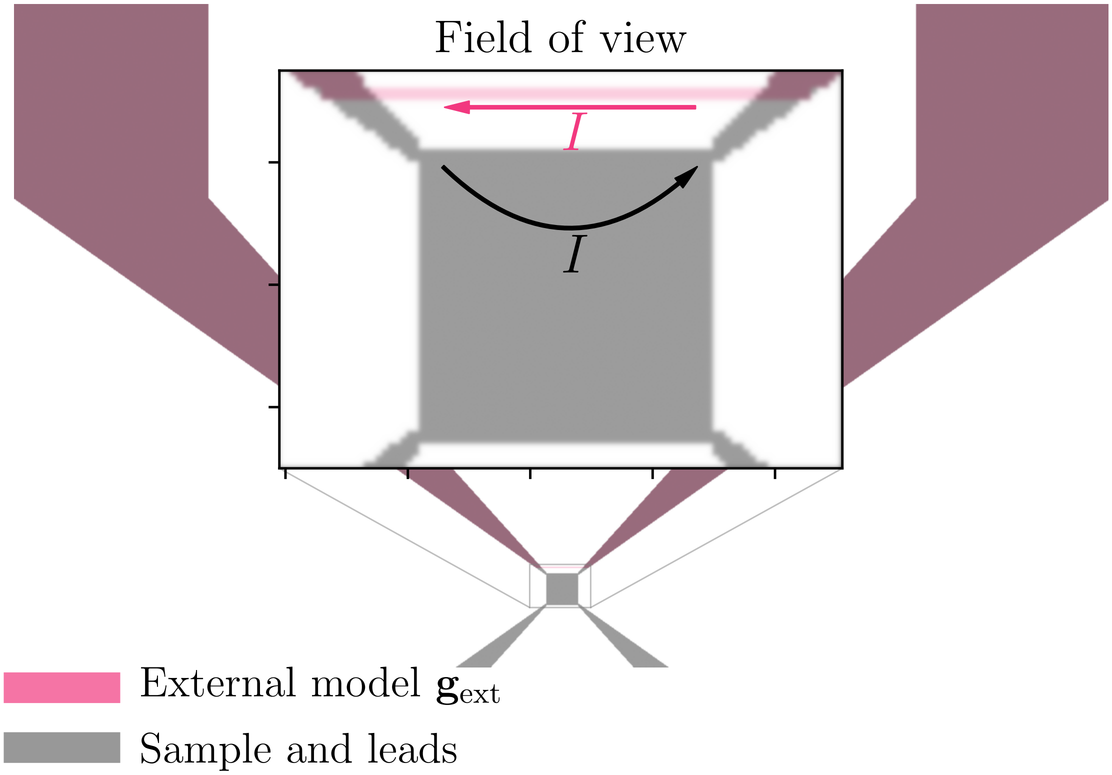


Figure 3.5: The Montgomery or van der Pauw geometry, a common experimental lithographic pattern for measuring transport properties of a sample. In this case, voltage is applied between the top two corners, producing a current through a material with uniform resistivity. The sample is the gray region examined in the zoomed inset. In order to remove currents crossing the image field of view boundary, we model the leads (pink), and subtract the resulting field $\boldsymbol{\phi}_{\text{ext}}$ from the data $\boldsymbol{\phi}$. The linearity of electromagnetism ensures that we are then also subtracting the current in the pink region. The pink external model currents are then accounted for in the reconstruction.

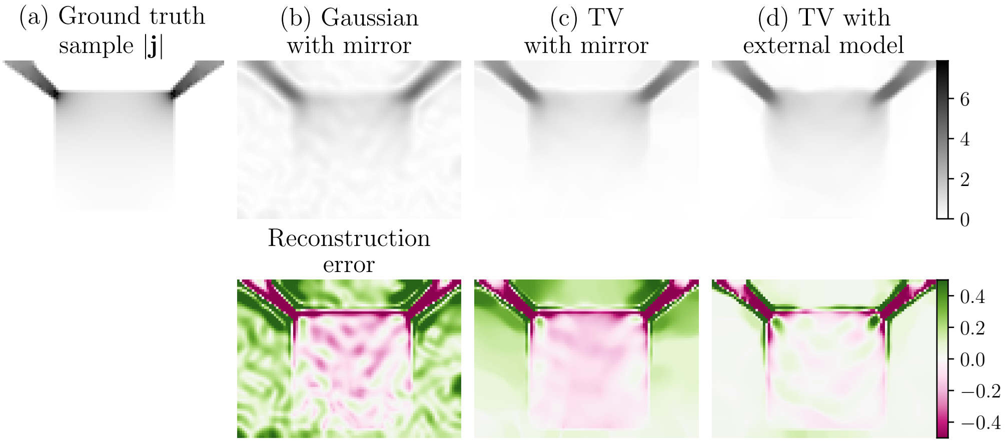


Figure 3.6: (a) Ground truth current density from Montgomery or van der Pauw geometry of a sample of uniform resistivity, with current entering and leaving through the top of the image. Synthetic data was computed with a kernel assuming 4 pixel widths separated the measurement and sample planes, and i.i.d. noise of $\sigma = 0.05$ was added. Reconstructions of the current density using (b) a Gaussian Frobenius prior with mirror-symmetric boundary conditions, (c) a total variation (TV) prior with mirror boundary conditions and (c) a total variation prior using an external model; below each reconstruction is its deviation from the ground truth in (a). The data $\boldsymbol{\phi}$ was re-scaled to have unit peak-to-peak range, and $\lambda = 0.9$ was used for both TVF prior reconstructions and $\lambda = 0.8$ was used for the GF prior.

current dipole field $\boldsymbol{g}_{\text{ext}}$, and subtract the resulting magnetic field of the external model $\boldsymbol{\phi}_{\text{ext}}$ from $\boldsymbol{\phi}$. The linearity of electromagnetism guarantees that we will be only trying to recover the difference $\boldsymbol{g} - \boldsymbol{g}_{\text{ext}}$, which should have conserved current in the field of view.

Our prescription for external modeling is almost complete, except for the fact that subtracting the external model creates extra variations in the currents which will then be unfairly penalized the prior. Therefore the prior needs to be a function of $\boldsymbol{g} + \boldsymbol{g}_{\text{ext}}$. And so, we modify the maximum likelihood inference of eqn. 3.5 as follows:

$$\begin{aligned} \boldsymbol{g}_\lambda = \; & \boldsymbol{g}_{\text{ext}} + \\ & \min_{\boldsymbol{g}} \frac{1}{2} ||M\boldsymbol{g} - (\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{ext}})||^2 + (\lambda\sigma)^2 \; \ell(\boldsymbol{g} + \boldsymbol{g}_{\text{ext}}), \end{aligned} \tag{3.15}$$

where ℓ is any log-prior. Appendix sec. 3.6.3 explains how we accommodate these modifications for the TVF priors.

Given a model of a van der Pauw geometry (in fig. 3.5), we now have a generative scheme for building external current leads into our model via eqn. 3.15. Figure 3.6(a) shows the ground truth current density that results from solving for the currents with voltage applied to the top two contact pads of the van der Pauw geometry in fig. 3.5 (using a tool built into our `pysquid` package). We then produce synthetic data $\boldsymbol{\phi}$ (assuming a height above the plane of 4 pixel widths) and add noise of magnitude $\sigma = 0.05$.

Figure. 3.6(b) shows a reconstruction using the Gaussian Frobenius prior of eqn. 3.11, with mirror boundary conditions to account for currents crossing the image boundary. The reconstruction suffers from excessive variations in current as we saw in fig. 3.1. Figure 3.6(c) shows the reconstruction using the TVF prior of eqn. 3.13 with mirror boundary conditions. The error image directly below displays a haze of pink, indicating excessive current density outside the sample. The mirror boundary conditions were not used to produce the data so the reconstruction inserted some current in order to make it fit. Finally, fig. 3.6(d) shows the reconstruction using the external model of eqn. 3.15 with the TVF prior, and the pink haze in the error (directly below) has been cured. With a generative model, the reconstruction does not need to place current outside the sample to fit the data.

3.5 Discussion and conclusion

Current reconstruction from magnetic images is an important problem: as magnetic imaging methods are approaching fundamental size limits, improving the resolution of current reconstructions will come from improved inference. An ill-posed deconvolution blurred by electromagnetism itself, current reconstruction requires strong regularization for stability. We followed the methods of the literature [Wijngaarden et al., 1998, Feldmann, 2004], and defined the current dipole field $\boldsymbol{g}$ such that $\boldsymbol{j} = \nabla \times \boldsymbol{g}$, and defined the Biot-Savart kernel M such that $\boldsymbol{\phi} = M\boldsymbol{g}$. Using Bayesian inference, we defined the negative log-posterior in eqn. 3.6, the maximization of which provides a solution.

We justified the importance of prior information $p(\boldsymbol{g})$, also called regularization, and derived a new, principled prior—the Frobenius of the Hessian—which improved the standard Gaussian prior. In many experimental situations regions edges of the imaged devices are captured in the field of view. Therefore the field of view necessarily contains sharp edges at which the current drops to zero (corresponding to the device edges) and areas of zero current (corresponding to areas where the device is absent). To improve the reconstruction of regions of constant current, we investigated a total variation prior and contrasted it with Gaussian priors which permits unnecessary oscillation. To be able to use information about the device geometry we developed a finite support prior which can include information about where currents are and are not allowed to flow. Finally, we extended our generative picture by building an external current model to accommodate the assumption of conserved currents in the current dipole field.

In addition to Bayesian inference, our approach used the principle of generative modeling: if we can make convincing data, we can use those models to make better inferences from data. Moving beyond tractable priors to TV, finite support, and using external models, we advocated for setting the regularization γ using a Bayesian discrepancy principle. This method explicitly involves human input, so we call it Bayesian because looking at the reconstructions, we as intelligent agents can tell if the model is effective or not, and respond accordingly.

In future work, we will develop methods to infer the point spread function of a given imaging device (SQUIDs in particular). If we image a known source of magnetic field, e.g. a magnetic dipole or a superconducting vortex, finding the PSF is a reconstruction problem in itself. The open-source code is already capable to take a finite PSF into account. The generative model approach combined with Bayesian inference presented here can be readily adapted to evaluate additional new priors for solving the current reconstruction problem. New priors of interest can be found e.g. in recent developments of machine learning. Deep priors [Ulyanov et al., 2018] use the restrictive structure of an untrained convolution neural network as an inductive bias, or implicit prior, and random projectors [Gupta et al., 2018] learn a lower dimensional subspace trained on latent and data-space pairs.

3.6 Appendix

3.6.1 Defining the Flux Model

Recall that our model for a measured magnetic flux is $\phi = Mg + \eta$, for flux image ϕ, current dipole field g, and circulant matrix M representing the convolution of the Biot-Savart law and the PSF. This section derives the Biot-Savart component of M, following the analysis performed by Wijngaarden [Wijngaarden et al., 1996]. The Biot-Savart law is written

$$\mathbf{B}(\mathbf{r}) = \frac{1}{4\pi}\int_V \mathbf{j}(\mathbf{s}) \times \frac{\mathbf{r}-\mathbf{s}}{|\mathbf{r}-\mathbf{s}|^3} d\mathbf{s}, \tag{3.16}$$

where $\mathbf{j}$ is the current density in some volume V.

When inferring $\mathbf{j}$ we will have to optimize over all currents, but because currents are conserved ($\nabla \cdot \mathbf{j} = 0$), this would have to be constrained optimization (which is more difficult than unconstrained). If $\mathbf{j}$ only varies in 2D, i.e. if $\mathbf{j}(\mathbf{r}) = \mathbf{j}(x, y)$, we can enforce conservation by writing

$$\mathbf{j}(x,y) = \nabla \times g(x,y)\hat{\mathbf{z}}, \tag{3.17}$$

then we can perform unconstrained optimization on g. After employing a number of vector identities, we can write the Biot-Savart law as a function of g:

$$\mathbf{B}(\mathbf{r}) = \frac{1}{4\pi}\int_V g(\mathbf{s})\frac{3\hat{\mathbf{n}}(\hat{\mathbf{z}}\cdot\hat{\mathbf{n}}) - \hat{\mathbf{z}}}{|\mathbf{r}-\mathbf{s}|^3}d\mathbf{s}, \tag{3.18}$$

where $g(\mathbf{s})$ only depends on two-dimensions and $\hat{\mathbf{n}} = (\mathbf{r}-\mathbf{s})/|\mathbf{r}-\mathbf{s}|$. The kernel convolved with g in Eqn. 3.18 is then recognizable as the magnetic field of a point dipole, which is why we call g the *current dipole field*; g is a decomposition of a 2D current sheet into circulations of current.

The data we take is discrete, pixelated, and so to proceed we express our current density as rectangular pixels centered below the point at which we measure flux. Squares of constant g are a square loop of constant current present only at the edges. Let us evaluate the field due to a square of constant g. We will only calculated the z-component, that is the component orthogonal to the plane of current as that is the component our probe measures. The field due to a g with a constant value of 1 in a rectangle is

$$B_z^1(\mathbf{x}) = \frac{1}{4\pi}\int_{x_0-\frac{a}{2}}^{x_0+\frac{a}{2}}\int_{y_0-\frac{b}{2}}^{y_0+\frac{b}{2}}\int_0^t \frac{3\mathbf{s}\cdot\hat{\mathbf{z}} - \mathbf{s}}{|\mathbf{s}|^{5/2}}d\mathbf{x}', \tag{3.19}$$

where $\mathbf{s} = \mathbf{x} - \mathbf{x}'$, a and b are the x and y-widths of the rectangle respectively, and t is the thickness of the current.

We can evaluate Eqn. 3.19 using the following indefinite integral:

$$\int \frac{2z^2 - x^2 - y^2}{(x^2+y^2+x^2)^{5/2}}d\mathbf{x} = -\tan^{-1}\frac{xy}{z|\mathbf{x}|}. \tag{3.20}$$

For this work we assume the thickness of the current is much smaller than the distance between the sample and measurement planes ($z \gg t$). Taking the limit $t \to 0$, we find that the magnetic field due to a rectangle of constant g is

$$\begin{aligned} B_z^1(\mathbf{x}, \mathbf{x}_0) = \frac{1}{4\pi} \quad & \left(\mathcal{I}(x_0 - x + \tfrac{a}{2},\ y_0 - y + \tfrac{a}{2}, z) \right. \\ - \quad & \mathcal{I}(x_0 - x + \tfrac{a}{2},\ y_0 - y - \tfrac{a}{2}, z) \\ - \quad & \mathcal{I}(x_0 - x - \tfrac{a}{2},\ y_0 - y + \tfrac{a}{2}, z) \\ + \quad & \left. \mathcal{I}(x_0 - x - \tfrac{a}{2},\ y_0 - y - \tfrac{a}{2}, z) \right), \end{aligned} \tag{3.21}$$

where x_0 and y_0 are the center of the rectangle, a and b are the widths, and

$$\mathcal{I}(x, y, z) = \frac{xy(2z^2 + x^2 + y^2)}{(z^2 + x^2)(z^2 + y^2)|\mathbf{x}|}. \tag{3.22}$$

We can now build up the entire magnetic field due to a current distribution by summing over the discrete current dipole field. Say $g_{ij} = g(x_i, y_i)$ for some set of pixel centers $\{x_i, y_j\}$. Using the linearity of electromagnetism we can write the magnetic field of this current distribution as a sum of the magnetic field due to each individual rectangle of constant g:

$$B(x_k, y_l, z) = \sum_i \sum_j B_z^1(x_k, y_l, z, x_i', y_j') g(x_i', y_j'). \tag{3.23}$$

Since Eqn. 3.21 only depends on the differences between, for instance x_k and x_i', we observe than Eqn. 3.23 is a discrete convolution. This means that we can write our model more succinctly as $b = Mg$, where b is the unraveled magnetic field image, M is a circulant matrix representing the discrete convolution, and g is our unraveled current dipole density image.

Because M is circulant it is diagonalized by plane waves, and matrix-vector products like Mg can be computed very efficiently with Fourier transforms. Once a model for the PSF has been defined, we include this in the definition of M, as two discrete convolutions are just a sequence of multiplications in Fourier space, and then our model is written

$\phi = Mg$, where ϕ is now a magnetic flux, as the addition of the PSF modifies the units of M.

3.6.2 Defining Linear Operators for Priors

It may not be immediately clear how to implement the Gaussian Laplacian prior of eqn. 3.10, the Gaussian Frobenius prior of eqn. 3.11, and the TV Frobenius prior of eqn. 3.13, so I will present here a description of the `pysquid` implementation.

Firstly, all these operators are composed of the partial derivatives ∂_x^2, ∂_y^2, and $\partial_x \partial_y$. These operators need to be generalized to our discrete problem. The most correct way to represent these operators is in Fourier space, where $\partial_x \to -ik_x$ for example is a perfect representation of the derivative operator assuming the operand has been Nyquist sampled. This representation is beautiful for its translation invariance, and it makes solving the psuedoinverse eqn. 3.9 very efficient, but it assumes mirror boundary conditions. Since we almost never come across mirror-symmetric data, we pad out data with zeros so that the right cannot 'see' the left side, but then the Fourier derivatives lead to priors which unfairly penalize variations at the edges of the image.

We overcome this by using finite-difference derivatives, encoded as a sparse matrix, where the interior of the image is computed using centered finite differences [Press et al., 1989], and the edges use forward or backward finite differences, moving away from the edges. This way we can estimate the derivatives using only information that we have. Writing the image $\boldsymbol{g}$ as a two-index matrix $g_{x,y}$, and take D_x^2 for example. For a pixel not at the edge, and assuming that the distance between adjacent pixels is Δx,

$$(D_x^2)_{x',y',x,y} = \frac{\delta_{x',x-1} - 2\delta_{x',x} + \delta_{x',x+1}}{(2\Delta x)^2}, \tag{3.24}$$

and at the left edge, for example,

$$(D_x^2)_{0,y',x,y} = \frac{\delta_{0,x-2} - 2\delta_{0,x-1} + \delta_{0,x}}{(2\Delta x)^2}. \tag{3.25}$$

Then generalizing appropriately, we have discrete linear operators D_x^2, D_y^2 and the cross-derivative D_{xy}^2, and we write matrix vector products $(D_x^2\boldsymbol{g})_i = \sum_i (D_x^2)_{i,j}\ g_j$, for appropriate $i = (x, y)$ and $j = (x, y)$. For specific implementations, see the `pysquid` source code.

With these linear operators defined, we can define the discrete representations of the Gaussian Laplace priors of eqn. 3.10 as

$$\log p(\boldsymbol{g}) \propto -\lambda^2 ||\Gamma\boldsymbol{g}||^2, \tag{3.26}$$

where $\Gamma = D_x^2 + D_y^2$, and $||\cdot||^2 = \sum_i \cdot_i^2$. Throughout this work we will ignore the dimensions of the differential in discretizing the integrals, as it is a uniform rescaling of all terms of the log-posterior eqn. 3.5 which does not change the location of the maximum.

Next, we represent the Frobenius of the Hessian, $\mathrm{Tr}H^T H$ for $H_{\alpha\beta} = \partial_\alpha\partial_\beta g$. Where as the Laplacian Gaussian integrand in eqn. 3.10 is a square of a sum, the Frobenius prior is a sum of squares. In order to write this in terms of some operator Γ, we must stack our operators thusly

$$\Gamma = \begin{pmatrix} D_x^2 \\ D_y^2 \\ \sqrt{2}D_{xy}^2. \end{pmatrix} \tag{3.27}$$

Written this way, then one can see that the discrete generalization of the Gaussian

Frobenius prior in eqn. 3.11 is

$$\begin{aligned}\log p(\boldsymbol{g}) \propto -\lambda^2 ||\Gamma \boldsymbol{g}||^2 &= -\lambda^2 (\Gamma \boldsymbol{g})^T \Gamma \boldsymbol{g} \\ &= -\lambda^2 \Big(\sum_i (D_x^2 g)_i^2 + \sum_i (D_y^2 g)_i^2 + \\ &\qquad \sum_i (D_{xy}^2 g)_i^2 \Big). \end{aligned} \tag{3.28}$$

Finally, the Total Variation prior simply modifies eqn. 3.28 by taking a square root of the summand to obtain

$$\begin{aligned}\log p(\boldsymbol{g}) \propto -\lambda^2 \Big(&\sum_i (D_x^2 g)_i^2 + \sum_i (D_y^2 g)_i^2 + \\ &\sum_i (D_{xy}^2 g)_i^2 \Big)^{1/2}. \end{aligned} \tag{3.29}$$

Note that following Vico et al.[Vico et al., 2016], it is possible to find truncated kernels, which allow the computation of convolutions in a finite domain. This would allow us to construct derivative translation-invariant operators which do not require zero-padding around the domain. This would allow direct solution of the psuedo-inverse, and speed up all of our calculations by a factor of 4×. Future work could explore this for improved efficiency of `pysquid`, and perhaps even allow analytic results for the new Frobenius Hessian prior.t

3.6.3 ADMM for Total Variation Deconvolution

Alternating Difference Method of Multipliers (ADMM)[Boyd et al., 2011] is a convex optimization algorithm which solves

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} \qquad & f(\boldsymbol{x}) + g(\boldsymbol{z}) \\ \text{subject to} \qquad & A\boldsymbol{x} + B\boldsymbol{z} = \boldsymbol{c}, \end{aligned} \tag{3.30}$$

for some scalar functions f and g, appropriately-sized matrices A and B, and vector $\boldsymbol{c}$. In this section only, g refers to the scalar function of the ADMM algorith, and not the scalar current dipole field g. The only requirement for ADMM to provably solve 3.30 is that f and g be convex in their arguments.

For solving current reconstruction we must optimize Eqn. 3.6. We can cast our problem into the standard form of Eqn. 3.30 by identifying $\boldsymbol{x} \equiv \boldsymbol{g}$, and by setting

$$f(\boldsymbol{g}) \quad = \tfrac{1}{2}||M\boldsymbol{g} - \boldsymbol{\phi}||^2. \tag{3.31}$$

Then the function g is the regularization term set by $-\log p(\boldsymbol{z})$. Our isotropic total variation prior penalizes second derivatives of the $\boldsymbol{g}$-field (penalizing changes in the current $\boldsymbol{j}$). Identifying A with the x- and y-derivative matrices D_x^2 and D_y^2 and D_{xy}^2,

$$A = \begin{pmatrix} D_x^2 \\ D_y^2 \\ D_{xy}^2 \end{pmatrix}, \tag{3.32}$$

setting $B = -\mathbb{I}$, and $\boldsymbol{c} = 0$, the stipulation that $A\boldsymbol{g} + B\boldsymbol{z} = \boldsymbol{c}$ is equivalent to

$$\begin{pmatrix} D_x^2 \\ D_y^2 \\ D_{xy} \end{pmatrix} \boldsymbol{g} = \boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_x \\ \boldsymbol{z}_y \\ \boldsymbol{z}_{xy} \end{pmatrix}, \tag{3.33}$$

where $\boldsymbol{z}$ is twice as long as $\boldsymbol{g}$, containing both the horizontal and vertical second derivatives of $\boldsymbol{g}$. The final piece is then the total variation of the Frobenius norm of the Hessian:

$$g(\boldsymbol{z}) = \lambda^2 \sum_i \sqrt{z_{x,i}^2 + z_{y,i}^2 + 2z_{xy,i}^2}. \tag{3.34}$$

We can modify ADMM to include finite support priors by replacing $\boldsymbol{g} \to F\tilde{\boldsymbol{g}}$, and optimizing $\tilde{\boldsymbol{g}}$ instead of $\boldsymbol{g}$:

$$\begin{aligned} \min_{\tilde{\boldsymbol{g}},\boldsymbol{z}} \quad & f(F\tilde{\boldsymbol{g}}) + (\boldsymbol{z}) \\ \text{subject to} \quad & AF\tilde{\boldsymbol{g}} = \boldsymbol{z}, \end{aligned} \tag{3.35}$$

where f is still Eqn. 3.31, g is Eqn. 3.34, and A is Eqn. 3.32. Equivalently, we can modify the kernel matrix $M \to MF$ and the second-derivative matrix $A \to AF$, which is how we actually implemented it.

We can also modify ADMM to use an external model as in eqn. 3.15, by simply setting $\boldsymbol{c} = -A\boldsymbol{g}_{\text{ext}}$ for the external model current dipole field in the field of view $\boldsymbol{g}_{\text{ext}}$.

Chapter 4

Information Geometry and Effective Hamiltonians of Spin Glasses

4.1 Introduction

Originally developed to study dilute solutions of manganese in copper, spin glass models have become the prototype for studying disordered complex systems, especially those with rough energy landscapes. Over the last four decades, variants of spin glasses have had important applications in computational complexity [Bachas, 1984, Fu and Anderson, 1986, Kirkpatrick, 1984, Mézard et al., 2002], neuroscience and machine learning [Amit et al., 1985a, Hertz et al., 1991], HIV drug resistance [Shekhar et al., 2013b,a], memory models [Amit et al., 1985b], the supreme court [Lee et al., 2015], and protein folding [Weigt et al., 2009, Marks et al., 2012, Stein, 1985]. Here we focus on the two-dimensional

The work constituting this chapter was done in collaboration with Danilo B. Liarte and James P. Sethna

Ising spin glass model, which is not only used to model physical systems but also is also applied to the study of deep neural network energy landscapes [Choromanska et al., 2015], which are famously difficult to interpret. Unlike other models, such as the NP-complete three-dimensional Ising spin glass [Barahona, 1982], the 2D model is amenable to polynomial-time, numerically exact sampling methods [Thomas and Middleton, 2009] and its behavior is relatively well understood [Hartmann, 2011, Arguin et al., 2010]. Thus the 2D Ising spin glass can be used as a playground for building interpretable theories of complex systems.

Here we apply recent general techniques for model reduction [Transtrum and Qiu, 2014] inspired by information geometry [Transtrum et al., 2011] to develop a reduced coarse-grained Hamiltonian for the two-dimensional Ising spin glass. First we demonstrate the sloppiness of spin glasses at low temperatures, elucidating the model's effective low-dimensionality. We then introduce a scheme for exploiting this property to find explicit coarse-grained lower-dimensional effective Hamiltonians. We employ information theoretic clustering to assign the coarse-grained block 'nuggets' of spins and use the inverse Ising algorithm to find effective couplings between the nuggets. We show that these effective Hamiltonians retain some of the low-temperature information by cooling them and comparing to the original Hamiltonian. Using ideas from machine learning and variational mean field theory, we then propose an improved dimension-reduction learning algorithm that has attractive connections to the Renormalization Group. We leave the implementation and analysis of this method to future work.

4.2 Sloppiness of Lattice Models

Sloppiness is the observation that multiparameter nonlinear models are usually sensitive to the values of only a few linear combinations of parameters, with each successive 'stiff' combination roughly two to ten times less important to the collective behavior [Brown and Sethna, 2003, Gutenkunst et al., 2007]. Information geometry and interpolation theory tells us that these models have a space of possible predictions (the 'model manifold') that is effectively low-dimensional [Transtrum et al., 2010, Sethna, 2015]. This manifold is usually thin in data-space directions corresponding to sloppy directions in parameter space, forming a *hyperribbon* with a geometrical hierarchy of widths corresponding to the hierarchy of sensitivities of the behavior near a particular parameter set [Transtrum et al., 2011].

Sloppiness is a ubiquitous feature of complicated models in science. It has been identified in models of gene regulation and cell-signaling [Waterfall et al., 2006], radioactive decay, particle accelerators, neural networks, and quantum wave functions [Transtrum et al., 2015, Gutenkunst et al., 2007]. Sloppiness has been found in coarse-grained models through a beautiful connection to renormalization group (RG) flows of the Ising model and discrete diffusion [Machta et al., 2013]. Here we will show that Gaussian spin glasses for $D = 3$ are sloppy at low temperature, due to the broad distribution of activation energies which characterize the glass phase.

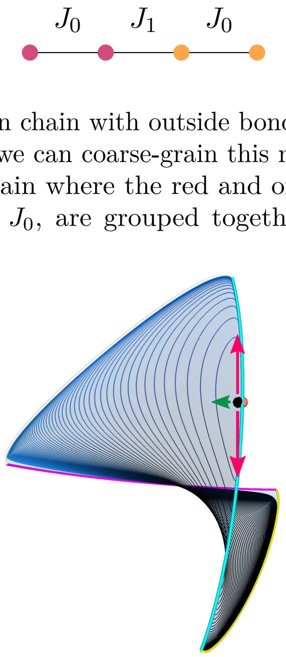
In what way are spin glasses illustrative of sloppiness? An Ising spin glass can be viewed as a model that predicts the probability $p(\mathbf{s}) = \exp(-\beta\mathcal{H}(\mathbf{s}))$ of the 2^N configurations of N spins $s_i = \pm1$, with

$$\mathcal{H} = -\sum_{\langle ij\rangle} J_{ij} s_i s_j + \sum_i h_i s_i. \tag{4.1}$$

Here the bonds J_{ij} are independent Gaussian random variables of zero mean and unit variance $P(J) = \mathcal{N}(0,1)$, connecting only nearest neighbor spins on the lattice. We shall set the random fields $h_i = 0$ (but later will consider the sensitivity of the predictions to varying them).

We consider first a set of $N = 4$ spins along a straight line (Fig. 4.1a), but (for illustrative purposes) fixing the outside bonds to be equal. There remain two parameters (bond strengths) in this model $\theta = \{J_0, J_1\}$, which act as coordinates for the model manifold. The model manifold is a Riemannian manifold, with a natural metric which can be understood as the relative entropy or local Kullback-Leibler divergence between nearby parameter sets. In the bond-strength coordinates this metric tensor is given by the Fisher information matrix (FIM) [Cover and Thomas, 2012]. The FIM can be understood as the flat Euclidean metric on the unit sphere restricted to the positive quadrant, where the restriction of lying on a sphere replaces the normalization condition. Therefore changing coordinates to $y_i = \sqrt{p_i}$ will result in a locally flat manifold. Figure 4.1b shows a projection of this isometrically embedded two-dimensional model manifold onto its two largest principle components.

We can understand some of the consequences of sloppiness by considering several dozen samples from the spin chain, choosing $J_0 \gg J_1$, plotted as a black dot in fig. 4.1b. We can infer the parameters θ which produced the data sample by maximizing the likelihood. This problem is often called the inverse Ising problem, and is equivalent to finding the point on the model manifold closest (as measured by the FIM) to the data. The solution of this optimization is plotted as a red dot in fig. 4.1b. Figure 4.2 shows contours of the log-likelihood (our 'cost' function) in parameter (θ) space, with the best-fit denoted by the block dot. These same contours are the blue lines on the model manifold in fig. 4.1b, explicitly showing how our spin chain maps bond-strengths to probabilities.



(a) Four-spin chain with outside bonds identified. If $J_0 \gg J_1$ we can coarse-grain this model into a two-spin chain where the red and orange spins, coupled by J_0, are grouped together into two 'nuggets.'

(b) Cross-section of an isometric embedding of the probabilities of the four-spin chain in fig. 4.1a, parameterized by the bonds J_0 and J_1. The black dot (slightly above the model manifold) is data from a finite sample where $J_0 \gg J_1$, the contours are lines of constant likelihood of this data point, and the red dot (partially hidden by the black dot) is the maximum likelihood point. The red arrows indicate the stiff directions, the green arrow indicates a sloppy direction, where both correspond to parameter directions in fig. 4.2.

Figure 4.1: Visualizing the model manifold of a small Ising model. The Ising model can be viewed as a mapping from bond strengths to state probabilities, where sloppiness can be understood by ordering the parameter directions by how much they effect model behavior. Information geometry [Transtrum and Qiu, 2014] suggests coarse-graining by approximating models by their boundary, in this case achieved by taking $J_0 \to \infty$.

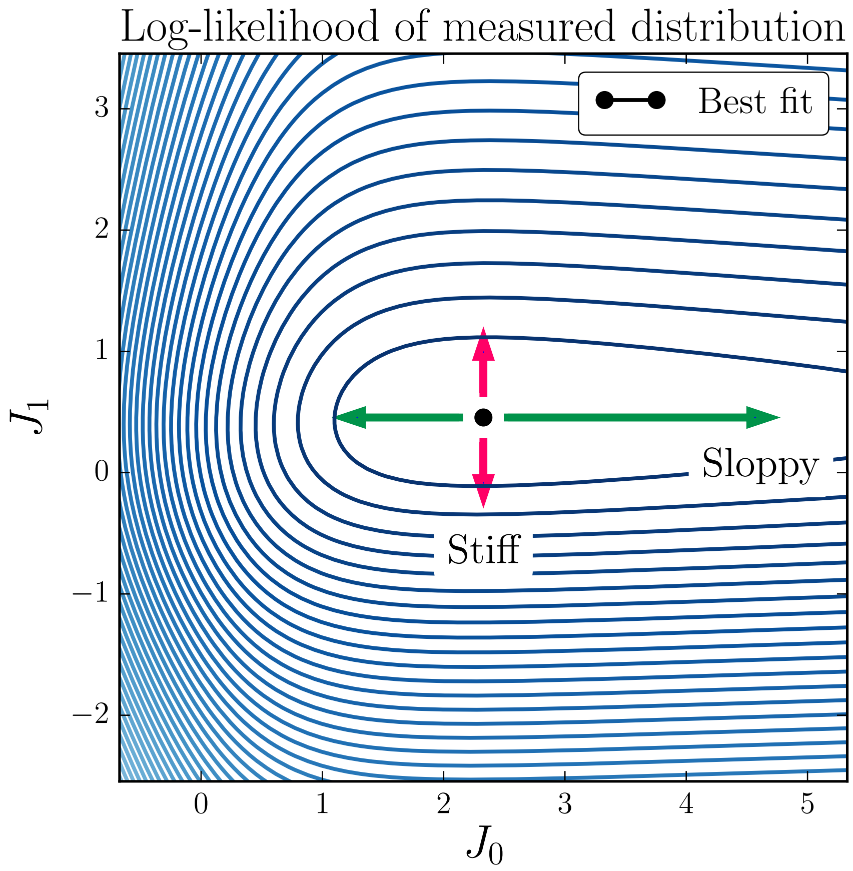


Figure 4.2: Contours of constant probability for a finite sample of the four-spin chain in fig. 4.1a. The black dot indicates the maximum likelihood point (best-fit), and the labeled arrows indicate the stiff and sloppy parameter directions. Information geometry and sloppiness suggests coarse-graining by approximating the model by its boundaries. This occurs in the limit $J_0 \to \infty$.

The nonlinearity of this mapping is important because it allows the model-manifold to have edges as either bond strength is taken to $\pm\infty$.

We chose $J_0 \gg J_1$ so that the spins connected by J_0 are almost always aligned. In this case there is less information to restrict J_0, so J_0 will be a sloppy direction and J_1 stiff. This is demonstrated in fig. 4.2, as the contours of constant likelihood are more closely packed in the J_1 direction. Examining fig. 4.1b we see that changing J_1 effects the likelihood more than J_0 because it moves farther along the model manifold. This relationship is indicated by the red arrows in figs. 4.1b and 4.2. The green arrows show the relationship between parameter and data space for the sloppy direction, except that the $+J_0$ arrow is invisible in fig. 4.1b as the best-fit is too close to the boundary.

Transtrum et al. [Transtrum and Qiu, 2014] suggest using proximity to the boundary to coarse-grain the model by taking $J_0 \to \infty$. For our four-spin model this procedure results in an effective two-spin model on the boundary of the model manifold which is 'close' to the original as measured by the FIM. This approximation is justified because the sloppy parameter direction is less important for model predictions. In general the sloppy and stiff parameter directions are linear combinations of parameters. In this case Transtrum et al. follow geodesics of the model manifold from the best-fit, in the sloppy direction. As the boundary is approached the components of the sloppy parameter direction go to zero or infinity, resulting in simpler formulas for the model. For lattice-spin models this parameter reduction is analogous to the Renormalization Group, where in real space 'block-spins' are found which have the same statistics of the microscopic model.

Integrating the geodesic equations of a lattice model would require a Monte Carlo sampling at each point along the geodesic, which would be prohibitively expensive for coarse-graining spin glasses. We suggest an alternative to following geodesics, which we argue will find a similarly near-optimal reduced model. In our four-spin example, instead of taking the limit $J_0 \to \infty$ to get a 'nearby' two-spin model, we notice that the spins connected by J_0 are highly correlated, identifying the two pairs of spins as 'nuggets,' indicated by the red and orange colors in fig. 4.1a. We then compute nugget 'samples' by projecting each nugget configuration onto its most common state, and infer the bonds connecting the nuggets by maximizing the likelihood of a two-spin chain. The resulting effective two-spin model will have the same statistics as the nuggets, and therefore be statistically close to the original chain in the same way that the data in fig. 4.1b is close to the boundary of the model manifold.

Our spin chain example is contrived, however it motivates our approach as spin glasses also contain groups of highly correlated spins called droplets. Droplets are complex, subsuming and overlapping at all length scales. With this in mind we define non-overlapping 'nuggets' to be the intersection of the most 'important' low-energy droplets of the spin glass. Like the droplets in the droplet theory of spin glasses [Fisher and Huse, 1986, 1988, Huse and Fisher, 1987, Bray and Moore, 1984, 1987b] the nuggets are well-approximated by two-state systems—individual spins—inspiring a nugget-spin renormalization. We find effective nugget Hamiltonians with an inverse-Ising algorithm, from which activation free energies and nugget interactions can be calculated. The coarse-grained nugget Hamiltonian is another spin glass from which we can sample. The nugget Hamiltonian can be viewed as the limit of the intra-nugget bonds going to infinity, approximating the spin glass model-manifold by its boundaries on which the nugget-Hamiltonian lives.

Among the development of exciting interdisciplinary work still lurks controversy over the appropriate description of realistic, 3D spin-glasses. Two competing theories are replica symmetry breaking (RSB) [Mézard et al., 1984] and the droplet theory [Bray and Moore, 1984, 1987b, Fisher and Huse, 1986, Huse and Fisher, 1987, Fisher and Huse, 1988, Newman and Stein, 1992, 1996, 1998]. Droplet theory predicts two ground states and physics dominated by compact regions of thermally excited spins. RSB—which is exact for mean-field—predicts a fantastic hierarchy of infinitely-many ground states. We will avoid the debate by discussing spin glasses in 2D where there is strong evidence that droplet theory is a good description [Arguin et al., 2010]. Though all of the methods we develop can be applied in higher dimensions, we leave that for future work as even obtaining samples of $D = 3$ spin glasses at low temperature is much more expensive.

The droplet theory is a scaling theory for low-energy, large scale excitations. These excitations are thought to be regions of size L which flip coherently and have a typical free energy asymmetry $F_L \sim L^\theta$ (also called the *activation energy*). This follows from the central ansatz that the density of states of F_L scales as $P(F, L) = \rho(F/L^\theta)/L^\theta$ where $\rho(0) \sim 1$. θ is called the stiffness exponent, and is less than zero in 2D which means that F_L generally decreases with L and thus there can be no ordering for $T > 0$. Two other exponents govern the behavior of droplets: the barrier height free energy exponent ψ [Huse and Fisher, 1987] and the fractal dimension d_f.

In two dimensions the existence of efficient algorithms for finding ground states have enabled precise studies of θ and d_f [Hartmann, 2011], while the NP-completeness of determining barrier heights has limited our knowledge of ψ [Middleton, 1999]. Existing methods find low-energy excitations at zero-temperature by computing the ground state for some disorder configuration, perturbing some of the bonds, then finding the perturbed ground state. θ is determined by associating the difference between these energies with the scale set by the system size, and the domains of the perturbed ground state can be used to measure d_f. These methods are considered a kind of real-space renormalization over the scale of the system size [McMillan, 1984]. In contrast with these existing approaches, we identify low-lying excitations by studying the correlation functions inspired by sloppiness and information geometry.

4.3 Spin Glasses are Sloppy!

4.3.1 Edwards-Anderson Model

We will study the model manifold of the 2D Edwards Anderson model with added random field, for which configurations of N spins $s_i \in \{+1, -1\}^N$ on a square lattice have energy prescribed by

$$\mathcal{H} = -\sum_{\langle ij \rangle} J_{ij} s_i s_j + \sum_i h_i s_i \tag{4.2}$$

where $P(J) = \mathcal{N}(0,1)$ and $\langle ij \rangle$ denotes nearest-neighbor sites. Throughout this paper we wil set $h_i = 0$. It will be useful to write our Hamiltonian in another notation: $\mathcal{H} = -\theta^\mu \phi_\mu$ where $\theta^\mu = J_{ij}$ and $\phi_\mu = s_i s_j$, where we have assumed Einstein's summation convention.

4.3.2 Fisher Information

The fundamental object for understanding sloppiness is the Fisher Information Matrix (FIM), which measures the sensitivity of a probability distribution to its parameters. We will consider the sensitivity of our model to changes in bonds (bond-FIM) and fields (field-FIM). The FIM of some probability distribution $P_\theta(x)$ parameterized by θ_μ is defined by

$$g_{\mu\nu} = -\sum_x P_\theta(x) \frac{\partial^2}{\partial\theta_\mu \partial\theta_\nu} \log P_\theta(x). \tag{4.3}$$

The FIM tells us which parameter combinations are stiff and sloppy; which move along the longest and shortest directions of the hyper-ribbon. Further, via the Cramer-Rao bound the *inverse* of the FIM tells us the minimum achievable variance on estimating parameters [Cover and Thomas, 2012]. In other words, g^{-1} tells us how well we could in principle estimate θ. Diagnosing sloppiness consists of computing the eigenvalues of $g_{\mu\nu}$ and checking that they span many decades and are roughly equally spaced in logarithm.

Eigenvalues of the FIM equally spaced in log is equivalent to a hierarchy of widths of the model-manifold [Transtrum et al., 2011]; stiff directions are eigenvectors associated with large eigenvalues and thick directions of the model manifold, sloppy with small eigenvalues and thin directions on the hyper-ribbon.

4.3.3 FIM of Spin Glasses

Naturally we are interested in the manifold swept out by the Boltzmann distribution $P_\theta(s) = \frac{1}{\mathcal{Z}} e^{-\beta \mathcal{H}_\theta(s)}$. We will consider for θ both nearest-neighbor bonds J_{ij} (bond-FIM) and local fields h_i (field-FIM). We can use Eq. 4.3 and the definition of the free energy $\mathcal{F} = -\beta^{-1} \log \mathcal{Z}$ to write

$$g_{\mu\nu} = -\beta \frac{\partial^2}{\partial\theta_\mu \partial\theta_\nu} \mathcal{F} = \beta^2 \left(\langle \phi_\mu \phi_\nu \rangle - \langle \phi_\mu \rangle \langle \phi_\nu \rangle \right), \tag{4.4}$$

where $\langle \cdot \rangle$ denotes averaging over $P_\theta(s)$, $\phi_\mu = s_i s_j$ if $\theta_\mu = J_{ij}$ (bond-FIM) and $\phi_i = s_i$ if $\theta_\mu = h_i$ (field-FIM). The field-FIM is $\beta^2 c_{ij}$, where c_{ij} is the two-point correlation function, and the bond-FIM is a four-spin correlation function, the covariance of nearest-neighbor bond orientations.

4.3.4 Signatures of Sloppiness

Using Eq. 4.4 we can compute the bond-FIM from samples of a spin glass with quenched disorder. Figure 4.3 shows the eigenvalues (left) and eigenvectors (right) of the bond-FIM of a $L = 16$ spin glass with Gaussian disorder. The bond-FIM in the figure was computed from 20,000 exact samples obtained from the method of Thomas and Middleton [Thomas and Middleton, 2009], which maps planar spin glasses onto dimer matching which can be sampled without Monte Carlo. The first thing to notice is that the eight stiffest eigenvectors trace out striking rings. We will argue that these rings are

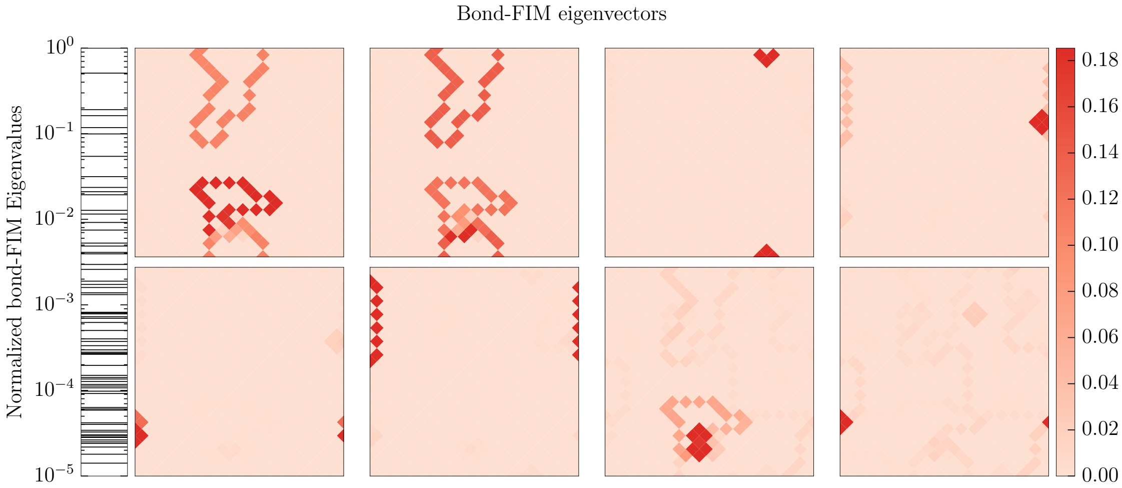


Figure 4.3: (left) FIM eigenvalues and eigenvectors for a periodic L =16 EA model at T=0.1. The top-left panel is the eigenvector associated with the largest eigenvalue, the lower-right panel is associated with the eighth-largest eigenvalue. The rings are the boundary of the thermally active droplets.

the boundaries of the lowest energy excitations, which are associated with the droplets of droplet theory[1]. The bond-FIM highlights the accidental degeneracies of the spin glass as those most sensitive to perturbation; changing bonds in these rings traverses the long and thick directions of the model manifold. The bonds inside the droplets are not sensitive, pushing these bonds moves along the thin direction of the model manifold. Our effective Hamiltonians will take these interior bonds to infinity, approximating the model manifold by the boundaries of the long and thick directions. An important feature of the eigenvectors is the 'ghosts' of adjacent droplets in multiple eigenvectors; the orthogonality of eigenvectors causes unnecessary superpositions of the boundaries, so we will study the droplets indirectly by clustering spins into nuggets.

The eigenvalue distribution is sloppy; the eigenvalues are spread uniformly in log

[1] We mean the rings in the eigenvectors are caused by droplet excitations, but the relationship between the droplets and the eigenvectors is not simple.

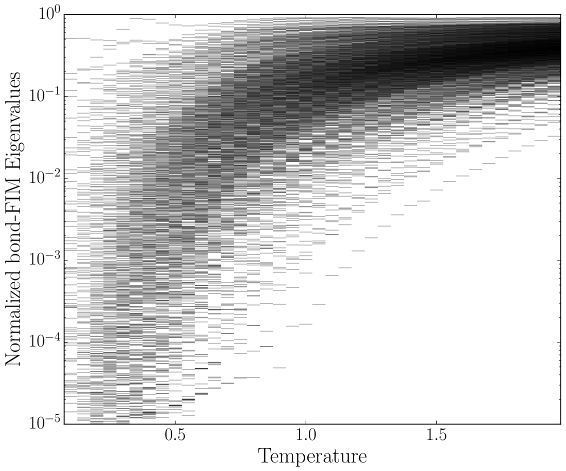


Figure 4.4: Normalized eigenvalues of the bond-FIM of a $L = 16$ EA model as a function of temperature. As the temperature is lowered the eigenvalues spread out and become sloppy.

scale across many decades and the model manifold of a spin glass is a hyper-ribbon. In this disorder sample the eighth eigenvalue is about fifty times smaller than the largest eigenvalue, so the excitations in the eighth eigenvector is fifty times ‘less important’ to the behavior of the spin glass than the droplet in the first. Sloppiness tells us what we already know, that the droplets dominate the physics at low temperature. Figure 4.4 shows how the eigenvalues of the bond-FIM change with temperature. The spectrum develops clear signatures of sloppiness at low temperature. We will argue that sloppiness in our model arises when the distribution of droplet activation energies is broad relative to the temperature; the spin glass phase is sloppy.

4.4 Bond-FIM for Independent Droplets

We can analytically compute the eigendecomposition of the bond-FIM in the case of non-interacting and non-overlapping droplets. We assume the only degrees of freedom are contiguous regions of spins (droplets) with only two states $c^i = \pm 1$ differing by an activation energy Δ^i. This scenario is inspired by the low-temperature results of fig. 4.3, and the seminal work describing glasses as collections of two-state systems [Anderson et al., 1972, Phillips, 1987]. We assume the droplets are non-overlapping and do not interact so that $P(\{c^i\}) = \prod_i P(c^i)$. We also assume the ground-state gauge, i.e. given the ground state s_i^0 we transform the spins as $s_i \to s_i^0 s_i$ and the bonds as $J_{ij} \to s_i^0 s_j^0 J_{ij}$.

Perturbing all the bonds by θ_μ, droplet i has two states with energies $E_{\text{inactive}} = E_0 + \sum_\mu \theta_\mu$ and $E_{\text{active}} = E_0 + \sum_\mu \theta_\mu (1 - \delta_\mu^i) - \sum_\mu \theta_\mu \delta_\mu^i$. E_0 is the ground state energy and δ_μ^i is 1 if bond μ is on the boundary of droplet i and zero otherwise. $P(c^i) = e^{-\beta E(c^i)} / \mathcal{Z}^i$ with $\mathcal{Z}^i = 1 + \exp -\beta(\Delta^i - 2\sum_\nu \theta_\nu \delta_\nu^i)$, from which we can directly calculate:

$$\begin{aligned} g_{\mu\nu} &= \sum_i \frac{\partial^2 \log \mathcal{Z}^i}{\partial\theta_\mu \partial\theta_\nu} \\ &= \sum_i \delta_\mu^i \delta_\nu^i \beta^2 \text{sech}^2 \frac{\beta\Delta^i}{2} && (4.5) \\ &\approx 4\beta^2 \sum_i \delta_\mu^i \delta_\nu^i e^{-\beta\Delta^i} \quad \text{for} \quad \beta >> \Delta^0, && (4.6) \end{aligned}$$

where in Eq. 4.5 we have taken the perturbation to zero and in Eq. 4.6 we have assumed β is much larger than the smallest activation energy Δ^0. The low-temperature bond-FIM $g_{\mu\nu}$ is a block diagonal matrix:

$$g_{\mu\nu} = 4\beta^2 \left(\begin{array}{c|cc} e^{-\beta\Delta^0} & & \\ \hline & \boxed{e^{-\beta\Delta^1}} & \\ & & \ddots \end{array}\right),$$

where the ith block is proportional to a square matrix of ones with dimension of the perimeter p^i of droplet i. We assume an ordering where $\Delta^i < \Delta^{i+1}$. Each block has one nonzero eigenvalue and $p^i - 1$ zero eigenvalues. The nonzero eigenvalues and associated eigenvectors of the entire matrix are

$$\lambda^i = 4\beta^2 p^i e^{-\beta\Delta^i}; \quad v^i_\mu = \frac{\delta^i_\mu}{\sqrt{p^i}}. \tag{4.7}$$

Defining $\tilde{\lambda} = \lambda/4\beta^2$ we see that $\log\tilde{\lambda}^i = -\beta\Delta^i + \log p^i$. In the glass phase the Δ^i are broadly distributed; we conclude that a spin glass model of independent droplets is sloppy due to the broad distribution of activation energies relative to temperature. At low enough temperature our spin glasses should be well-described by this theory, and so we conclude that the onset of sloppiness shown in fig. 4.4 is due to the broadening of the activation energies. The eigenvectors also uniformly highlight the boundaries of each droplet, unlike fig. 4.3 which has ‘ghosts’ of droplets across eigenvectors. This difference is clearly because we assumed independent droplets. The orthogonal basis of a correlation matrix is formed by statistically independent combinations. When there are correlations between droplets the statistically independent field combinations become a complicated combination of the interacting droplets.

The largest eigenvalue λ^0 of Eq. 4.7 likely corresponds to the droplet with the smallest Δ^i. Since we are confining ourselves to two dimensions, $\theta < 0$ and the droplet with the smallest energy may span the finite size of the system. In this case $l^0 = L$ and $\Delta^0 \sim L^\theta$ [Fisher and Huse, 1988], so

$$\log\tilde{\lambda}^0 \sim -\beta L^\theta + \log p^i. \tag{4.8}$$

The largest bond-FIM eigenvalue may therefore be a measure of droplet excitations on the scale of the system size. At low temperature Eq. 4.8 is likely similar to other measurements of zero-temperature droplet excitations [Hartmann and Young, 2002, Hartmann and Moore, 2004, Kawashima, 2000, Kawashima and Aoki, 1999]. Such measurements are well-known to suffer from large corrections to scaling due to the relative independence of droplets of different size [Middleton, 2001, Hartmann and Moore, 2003]; the lowest energy excitation does not always span the system size. The extraction of the stiffness exponent θ from the largest bond-FIM eigenvalues are likely corrupted by these corrections as well.

The same calculation can be performed for the field-FIM of independent droplets. The dominant field-FIM eigenvalue will correspond to the breaking of the ground state symmetry (which has zero activation energy). In this case the probability of droplet configurations does not factor; the field-FIM is complicated for independent droplets, just as complicated as the bond-FIM for interacting droplets.

In summary, at low temperature, sloppiness is equivalent to $e^{-\beta\Delta^i}$ being equally spaced in logarithm, which is caused by the independent low-energy excitations of the spin-glass phase. The orthogonality of eigenvectors leads to droplets being mixed up, we will avoid this problem by finding the droplets with a clustering method.

4.5 Finding the 'Nuggets' Within the Boundaries

Our results suggest the rings in the bond-FIM are the boundaries of the active droplets, so naturally we want to study the volume of spins they contain. Unfortunately the 'ghosts' in the eigenvectors make it difficult to isolate the droplets, or even be sure what exactly constitutes a droplet. Droplet theory allows the active droplets to overlap and subsume each other; in order to proceed we will study 'nuggets.' We define nuggets to be the volumes separated by the intersection of all the rings of the stiffest bond-FIM eigenvectors. Nuggets are intersections of the subset of 'important' droplets as defined by

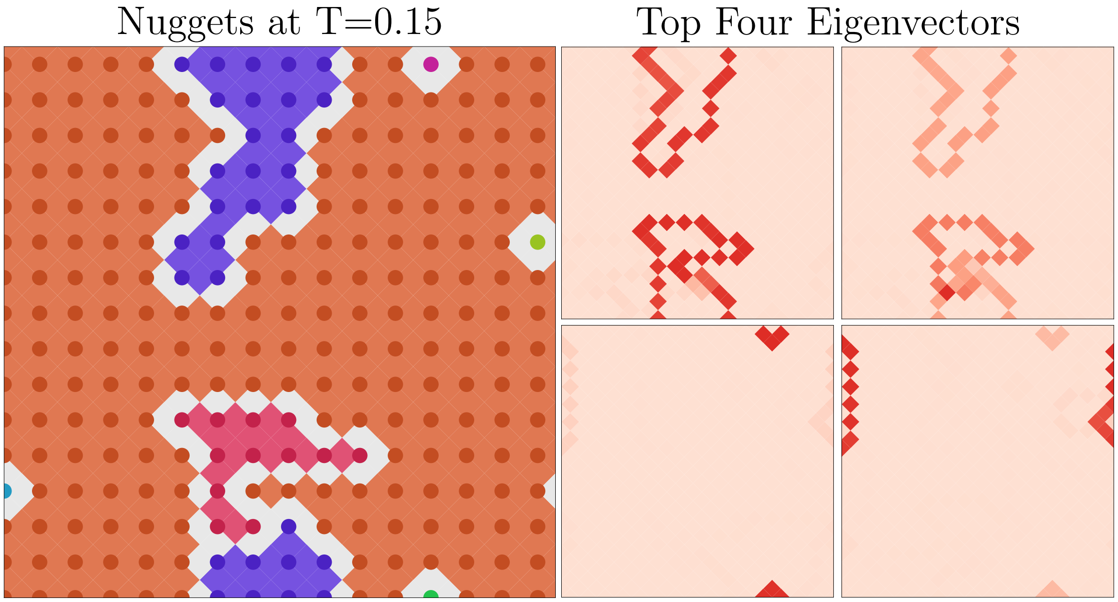


Figure 4.5: Left: Nuggets (spins of constant color) from the information theoretic clustering, with the stiffest eigenvector behind. Right: Top four bond-FIM eigenvectors to show the relationship between nuggets and droplet boundaries.

the stiffest eigenvectors of both the FIMs. Some of the nuggets will be active droplets and some will be chunks of an active droplet.

We will identify the nuggets by clustering spins according to the *field-FIM* $\beta^2 c_{ij}$. More precisely we will cluster spins according to the similarity measure $|c_{ij}|$. The spin-spin correlation function is the sufficient statistic for the bonds of our spin glass model. Sufficient statistics are measures of samples of a probability distribution which are maximally informative about the parameters of the distribution [Cover and Thomas, 2012]. In other words, given c_{ij}, no other correlation function will contain more information about the bonds of our system, from which the droplets must emerge. Therefore it should not be surprising that we only need the field-FIM $\beta^2 c_{ij}$, and not the bond-FIM to identify the nuggets/droplets.

Figure 4.5 (left) shows the nugget decomposition of the same sample of disorder as in figs. 4.3 and 4.4, where circles of a constant color indicate nugget membership, shown with the stiffest eigenvector in the background. The right side of fig. 4.5 shows the top four eigenvectors, so that we can compare the droplet boundaries with nuggets. Notice that the rings in the top two eigenvectors could be two correlated droplets or one large droplet with a sub-droplet; there are precisely two nuggets in the same volume.

We obtain the nugget decomposition by clustering spins according to the spin-spin similarity measure $s_{ij} = |c_{ij}|$ using information-based clustering [Slonim et al., 2005]. This is a 'soft' clustering: it finds the probability that an item is assigned to a cluster $P(i|c)$ by optimizing $\langle s \rangle_P - \lambda I[P]$, the average similarity less the information cost I of specifying the clusters. λ is a temperature-like variable, in analogy with free energies, and weights the clustering complexity against the average similarity.

Similar nuggets can be obtained through more convential agglomerative clusterings, but the information clustering has certain conveniences. We assign spin i to whichever cluster has the highest probability in $P(i|c)$, then divide the assigned clusters into simply connected regions. In this way we remove sensitivity to the arbitrary choice of the number of clusters. Figure 4.5 shows the result of this procedure for the same disorder as in figs. 4.3 and 4.4. The clustering temperature is also not very sensitive, in the sense that a wide range of values identifies the 'important' nuggets highlighted by the stiffest bond-FIM eigenvectors. We used our own GPU implementation of the information clustering algorithm for the analysis in this paper [Clement, 2016].

4.5.1 Properties of nuggets

Clustering is a form of compression, wherein individual spin labels are forfeited to the label of the entire nugget, a more useful and expressive unit. How effectively can the states of our nuggets compress the states of our spin glasses? Given a spin configuration of a spin glass s^i we can look at the magnetization of each nugget with respect to the ground state, $m_c^t = \sum_{i \in c} s_i^0 \cdot s_i^t / |c|$, where s_i^t are the spins in configuration t and $|c|$ is the number of spins in nugget c. We take the most common spin configuration as the ground state because we are using exact sampling. If we make a histogram of m_c^t for each nugget, we see that almost every nugget state has $m_c^t = \pm 1$; the nuggets are effectively two-state systems.

We can see this more clearly for all nuggets at all temperatures if we examine the entropy of each nugget. Let $S[P(m_c)]$ be the entropy of the probability of the nugget magnetization. Figure 4.6 shows the number of nuggets with some entropy as a function of entropy and temperature. For all temperatures we studied (even high temperatures away from the glass phase) most nuggets have an entropy of 1-bit, and so are well described by a two-state system. Further, the largest nugget entropy is about 1.5 bits, equivalent to a 3-state system. More concretely, our original $L = 16$ system has 2^{256} possible states, whereas the nuggets have only about 2^9 states (there are 9 nuggets at T=0.14), an exponential compression. We conclude that the nuggets are an effective compression of the spin glass, even at high temperature.

4.6 Effective Nugget Hamiltonians

Early work on the renormalization group by Kadanoff studied the flows of the couplings of an Ising model under coarse-graining by grouping spins in a block and compressing the block state into a single summary of its members [Kadanoff, 1967]. So-called block-spin renormalization, the procedure is intuitive and predicts the scaling laws observed near critical phenomenon. As Wilson pointed out though, the exact region it is meant to

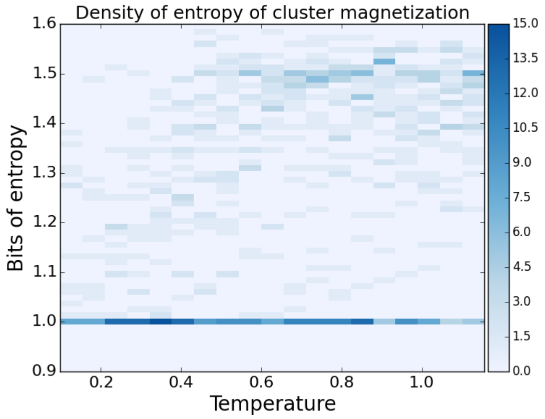


Figure 4.6: The distribution of $S[P(m_c)]$, the entropy of the probability of nugget magnetizations. The majority of nuggets lie on the 1 bit line, and 1 bit is two states. Even at a high temperature of 1 the highest entropy a nugget achieves is 1.6, which is about three states. Therefore the clustering is performing very effective compression.

study—near the phase transition—has zero magnetization and so treating a block of spins as 'up' or 'down' is not a good approximation [Wilson, 1971]. We have shown that our nuggets are well-approximated as two-state systems, so we propose a 'nugget-spin renormalization,' shown schematically in fig. 4.7, where on the left, all spins of a constant color are treated as a block-spin, and the graph on the right is the structure of the effective Hamiltonian after tracing out the internal degrees of freedom. We are following the results of information geometry by approximating the model manifold only by the long and thick boundaries. Each m-nugget configuration is an m-dimensional 'edge' of the model manifold.

The geometric irregularity of our nuggets means we cannot pursue any known analytic renormalization schemes. Instead, we use a solution to the inverse Ising problem, which is: given spin configurations $\{s_i\}$ presumed to be sampled from a Hamiltonian of the form $-\sum_{\langle ij\rangle} J_{ij}s_is_j$, what are the most likely couplings J_{ij}? We solve this problem using an algorithm called Minimum Probability Flow learning [Sohl-Dickstein et al., 2011], and

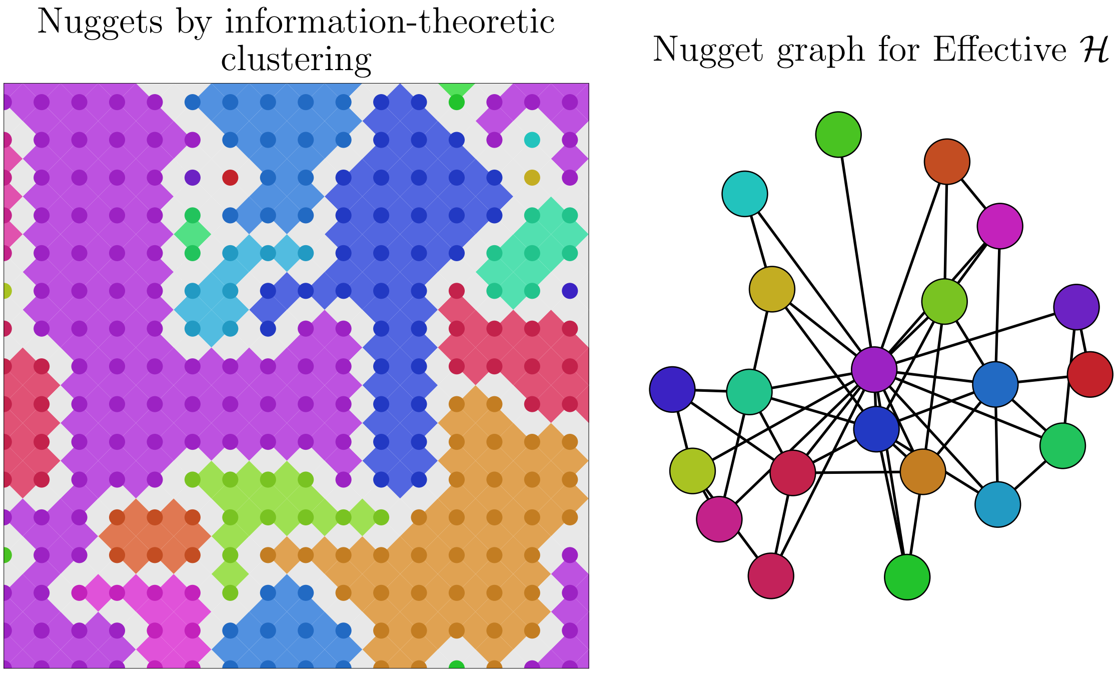

Nuggets by information-theoretic clustering
Nugget graph for Effective $\mathcal{H}$

Figure 4.7: Effective nugget Hamiltonian schematic. (Left) Colors identifying nuggets. (Right) Graph of effective Hamiltonian which treats each nugget as one spin and couples adjacent nuggets.

our own efficient GPU implementation capable of solving a problem with thousands of spins in a few minutes.

The inverse Ising problem is a parameter estimation problem, a model fitting problem, exactly the problem in which sloppiness is originally framed. Sloppiness tells us that only the stiff parameter combinations of a model may be accurately estimated, where the limit on accuracy is determined by the inverse FIM. Therefore since the boundaries of the active droplets are stiff—and by design all boundaries of nuggets are along these stiff directions—the only parameters we have any hope of inferring are the interactions between nuggets. The dual consequence of sloppiness is that only the stiff directions matter to model behavior, so again *sloppiness tells us that the droplets dominate the* *physics*. For this reason we assume the bonds between nuggets which do not share a boundary are zero. In other words, we assume the nugget interactions can be represented by a planar graph.

We find the effective nugget Hamiltonian through the following procedure. Given a spin configuration and nugget assignment, we define the nugget configuration to be the sign of the magnetization of each nugget with respect to the ground state. The ground state is assumed to be the most common observed configuration since we are using exact sampling. In other words, the state of nugget c in sample i is defined to be $S_c^i = \text{sign}(s_c^0 \cdot s_c / |c|)$, where s_c denotes the spins in nugget c. We then find the effective nugget Hamiltonian $\mathcal{H}'$ by solving the inverse Ising problem given the nugget 'time series.'

It would be interesting to test the predictions of droplet theory with $\mathcal{H}'$ in two ways. The energy difference between the ground state energy and the effective energy of activation of a nugget should be similar to the activation free energy of the original nugget. The scaling of the nugget activations with their size should let us estimate θ. The boundaries between nuggets are also domain walls, and so the energy of interaction between nuggets should scale with the boundary length, similar to the Domain-Wall Renormalization Group [McMillan, 1984].

4.6.1 Sampling Nugget Hamiltonians

The effective nugget Hamiltonian $\mathcal{H}'$ can be sampled just like its parent Hamiltonian $\mathcal{H}$, and the resulting correlations of $\mathcal{H}'$ can also yield nuggets and nugget Hamiltonians. Of course $\mathcal{H}'$ is a compressed version of $\mathcal{H}$, so it is natural to ask what properties are preserved. In order to understand this question, we took samples of the same $L = 16$ spin glass as in previous examples at $T = 0.33$ and found $\mathcal{H}'$ as described above. We then sampled both $\mathcal{H}$ and $\mathcal{H}'$ at a cooler temperature of $T = 0.14$ and compared.

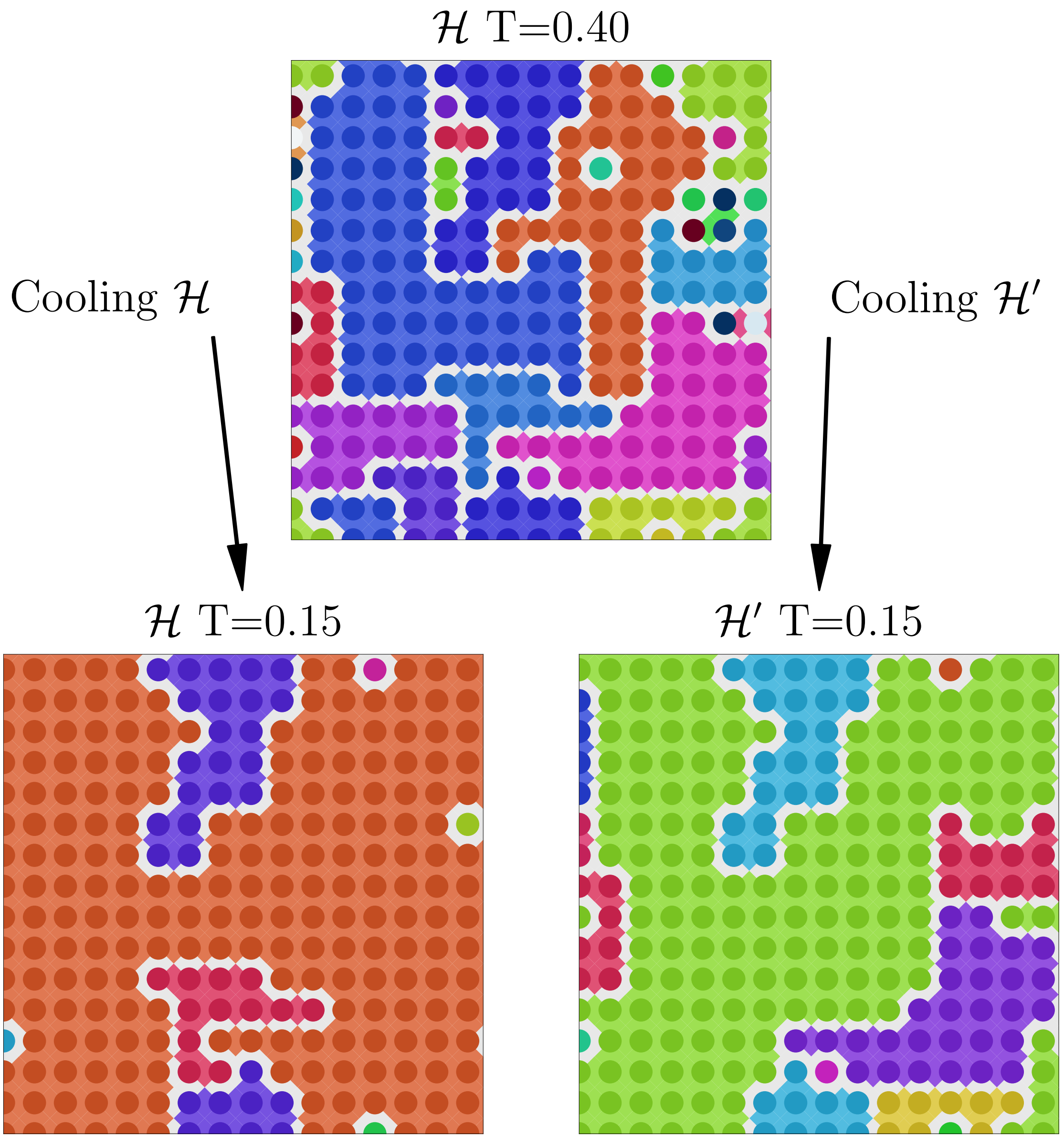


Figure 4.8: Studying the flows of the nugget Hamiltonian. Top: nuggets of $\mathcal{H}$ at T =0.4. Bottom left: nuggets of $\mathcal{H}$ at T =0.14. Bottom right: nuggets of $\mathcal{H}'$ at T =0.14. This shows that the nugget Hamiltonian $\mathcal{H}'$ retains most of the low-temperature correlations.

Figure 4.8 shows that the nugget Hamiltonian retains most of the low-temperature behavior of its parent model. This is because the nuggets of the nugget Hamiltonian preserve the same ‘important’ nuggets as determined by the bond-FIM eigenvector in Fig. 4.3. We can be more quantitative by comparing the explicit spin-spin correlation matrices, where the spins inside a nugget are defined to be perfectly correlated. Figure 4.9 shows the mean absolute difference between the correlation matrix of the parent model and nugget Hamiltonian. The red dashed line indicates the temperature at which the nugget Hamiltonian was formed; the difference is primarily due to throwing away the interior nugget degrees of freedom. As the temperature is lowered, the difference barely increases, the nugget Hamiltonian encodes most of the low temperature physics of its parent model.

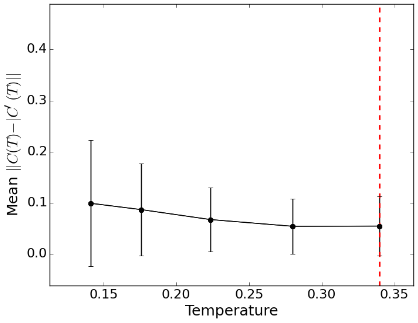


Figure 4.9: Measuring the difference in the correlation functions of the nugget and parent Hamiltonian as a function of temperature. The dotted line indicates the temperature at which $\mathcal{H}'$ was formed from $\mathcal{H}$.

This inheritance of low temperature behavior in high temperature nuggets seems inconsistent with temperature chaos. Temperature chaos is a prediction of droplet theory which states that small changes in temperature will dramatically change the equilibrium behavior, i.e. the active droplets should be changed [Bray and Moore, 1987a, Fisher and Huse, 1988]. This prediction is due to the precarious sensitivity of the active droplets:

their nearly accidental free energy degeneracy allows them to be ‘active,’ but since $F = E - TS$ is small, small changes in T quantity are likely to tip the balance. There are small differences between the nuggets of $\mathcal{H}$ and $\mathcal{H}'$, but the ‘important’ nuggets are retained. Our nuggets are not precisely active droplets, so that nuggets might be quite persistent, whereas the nuggets composing a droplet may change rapidly with temperature. It is also possible that temperature chaos is associated with varying correlations between nuggets.

4.7 Learning Rule for Effective Hamiltonians

Given the effectiveness of our approach, we now seek to simplify its conceptual complexity and efficiency using a method that combines the steps of model reduction. Our current algorithm is a two-step process, which first finds the nuggets and then the effective Hamiltonian – an approach wherein the nugget assignment and effective Hamiltonian influence each other would be prefereable. Additionally, relying on the inverse Ising model is sub-optimal. Knowing the original Hamiltonian should inspire or inform a less general effective Hamiltonian. What follows is a proposed attempt to improve the coarse-graining algorithm used in this paper by searching for an algorithm which pursues these two features.

Given a number of nuggets, we can define a cluster assignment matrix $U_{\alpha i} = \pm\delta_{\alpha p(i)}$, where i labels a spin in the original Hamiltonian, α labels the nuggets in the effective Hamiltonian, and $p(i) = \alpha$ is an assignment or placement of spin i in nugget α. The ± 1 is to accomodate spins which point in or against the direction of their neighbors, and can be set to 1 by an appropriate gauge choice, including defining the ground state to be all $s_i = 1$. We can thus expect that given a spin configuration sampled from the original Hamiltonian s_i, the resulting nugget spin will be $\tilde{s}_\alpha(s_i) = \sigma(\sum_i U_{\alpha i} s_i)$, where $\sigma(x) = \text{sign}(x)$. Now, we can write the probabillity of a given nugget spin configuration $\tilde{p}(s_\alpha(s_i))$ in terms of its parent configuration as

$$\begin{aligned}\tilde{p}(s_i) = \frac{1}{\tilde{\mathcal{Z}}} e^{-\beta \tilde{\mathcal{H}}(s_i)} &\propto \exp\left(-\beta \sum_{\alpha\gamma} \tilde{J}_{\alpha\gamma}\ \tilde{s}_\alpha(s_i) \tilde{s}_\gamma(s_i)\right), \\ &= \exp\left(-\beta \sum_{\alpha\gamma} \tilde{J}_{\alpha\gamma}\ \sigma\left(\sum_i U_{\alpha i} s_i\right) \sigma\left(\sum_i U_{\gamma i} s_i\right)\right), \end{aligned} \tag{4.9}$$

for some inverse temperature β and appropriate nugget-coupling $\tilde{J}_{\alpha\gamma}$.

Assuming we have some nugget assignment operator U (we will discuss choosing this later), what criterion should we use to choose the effective nugget couplings? Since we know the original Hamiltonian we know that the probability of the original spin configuration s_i is

$$p(s_i) = \frac{1}{\mathcal{Z}} e^{-\beta \mathcal{H}} = \exp\left(-\beta \sum_{ij} J_{ij}\ s_j s_j\right), \tag{4.10}$$

for some inverse temperature β, and J_{ij} is the original system couplings which we assume are known or given. Then, it seems sensible that we should choose the coupling $\tilde{J}$ such that $\tilde{p} \sim p$. Framing the problem as an optimization problem then, we seek $\tilde{J}$ such that the Kullback-Leibler (KL) divergence between p and $\tilde{p}$ is minimized, or formally,

$$\tilde{J} = \min\nolimits_{\tilde{J}'} D_{KL}(p||\tilde{p}(\tilde{J}')). \tag{4.11}$$

Writing the KL divergence as a sum over sampled spin configurations s^n, we find

$$\begin{aligned} D_{KL}(p||\tilde{p}) &= -\sum_n p(s^n) \log \frac{\tilde{p}(s^n)}{p(s^n)}, \\ &= -\sum_n p(s^n) \log \frac{\mathcal{Z}}{e^{-\beta \mathcal{H}(s^n)}} \frac{e^{-\beta \tilde{\mathcal{H}}(s^n)}}{\tilde{\mathcal{Z}}}, \\ &= \log \frac{\tilde{\mathcal{Z}}}{\mathcal{Z}} + \beta \left\langle \tilde{\mathcal{H}} - \mathcal{H} \right\rangle_{\mathcal{H}}, \end{aligned} \tag{4.12}$$

where $\langle\cdot\rangle_{\mathcal{H}}$ is the expectation value over states sampled from the original Hamiltonian. As we will show, this quantity can in fact be estimated by taking expectations as sums over states sampled from $\mathcal{H}$, and the ratio of the partition functions can even be estimated by importance sampling.

In order to optimize the KL divergence, we will need to compute its gradient

$$\frac{\partial}{\partial \tilde{J}_{\delta\gamma}} D_{KL}(p||\tilde{p}) = \beta \left(\langle \tilde{s}_\delta \tilde{s}_\gamma \rangle_{\tilde{\mathcal{H}}(\tilde{J})} - \langle \tilde{s}_\delta \tilde{s}_\gamma \rangle_{\mathcal{H}} \right), \tag{4.13}$$

which has the nice physical interpretation that the optimal $\tilde{J}$ (which will set the gradient to zero) will cause the correlation functions of the nugget spins under the original and effective Hamiltonians to match.

How shall we evaluate the KL divergence and its gradient? It seems reasonable to assume that given samples $\{s\}$ of the original Hamiltonian, the set of nugget samples $\{\tilde{s}(s)\}$ computed from these should have frequencies similar to those that an effective Hamiltonian would produce. Therefore we follow the procedure of importance sampling [Neal, 2001], reweighting samples from $\mathcal{H}$ in order to use $\{s\}$ to compute expectations with respect to the Boltzmann distribution of $\tilde{\mathcal{H}}$.

Say we have some distribution $q(x)$ on x, from which samples can be efficiently drawn $x \sim q(x)$, and for which some function $g(x) \propto q(x)$ can be evaluated efficiently. Say that we really wish to compute expectations with respect to a different distribution $p(x)$, where we can efficiently evaluate some function $f(x) \propto p(x)$, but which we cannot directly sample from. Importance sampling is the re-weighting of samples from q, so that as the number of samples increases, expectations values can be computed with respect to p. In particular, defining the weights $w(x) = f(x)/g(x)$, and given samples $x^i \sim q(x)$, as the number of samples goes to infinity, we can compute

$$\langle a(x) \rangle_p = \sum_i w(x^i) a(x^i) / \sum_j w(x^j), \tag{4.14}$$

the expectation value of some function $a(x)$ over the desired distribution $p(x)$. Further, we can estimate the ratio of the partition functions

$$\lim_{N\to\infty} \frac{1}{N}\sum_i w(x^i) = \frac{\mathcal{Z}_f}{\mathcal{Z}_g}, \tag{4.15}$$

where for example $\mathcal{Z}_f = \int \mathrm{d}x f(x)$.

Inspired by importance sampling, and understanding that we can sample from $p(s) \propto \exp(-\beta\mathcal{H}(s))$, but want to compute expecations with respect to $\tilde{p}(s) \propto \exp(-\beta\tilde{\mathcal{H}}(s))$, we can define weights

$$w(s) = \exp\left(-\beta(\tilde{\mathcal{H}}(s) - \mathcal{H}(s))\right) \tag{4.16}$$

such that given samples $s^n \sim p(s^n)$, we can compute expectations of functions of the spin configuration $C(s)$ over $\tilde{p}$ as

$$\langle C(s)\rangle_{\tilde{\mathcal{H}}} = \frac{1}{\sum_n w(s^n)}\sum_n w(s^n)C(s^n) \tag{4.17}$$

Further, we can compute the ratio of the partition functions as

$$\frac{\tilde{\mathcal{Z}}}{\mathcal{Z}} = \langle w(s^n)\rangle_{\mathcal{H}} = \langle e^{-\beta(\tilde{\mathcal{H}}-\mathcal{H})}\rangle_{\mathcal{H}}. \tag{4.18}$$

Putting all this together, we can express the KL divergence in eqn. 4.12 in terms of some set of N samples $\{s^n\} \sim p(s^n)$ drawn from the original Hamiltonian as

$$D_{KL}(p||\tilde{p}(\tilde{J})) = \frac{\beta}{N}\sum_n \Delta\mathcal{H}(s^n) + \log\left(\frac{1}{N}\sum_n \exp(-\beta\Delta\mathcal{H}(s^n))\right), \tag{4.19}$$

where $\Delta\mathcal{H}(s) = \tilde{\mathcal{H}}(s) - \mathcal{H}(s)$. Finally, the gradient can be expressed as

$$\frac{\partial}{\partial\tilde{J}_{\delta\gamma}} D_{KL}(p||\tilde{p}) = \beta\sum_n \tilde{s}_\delta(s^n)\tilde{s}_\gamma(s^n)\left(\frac{e^{-\beta\Delta\mathcal{H}(s^n)}}{\sum_m e^{-\beta\Delta\mathcal{H}(s^m)}} - \frac{1}{N}\right). \tag{4.20}$$

Note that if we we happen to choose the number of nuggets to match the number of spins in $\mathcal{H}$, the gradient of the KL divergence will be zero if the weights $w(s) = 1/N$, which is true if and only if the Hamiltonians match.

Therefore with eqn. 4.19 and eqn. 4.20 we have a well-defined learning rule for obtaining an effective Hamiltonian $\tilde{\mathcal{H}}$ given a starting Hamiltonian $\mathcal{H}$. The problem remaining is to set the nugget assignment operator $U_{\alpha i}$.

4.7.1 Connection to the Real Space Renormalization Group

Before discussing how we might set the nugget spin operator U, let us consider the implications of this learning algorithm on a simpler case: the $N = L \times L$ 2D Ising model. In this case, we know a reasonable nugget assignment is that of block-spins in Kadanoff's Real-Space Renormalization scheme [Efrati et al., 2014]. In this case the spatial symmetry suggests that we choose a number of nuggets $N_c = N/b^2$ for some integer $b > 1$, assigning squares of adjacent spins in $b \times b$ blocks to a single nugget or block spin. In the standard block spin renormalization one defines a new hamiltonian $\tilde{\mathcal{H}}(\tilde{s})$ on the block spins and requires that the partition function of the new effective Hamiltonian matches the original partition function:

$$\tilde{\mathcal{Z}} = \sum_{\{\tilde{s}\}} e^{-\beta \tilde{\mathcal{H}}(\tilde{s})} = \sum_{\{s\}} e^{-\beta \mathcal{H}(s)} = \mathcal{Z}. \tag{4.21}$$

Does our learning objective seek a relationship between the partition functions at all reminiscent of RG? The KL divergence $D_{KL}(p||\tilde{p}) \geq 0$, and after some manipulation we can turn eqn. 4.12 into

$$\mathcal{Z} \geq \tilde{\mathcal{Z}}(\tilde{J}) e^{-\beta \langle \tilde{\mathcal{H}}(\tilde{J}) - \mathcal{H} \rangle_{\mathcal{H}}}, \tag{4.22}$$

which is the Gibbs-Feynman-Bogoliubov (GFB) inequality. If we chose $\tilde{\mathcal{H}}$ to be the mean field Hamilltonian coupling each spin to a field, optimizing this inequality would yield the mean field solution to the Ising model [Nishimori, 2001]. As we optimize the effective Hamiltonian, this tells us that $\tilde{\mathcal{Z}}$ is bounded from above by $\mathcal{Z}$, assuming $\langle \Delta\mathcal{H} \rangle$ is small.

Writing the GFB inequality another way we find

$$e^{-\beta \mathcal{F}} \leq e^{-\beta \tilde{\mathcal{F}}(\tilde{J}) + \langle \tilde{\mathcal{H}}(\tilde{J}) - \mathcal{H} \rangle_{\mathcal{H}}}, \tag{4.23}$$

which is striking in its resemblance to the RG criterion relating the *singular* parts of the free energy

$$e^{-\beta \mathcal{F}} = e^{-\beta \tilde{\mathcal{F}}(\tilde{J}) + g(\tilde{J})}, \tag{4.24}$$

for some analytic function g of the effective couplings $\tilde{J}$.

4.7.2 Choosing the Nugget Spin Operator

Emboldened by a plausible relationship to block-spin renormalization, but sobered by the lack of spatial symmetry in spin glasses, we seek a method for assigning spins to blocks or nuggets. Since the KL divergence is non-negative, a brute-force method would be to perform Monte Carlo on the nugget spin operator $U_{\alpha i}$, treating it as a kind of multi-spin. We are guaranteed by this fact, explored in the previous section, that the lower the KL divergence, the better the approximation.

We could also interpret the spin operator as a probability of assigning a spin to a nugget $U_{\alpha i} = p(\alpha|i)$, softening the spin assignment. Then we could simply optimize eqn. 4.19 over $\tilde{J}$ and $U_{\alpha i}$. Secure in the fact that the KL divergence is bounded below by zero, a superior optimization is always better. We can even use stochastic gradient descent to efficiently optimize this functional, as it is a sum of terms. Thorough numerical study of this newly defined learning problem will be reserved for future work.

4.8 Conclusion

Sloppiness is a geometric interpretation of high-dimensional model fitting, which leads to a coherent description of the spin glass phase consistent with droplet theory in 2D. It inspires a new method of analyzing spin glasses capable of probing many length scales at finite temperature, and a method of coarse-graining disordered systems. Many systems from neuroscience and biology have been modeled as spin glasses, and our method could provide coarse-grained and more easily interpretable description of these results without sacrificing the ability to make good predictions. Future work could apply these methods to protein folding, binary voting systems like the supreme court, DNA mutation, or be devoted to finding a method for unifying the clustering and effective Hamiltonian procedure. Finally, this method could be applied to 3D spin glasses, where sloppiness and information geometry might inspire a new perspective in an old debate.

Acknowledgments

The authors would like to thank Ben Machta, Mark Transtrum, Alex Alemi, and Lorien Hayden for useful conversations. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

4.9 Appendix

4.9.1 Bond-FIM for two interacting droplets

We assume two adjacent droplets (still two-state systems) interact only through their shared bonds. In the ground-state gauge their interaction energy is the sum of the bonds on their boundary, which we will denote $J = \sum_\mu \theta^\mu$ where μ are bond indices on the boundaries of two droplets c^A and c^B. Each of these droplets has an activation energy Δ_A and Δ_B respectively. With two droplets we have four states to consider: We have assumed the sign of J is such that positive J makes activating one droplet easier if the other is active. The probability of any one of these states is for instance $P(\uparrow\uparrow) = \exp -\beta E(\uparrow\uparrow)/Z$ where $Z = \sum_{AB} \exp -\beta E(AB)$ is the sum over all four states of A and B. The calculation is similar to the above independent case, except now the FIM is not block diagonal. We can still compute the total bond-FIM from sums of independent pairs of coupled droplets:

A/B state	Energy $-(E_0 - \sum_\mu \theta_\mu)$
↓↓	0
↑↓	$\Delta_A - 2\sum_\mu \theta_\mu \delta_\mu^A$
↓↑	$\Delta_B - 2\sum_\mu \theta_\mu \delta_\mu^B$
↑↑	$\Delta_A + \Delta_B - J - 2\sum_\mu \theta_\mu (\delta_\mu^A - \delta_\mu^B)$

$$\begin{aligned} g_{\mu\nu}^{AB} = \frac{4\beta^2}{N} \Big[& \delta_\mu^A \delta_\nu^A e^{\beta\delta_A} \left(1 + e^{\beta\delta_B}\right) \left(e^{\beta J} + e^{\beta\delta_B}\right) \\ & + (\delta_\mu^A \delta_\nu^B + \delta_\mu^B \delta_\nu^A) e^{\beta(\Delta_A + \Delta_B)} \left(e^{\beta J} - 1\right) \\ & + \delta_\mu^B \delta_\nu^B e^{\beta\delta_A} \left(1 + e^{\beta\delta_B}\right) \left(e^{\beta J} + e^{\beta\delta_B}\right) \Big] \end{aligned} \tag{4.25}$$

Where $N = e^{\beta J} + e^{\beta\Delta_A} + e^{\beta\Delta_A} + e^{\beta(\Delta_A + \Delta_B)}$. This matrix still has large blocks proportional to a matrix of ones, except with off-diagonal blocks corresponding to the interactions. It can be diagonalized by noting that it has only two linearly independent columns, following the procedure outlined in appendix II. We won't present the results of the exact eigenvalues and vectors of $g_{\mu\nu}^{AB}$ as they are too complicated to be useful. Instead we will expand in the limit of small coupling J. To first order in J, the field-FIM can be written

$$g^{AB} = \begin{pmatrix} \beta^2 \text{sech}^2 \frac{\beta\Delta_A}{2} & 0 \\ & \\ 0 & \beta^2 \text{sech}^2 \frac{\beta\Delta_B}{2} \end{pmatrix} + J\beta^3 \begin{pmatrix} \frac{\text{sech}^2 \frac{\beta\Delta_A}{2} \tanh \frac{\beta\Delta_A}{2}}{1+\exp\beta\Delta_B} & \text{sech}^2 \frac{\beta\Delta_A}{2} \text{sech}^2 \frac{\beta\Delta_B}{2} \\ & \\ \text{sech}^2 \frac{\beta\Delta_A}{2} \text{sech}^2 \frac{\beta\Delta_B}{2} & \frac{\text{sech}^2 \frac{\beta\Delta_B}{2} \tanh \frac{\beta\Delta_B}{2}}{1+\exp\beta\Delta_A} \end{pmatrix} \tag{4.26}$$

where it should be understood that Eq. 4.26 contains block matrices with the number of constant elements equal to the squared perimeters of droplets A and B. The block diagonal matrix represents the bond-FIM of the uncoupled droplets. The first-order correction in the eigenvalues of Eq. 4.26 are

$$\begin{aligned} \lambda^A &= p_A \beta^2 \text{sech}^2 \frac{\beta\Delta_A}{2} \left(1 + \beta J \frac{\text{sech}^2 \frac{\beta\Delta_A}{2} \tanh \frac{\beta\Delta_A}{2}}{1 + \exp\beta\Delta_B}\right) \\ \lambda^B &= p_B \beta^2 \text{sech}^2 \frac{\beta\Delta_B}{2} \left(1 + \beta J \frac{\text{sech}^2 \frac{\beta\Delta_B}{2} \tanh \frac{\beta\Delta_B}{2}}{1 + \exp\beta\Delta_A}\right) \end{aligned} \tag{4.27}$$

Since g^{AB} is singular, obtaining the eigenvectors through perturbation theory is not straight-forward, however they can be obtained by solving Eq. 4.29 for the exact expression and then expanding the two solutions in m. They are

$$\begin{aligned} m_A &= \frac{2}{\beta J}\left((1 + \cosh\beta\Delta_B) - \frac{N_A}{N_B}(1 + \cosh\beta\Delta_A)\right) \\ m_B &= \frac{\beta J}{2}\left((1 + \cosh\beta\Delta_A) - \frac{N_A}{N_B}(1 + \cosh\beta\Delta_B)\right)^{-1} \end{aligned} \tag{4.28}$$

The normalized eigenvectors are $\vec{v}_A = \frac{1}{\sqrt{N_A m_A^2 + N_B}}(m_A \vec{1}_{N_A}, \vec{1}_{N_B})^T$, and similarly for the eigenvector corresponding to the eigenvalue dominated by B. The numbers m tell us how much the volumes of the droplets get smeared across eigenvectors, explaining the shadows we observe in the numerical eigenvectors.

4.9.2 Eigenvalues of block-constant matrix

We need to find the eigenvalues and eigenvectors of a block matrix with constant values in each block, for example below is such a symmetric block-constant matrix. We propose an ansatz eigenvector of the form $\vec{v} = (m\vec{1}_{N_A}, \vec{1}_{N_B})$ where m is a constant to be determined, $\vec{1}_{N_A}$ is a vector of ones of length N_A and likewise for N_B.

$$\begin{pmatrix} c_{AA}\mathbf{1}_{A\times A} & c_{AB}\mathbf{1}_{A\times B} \\ \\ c_{AB}\mathbf{1}_{\boldsymbol{B\times A}} & c_{BB}\mathbf{1}_{B\times B} \end{pmatrix} \begin{pmatrix} m\vec{1}_A \\ \\ \vec{1}_B \end{pmatrix} = \begin{pmatrix} (c_{AA}Am + c_{AB}B)\vec{1}_A \\ \\ (c_{AB}Am + c_{BB}B)\vec{1}_N \end{pmatrix}$$

Where $\mathbf{1}_{n\times n}$ is an $n \times n$ matrix of all ones. If $\vec{v}$ is an eigenvector with eigenvalue λ the following system of equations must be satisfied:

$$\begin{aligned} \lambda m &= c_{AA}Am + c_{AB}B \\ \lambda &= c_{AB}Am + c_{BB}B \end{aligned} \tag{4.29}$$

We can solve for λ and m, and the two solutions of m give orthogonal eigenvectors. Since this matrix is rank-2, we have found the only non-zero eigenvalues and eigenvectors.

Chapter 5

Normal Form of the Two-Dimensional Ising Model Renormalization Group Flows

5.1 Introduction

We explore the scaling behavior of the two-dimensional Ising model with normal form theory [Raju et al., 2019], applied to its renormalization group flows under coarse-graining. The Ising model has a logarithmic singularity in the specific heat, which is well known to arise from a nonlinear term in the renormalization group flow [Wegner, 1972]. Bifurcation theory studies how coordinate transformations can be used to cast nonlinear differential equations near fixed points into a simplest normal form. For the Ising model, this change of

The work constituting this chapter was done in collaboration with Archishman Raju and James P. Sethna variables leads to a flow with positive linear eigenvalues correspond to the universal critical exponents for temperature and field, negative eigenvalues for various irrelevant parameter combinations, and a variety of allowed universal nonlinear resonant terms including the known one leading to the specific heat singularity. We shall use the Ising model to investigate several outstanding questions. Can we deduce the coordinate transformation which takes the exact Onsager solution into the normal form? Are the corrections to scaling in the Onsager solution due to irrelevant variables, or to analytic corrections to scaling introduced by the coordinate transformation? (Or can these be distinguished?) How do the renormalization-group flows behave under Legendre transformations – in particular, changing from temperature and field as control variables to the microcanonical energy and entropy variables? Although the exact results and extensive previous research on the 2D Ising model provide a firm ground for these explorations, the 2D Ising model is special in many ways that prevent us from developing definitive answers to many of these questions. We provide several conjectures and predictions from our analysis, however, which should have strong implications for corrections to scaling and Legendre transformations for more general critical points.

The celebrated Renormalization Group (RG) casts studies of critical phenomena into dynamical systems, flows of the free energy and control parameters under coarse-graining, and predicts power laws, scaling relations, and universal scaling functions. The RG allows the classification of many systems into universality classes which share critical exponents, amplitudes, and scaling functions. The two-dimensional Ising model is an important analytically tractable model which to this day has unresolved questions regarding its corrections to scaling as predicted by the RG. For example, while the free energy of the 2D lattice Ising model is known to be $f(t) = a(t)\log t + b(t)$, and the magnetization $m(t) = t^{1/8}c(t)$ for analytic functions a, b, and c of the reduced temperature

$t = (T - T_c)/T_c$, it is still unclear, whether the susceptibility $\chi(t)$ has one or more powers of $\log t$ [Boukraa et al., 2008].

Open questions remain regarding the role of irrelevant variables in the 2D Ising model. The most dangerous irrelevant eigenvalues as predicted by Conformal Field Theory (CFT) [Caselle et al., 2002] are associated with breaking rotational invariance: -2 and -4. There is also evidence that an irrelevant variable with eigenvalue $-3/4$ exists [Nienhuis, 1982, Barma and Fisher, 1985], which could explain features of the zero-field magnetic susceptibility [Orrick et al., 2001]. Calabrese et al. [Calabrese et al., 2000] suggest that relaxing the unitarity or 'reflection positivity' constraint allows CFT to predict a $\lambda_u = -3/4$ irrelevant variable. Caselle et al. [Caselle et al., 2002] support a conjecture—originally proposed by Aharony and Fisher [Aharony and Fisher, 1983]—that the leading irrelevant variables do not contribute to corrections to scaling. Here we discuss how this may be so.

In recent work we have shown that non-linear RG flows can be systematically classified by the application of normal form theory (NFT) [Raju et al., 2019]. Normal form theory uses sequences of polynomial coordinate transformations to simplify systems of ordinary differential equations, reducing them to a *normal form.* Using normal forms, flow equations of critical systems can be classified into *universality families* characterized by a small number of universal terms. For example, linearizable systems all fall under the *hyperbolic universality family* (e.g. $3D$ Ising). In this way, one can predict the complete form of the singularity near the fixed point by finding which nonlinear terms in the flows are universal, leading to a systematic and general method for studying corrections to scaling.

In sec. 5.2 we study the scaling behavior and corrections to scaling in the context of the hyperbolic universality family (e.g. 3D Ising), where the normal form linearizes the

RG flow. In sec. 5.3 we introduce the normal form of the 2D Ising model ignoring the irrelevant variables, exhibiting the irremovable nonlinear term due to a resonance that leads to the logarithmic singularity in the specific heat. We derive the scaling ansatz commonly used, and argue that there should not be a scaling function multiplying the log term in a scaling ansatz (agreeing with ref. [Wegner, 1972]). In sec. 5.3.1 we show that the coordinate transformation leading to the normal form leads to a predicted Ising free energy of the form $f(t) = a(t)\log t + b(t)$, where the analytic functions $a(t)$ and $b(t)$ reflect analytic corrections to scaling. We use Onsager's exact solution for the free energy on the nearest-neighbor square lattice 2D Ising model to calculate the coordinate transformation, showing that corrections to scaling from irrelevant variables are not needed to explain the exact results, in agreement with previous researchers [Caselle et al., 2002, Calabrese et al., 2000, Aharony and Fisher, 1983]. In sec. 5.4 we consider the question of irrelevant variables. The known irrelevant eigenvalues [Nienhuis, 1982, Barma and Fisher, 1985, Caselle et al., 2002] are integers and simple fractions. The corrections to scaling due to these irrelevant variables would naively take the same form as the analytic corrections to scaling discussed above, posing the question of whether these irrelevant variables also represent invariance under coordinate tranformations (perhaps 'gauge irrelevant' variables [Raju and Sethna, 2018] representing 'redundant' operators [Wegner, 1974] associated with changing the definition of the local order parameter). The universal resonant terms associated with these irrelevant variables, if they are non-zero, would lead to logarithmic contributions to the free energy which are inconsistent with the exact solution. We shall conclude that either all the amplitudes of the irrelevant variables happen to be zero for the exactly solvable Ising models (as they are for the non-conformal -3/4 eigenvalue), or the universal values of their resonant terms must all be zero (as expected for gauge-irrelevant eigenvalues, which could explain why Caselle et al. [Caselle et al., 2002] see no such corrections). In sec. 5.7 we investigate the normal form including finite size, and see modest improvement in the scaling collapse of the specific heat using the analytic corrections found in sec. 5.3.1. In sec. 5.8 we discover the surprising fact that the flow equations in the microcanonical ensemble (entropy as a function of energy) are generically non-analytic. We investigate the relationships between observables and control variables using the Legendre transformation.

5.2 Linearized RG Flows: the Hyperbolic Family

The classic (pure power law) predictions of the RG are obtained by linearizing the flow equations [Wilson, 1971]. (Simply linearizing the flows fails to capture logarithmic singularity of the specific heat of the $2D$ Ising model, which we remedy in sec. 5.3.) In general, the linearized RG predicts [Cardy, 1996] that the singular part of the free energy f of any thermodynamic system in the vicinity of a critical point should obey the following form

$$\mathcal{F}(t,h,u) = |\tilde{t}|^{D/\lambda_t}\ \mathcal{F}_{\pm}\left(\tilde{h}/|\tilde{t}|^{\lambda_h/\lambda_t}, \tilde{u}/|\tilde{t}|^{\lambda_u/\lambda_t}\right), \tag{5.1}$$

where t is the reduced temperature $t = (T - T_c)/T_c$ in terms of the model control temperature T and the critical temperature T_c, h is the ordering field, and u represents one of many irrelevant control variables, like a second nearest neighor interaction.

The power-law contribution $|\tilde{t}|^{D/\lambda_t}$ is the most famous prediction of this linearized RG, where the dimension-dependent critical exponent $2-\alpha$ is the most salient number which can define a universality class. For example, the 2D ($D = 2$) Ising university class features $\lambda_t = 1$. The other critical exponents defining the universality class are $\lambda_h > 0$ and $\lambda_u < 0$. The function $\mathcal{F}_{\pm}(x,y)$ is another universal prediction of the RG, where ± denotes a different function above or below T_c or $t = 0$.

The careful reader will note that the right-hand side of eqn. 5.5 is actually a function of $\tilde{t}$, $\tilde{h}$ and $\tilde{u}$. These are called non-linear scaling variables [Wegner, 1972], for reasons which are made clear below, and are in general analytic functions of the control variables t, u, and h. They arise due to non-universal details of each physical system, like the box containing the experiment, an imperfect thermometer, or some coupling between parameters which do not change essential features of the critical point.

While the power law is the most famous and beautiful prediction of the RG, there are many real and significant modifications to the pure power law $|\tilde{t}|^{D/\lambda_t}$, even for hyperbolic fixed points (that can be linearized). These corrections to scaling generally come from two sources, and yield two types of corrections named for types of terms they cause: analytic corrections and singular corrections. Imagining that we are close to the critical point (and that $h = 0$ for clarity), and writing for example $\tilde{t} = t + at^2 + \ldots$, eqn. 5.5 can be expanded around $t = 0$ and $\tilde{u} = 0$, yielding

$$\begin{aligned}\mathcal{F}(t, 0, u) = |t|^{D/\lambda_t}\Big(&A_1 + A_2 t + A_3 t^2 + \ldots + \\ &B_1 \tilde{u} t^{\frac{\lambda_u}{\lambda_t}} + B_2 \tilde{u} t^{1+\frac{\lambda_u}{\lambda_t}} + \ldots\Big),\end{aligned} \tag{5.2}$$

where the constants A_i and B_i are appropriate constants depending on λ_t, λ_u, $\mathcal{F}'(0)$. There are integral powers of t due to expanding $\tilde{t}$, the analytic corrections, and terms like t^{λ_u/λ_t}, which are non-analytic or singular, due the irrelevant variable u. Note that since conformal field theory (CFT) predicts for the 2D Ising model that $\lambda_u = -2n, -(2n + 1/8)$ for appropriate integers n, the singular corrections due to the negative integer irrelevant variables will be indistinguishable from analytic corrections.

5.3 2D Ising RG Flow Normal Form: Resonance

The most general form of the flow equations of the $2D$ Ising model can be written

$$\begin{aligned} df/d\ell =& \quad D\ f(1+g_f(t,h,u,\ldots)) \\ dt/d\ell =& \quad \lambda_t\ t(1+g_t(t,h,u,\ldots)) \\ dh/d\ell =& \quad \lambda_h\ h(1+g_h(t,h,u,\ldots)) \\ du/d\ell =& -\lambda_u\ u(1+g_u(t,h,u,\ldots)), \\ &\vdots \end{aligned} \tag{5.3}$$

where f is the free energy, t is the reduced temperature, u represents the many irrelevant variables (with $\lambda_u < 0$), the dimension is $D = 2$, and $\lambda_h = 15/8$. The functions g_f, g_t and g_u are analytic function which vanish at the critical point. Normal form theory of dynamical systems tells us we can find a coordinate change $f(\tilde{f}, \tilde{t}, \tilde{u}, \ldots)$, $t(\tilde{t}, \tilde{u}, \ldots)$, and so on which can lead to a set of differential equations which are *simplest* in the sense that only a small number of terms (maybe even zero) in the power series of the g-functions remain. According to NFT the terms that cannot be removed satisfy a resonance condition, an integer linear relationship between eigenvalues.

Ignoring resonances due to irrelevant variables for now, the normal form of the $2D$ Ising model is

$$\begin{aligned} &\frac{d\tilde{f}}{d\ell} = D\tilde{f} - \tilde{t}^2, \qquad \frac{d\tilde{t}}{d\ell} = \tilde{t}, \qquad \frac{d\tilde{h}}{d\ell} = \frac{15}{8}\tilde{h}, \\ &\frac{d\tilde{u}}{d\ell} = \lambda_u \tilde{u}, \end{aligned} \tag{5.4}$$

where we have rescaled $\tilde{t}$ to set the coefficient of $\tilde{t}^2$ to unity[1]. The $\tilde{t}^2$ term in eqn. 5.4

[1] Rescaling cannot change the sign, which is negative in the exact solution

cannot be removed by smooth coordinate changes because of a resonance [2] between $\tilde{f}$ and $\tilde{t}$, which is well-known to cause logarithmic singularities [Wegner, 1972]. Coarse-graining til $\tilde{t} = 1$, the prediction for the free energy (the initial conditions of the flows) is found to be

$$\tilde{f}_0(\tilde{t}_0) = \tilde{t}_0^2 \mathcal{F}\left(\tilde{h}_0 \tilde{t}_0^{1/y_h}, \tilde{u}_0 \tilde{t}_0^{1/y_u}\right) - \tilde{t}_0^2 \log \tilde{t}_0. \tag{5.5}$$

Some have speculated on the existence of another scaling function in eqn. 5.5 multiplying the logarithm [Caselle et al., 2002], but our normal form calculation above argues that none is needed. As we note elsewhere, since many integer eigenvalues λ_u exist for the 2D Ising model, it is possible that singular and analytic corrections will be indistinguishable. In sec. 5.4 we will show that resonances with irrelevant variables will modify the flows of eqn. 5.4 by adding a number of terms due to resonances between f, t, h, and the irrelevant variables with integer or simple fraction eigenvalues like $\lambda_u = 2, 4, 3/4, \ldots$.

In sec. 5.5 we will show that these resonant terms will modify the scaling ansatz of eqn. 5.5 by allowing terms of the form $(\log t)^n$ and $\log \log t$. The exact solution of the 2D Ising model is of the form $f(t) = a(t) \log t + b(t)$ for analytic functions a and b, so we will conclude that these resonances cannot contribute to the corrections to scaling in the exactly–solved lattice model. There are in fact two distinct ways that irrelevant variables can be neglected: a zero value $u_0 = 0$ of the irrelevant variable for the system being studied, or the universal coefficients of its normal-form nonlinear terms are identically zero. Caselle et al. [Caselle et al., 2002] support the former assertion that $u_0 = 0$ at the critical point in the case of the leading irrelevant operator in CFT; Barma and Fisher [Barma and Fisher, 1985] conjecture that it is zero for the non-conformal operator with eigenvalue $-3/4$.

[2] In this case the resonance condition is the integer relation $2 \cdot \lambda_t = \lambda_f$, with $\lambda_t = 1$ and $\lambda_f = 2$, for which a change of variable calculation shows that the $\tilde{t}^2$ term cannot be removed from the $\tilde{f}$ flow.

5.3.1 Unique coordinate transformation

The exact free energy of the 2-D Ising model is of the form $f(t) = a(t)\log t + b(t)$ where $a(t)$ and $b(t)$ are analytic functions of t. This can be seen by carefully studying Osager's exact solution [McCoy and Wu, 2014], or by simply adding analytic corrections to the normal form of eqn. 5.4. The normal form solution of the free energy in terms of a coordinate transforms $\tilde{t}(t)$ and $\tilde{f}(f,t)$ is

$$\tilde{f}(\tilde{t}) = -\tilde{t}^{2}\log\tilde{t}, \tag{5.6}$$

where we have chosen to coarse-grain to a point which removes the lone $\tilde{t}^{2}$ term. We now show that one can explain the exact solution purely in terms of analytic corrections to scaling (presuming that all irrelevant variables have unrenormalized initial values $u_0 = 0$). The most general transformation is $t = \tilde{t}(1+\tau(\tilde{t}))$, for arbitrary analytic function τ. Transforming the exact free energy we find

$$f(\tilde{t}) = a(\tilde{t})\log\tilde{t} + a(\tilde{t})\log(1+\tau(\tilde{t})) + b(\tilde{t}), \tag{5.7}$$

where $a(\tilde{t}) = a(t(\tilde{t}))$, and similarly for f, and b. The last two terms of $f(\tilde{t})$ are analytic; we can see that our coordinate transformation for f must be

$$\tilde{f}(\tilde{t}) = f(\tilde{t}) - a(\tilde{t})\log(1+\tau(\tilde{t})) - b(\tilde{t}). \tag{5.8}$$

In order for the transformation to match the normal form solution we then require that $\tilde{t}^{2} = -a(\tilde{t}(1+\tau(\tilde{t}))$. Thus if we assume that the irrelevant variables do not contribute to the free energy, this uniquely specifies the change of variables to the normal-form $\tau(\tilde{t})$.

What if we first transform f by an analytic function in t? The most general transformation we are allowed is $\tilde{f}^{(1)} = f + c(t)$. Then applying $\tilde{t}$ as above, $\tilde{f}^{(2)} = \tilde{f}^{(1)}(t(\tilde{t})) = a(\tilde{t})\log\tilde{t} + a(\tilde{t})\log(1+\tau(\tilde{t})) + b(\tilde{t}) + c(\tilde{t})$. Then to obtain the normal form, $\tilde{f}^{(3)} = \tilde{f}^{(2)} - a(\tilde{t})\log(1+\tau(\tilde{t})) - b(\tilde{t}) - c(\tilde{t})$, which if we simplify, cancels our original transformation of $c(t)$, yielding the same form for $\tilde{f}$ as above.

5.3.2 Nonlinear Scaling Fields

Using the results of the previous section and Onsager's exact solution to the Ising model, we can find the nonlinear scaling field $\tilde{t}(t)$ which will account for analytic corrections in the vicinity of $t = 0$. First, we find the singular part of the free energy from Onsager's solution, using an analytic formula for the energy or df/dt [McCoy and Wu, 2014]:

$$\frac{df}{dt} = -\frac{T_c}{T(t)} \coth\left(\frac{2}{T(t)}\right)\left(1 + \frac{2}{\pi} q(t) K(p(t))\right), \tag{5.9}$$

where $K(m) = \int_0^{\pi/2} d\theta/\sqrt{1 - m\sin^2\theta}$ is the elliptic integral of the first kind,

$$q(t) = 2\tanh\left(\frac{2}{T(t)}\right)^2 - 1, \tag{5.10}$$

and

$$p(t) = 2\frac{\sinh(2/T(t))}{\cosh(2/T(t))^2}, \tag{5.11}$$

and $T(t) = T_c(t+1)$. K has a branch cut and a logarithmic singularity at $K(1)$. Using an asymptotic expansion for $K(m)$ around $m = 1$, we can find an expansion for the energy as

$$\frac{df}{dt} = g(t)\log(t) + h(t), \tag{5.12}$$

in the form of Taylor series $g(t)$ and $h(t)$. We can then integrate eqn. 5.12 over t, obtaining

$$f(t) = \int \frac{df}{dt} = a(b)\log t + b(t), \tag{5.13}$$

the form of the free energy predicted in sec. 5.3.1 by normal form theory.

Using the expansion of the free energy in eqn. 5.13, and following the prescription of sec. 5.3.1, we can find the nonlinear scaling field $\tilde{t}$ to arbitrary order. The first 6 terms are

$$\begin{aligned}\tilde{t}(t) = \; & t + t^2\left(\frac{\beta}{\sqrt{2}} - 1\right) + t^3\left(\frac{11\beta^2}{12} - \sqrt{2}\beta + 1\right) + \\ & t^4\left(\frac{25\beta^3}{12\sqrt{2}} - \frac{11\beta^2}{4} + \frac{3\beta}{\sqrt{2}} - 1\right) + \\ & t^5\left(\frac{1289\beta^4}{480} - \frac{25\beta^3}{3\sqrt{2}} + \frac{11\beta^2}{2} - 2\sqrt{2}\beta + 1\right) + \\ & t^6\left(\frac{10399\beta^5}{1440\sqrt{2}} - \frac{1289\beta^4}{96} + \frac{125\beta^3}{6\sqrt{2}} - \frac{55\beta^2}{6} + \frac{5\beta}{\sqrt{2}} - 1\right) \\ & + \mathcal{O}(t^7), \end{aligned} \tag{5.14}$$

where $\beta = \log(1+\sqrt{2})/2$ is the inverse critical temperature. Equation 5.14 agrees with Caselle et al. [Caselle et al., 2002] to the order they report.

Note that technically the coefficients in eqn. 5.14 are functions of the value of all irrelevant variables, and that singular corrections from irrelevant variables with integer eigenvalues will also in general contribute to the coefficients. See Appendix sec. 5.11.1 for details. Further note that $\tilde{t}(t)$ can be combined with the normal form of eqn. 5.4 to find the flow equations of the 2D square-lattice Ising model in the original variables

5.4 Normal Form for Irrelevant Variables

Equation 5.4 is the normal form of the 2D Ising model, ignoring irrelevant variables. The t^2 term in the free energy flow, responsible for the famous logarithmic singularity, arises due to the resonance $0 \cdot D + 2 \cdot \lambda_t = D$: The right-hand side indicates that the equation with eigenvalue D (the free energy) will have a term t^2 which cannot be removed by smooth coordinate transformations. Since all so far proposed irrelevant variables of the 2D Ising model are negative integers or simple fractions like $\lambda_u = -3/4, -2, -4, \dots$ [Caselle et al., 2002, Calabrese et al., 2000], and since the relevant eigenvalues $\lambda_f = D$, $\lambda_t = 1$, and $\lambda_h = 15/8$ are also simple fractions or integers, we must expect many more resonances in the flows of eqn. 5.3, and thus more logarithmic corrections.

Consider a model resonance added to the flows of some variable y, caused by a resonance between a relevant variable x, and an irrelevant variable u:

$$\lambda_y + n(p\lambda_x - q\lambda_u) = \lambda_y, \tag{5.15}$$

If $p\lambda_y + q\lambda_u = 0$ for some non-negative integers p and q then the above resonance condition is satisfied for any n. For example, if $\lambda_u = -3/4$, then there is a resonance between u and t of the form $4\lambda_t - 3\lambda_u = 0$. In this case, NFT tells us that the normal form flows of any variable y will have terms $y(t^4u^3)^n$ which cannot be removed by analytic coordinate transformations for all $n > 0$, or in general some analytic function $g_y(t^4u^3)$ is essential to the singularity as

$$d\tilde{y}/d\ell = \lambda_y y(1 + g_y(\tilde{t}^4\tilde{u}^3)). \tag{5.16}$$

Our previous work [Raju et al., 2019] found that hyper-normal form theory or simplest NFT [Yu and Leung, 2007]— using a sequence of low-order polynomial coordinate transformations— can remove all but a couple of these resonant terms in the analytic function $g_y(t^4u^3)$, as we now demonstrate.

Let $w = x^p u^q$, with some relevant x and irrelevant y, the eigenvalues of which satisfy $p\lambda_y + q\lambda_u = 0$. We assume first that all analytic terms which are not resonances between

x and u have been removed, writing our flows as

$$\begin{aligned} dx/d\ell =&\lambda_x\ x(1+g_x(x^p u^q)),\\ dy/d\ell =&\lambda_y\ y(1+g_y(x^p u^q)),\\ du/d\ell =&\lambda_u\ u(1+g_u(x^p u^q)), \end{aligned} \tag{5.17}$$

where the infinite number of resonances are summarized in the analytic g-functions.

First consider the y-flows as they are uncoupled from x and u. We can start by assuming the form $dy/d\ell = \lambda_y y(1 + cw + dw^2)$. Proposing a coordinate change $y = y_1(1 + Aw)$, we find that no solution exists for A which can remove the wy resonance. Mercifully though, there is a nontrivial A which can eliminate the dyw^2 term. We can then proceed perturbatively, removing all higher order terms in w, concluding that the normal form for the flows of y is $d\tilde{y}/d\ell = \lambda_y \tilde{y}(1 + cw)$.

We now proceed to analyze the coupled resonances of the x and u flows. Starting by writing out a few terms of the g-functions perturbatively as

$$\begin{aligned} dx/d\ell =&\ \ \lambda_x\ x(1+a_1w+a_2w^2+a_3w^3),\\ du/d\ell =&-\lambda_u\ u(1+b_1w+b_2w^2+b_3w^3),\\ dw/d\ell =&\ \ \ c_2\ w^2+c_3w^3, \end{aligned} \tag{5.18}$$

where we have used the fact that $p\lambda_x = q\lambda_u$, and c_2 and c_3 are functions of the a_i and b_i. Since x and u are coupled we must propose a simultaneous coordinate change

$$\begin{aligned} x(x_1,u_1) &= x_1(1+A_1w_1+A_2w_1^2),\\ u(x_1,u_1) &= u_1(1+B_1w_1+B_2w_1^2),\\ w_1 &= x_1^p u_1^q. \end{aligned} \tag{5.19}$$

Inserting these new coordinates into Eqn. 5.18 and collecting terms in powers of w_1 yields

$$\begin{aligned}
dx_1/d\ell = {} & \lambda_x \; x_1(1 + a_1 w_1 + \\
& \qquad (a_2 + q(a_1 B_1 + A_1 b_1 \lambda_u/\lambda_x))w_1^2 + \\
& \qquad (a3 + f_1)w_1^3 + \mathcal{O}(w_1^4)), \qquad (5.20) \\
du_1/d\ell = {} & -\lambda_u \; u_1(1 + b_1 w_1 + \\
& \qquad (b_2 + q(b_1 A_1 + B_1 a_1 \lambda_x/\lambda_u))w_1^2 + \\
& \qquad (b_3 + g_1)w_1^3 + \mathcal{O}(w_1^4)), \qquad (5.21)
\end{aligned}$$

where f_1 and g_1 are some polynomials of all the parameters. The first fact to note is that the terms linear in w_1 do not depend on the new coordinates; the xw and uw terms cannot be removed. Moving on to the quadratic terms, we see that we can choose to remove either the xw^2 or the uw^2 terms, but not both. For instance, if we choose $B_1 = -\lambda_u(b_2 + A_1 b_1 p)/a_1 p \lambda_u$ we can remove the $u_1 w_1^2$ term, but inserting this choice into the x_1 flows yields a coefficient of $x_1 w_1^2$ equal to $a_2 - b_2 q \lambda_u / p \lambda_x$; A_1 has been cancelled so we cannot remove that term. This limitation does not apply to the higher order terms in w_1 as A_1, A_2, and others are free parameters. Therefore we can find some sequence of polynomial coordinate transformations which yields the normal form

$$\begin{aligned}
d\tilde{x}/d\ell = {} & \lambda_x \; \tilde{x}(1 + a\tilde{w} + b\tilde{w}^2), \\
d\tilde{y}/d\ell = {} & \lambda_y \; \tilde{y}(1 + c\tilde{w}), \\
d\tilde{u}/d\ell = {} & -\lambda_u \; \tilde{u}(1 + d\tilde{w}), \qquad (5.22)
\end{aligned}$$

where, since they cannot be changed by smooth coordinate transformations, there exist ratios of a, b, c,and d which are universal.

Note that this sequence of calculations leading to our normal form is not *ad hoc* — it is a prescribed, systematic mathematical procedure developed by the dynamical systems community [Yu and Leung, 2007] to study bifurcations.

5.5 Free energy solution including a resonance

We now explore the free energy singularities caused by resonances with irrelevant variables, namely terms like $t^n(\log t)^m$ and $t^n \log\log t$, for some integers n and m. Note that the latter is not predicted to be a part of any known exact results, and the former is not a part of the exact free energy (though some claim it appears in the susceptibility) [Orrick et al., 2001]. We claim that if terms in the flows lead to singularities like these in the free energy, the amplitude of those resonant terms in the flows must be zero.

Consider some irrelevant variable u such that $p\lambda_t = q\lambda_u$ for positive integers p and q, where, as before $\lambda_t = 1$ and λ_u is the absolute value of the eigenvalue of u. We have shown that in this case the normal form of the flows must be

$$\begin{aligned}
df/d\ell =& \quad 2\ f - t^2, \\
dt/d\ell =& \qquad t(1+aw), \\
dh/d\ell =& \ \ \lambda_h\ h(1+bw), \\
du/d\ell =& -\lambda_u\ u(1+a_1 w + a_2 w^2),
\end{aligned} \tag{5.23}$$

where $w = t^p u$. It will be more convenient to work directly with the flow of w, which is

$$dw/d\ell = cw^2 + dw^3, \tag{5.24}$$

where $c = ap$ and $d = q\lambda_u a_1$. Defining $s(l) = c/dw(l) + 1$, we find that $s(l) = W(s_0 \exp(s_0)\exp(-c^2 l/d))$ where $s_0 = s(0)$ and W is the 0 or -1 branch of the Lambert-W function depending on the sign of t_0.

We could solve for $t(l)$ and then $f(l)$ directly now that we have $w(l)$, but since coarse-graining till $t = 1$ is more complex now, we proceed a different way. Integrating f directly and solving for the initial condition, our predicted free energy should be

$$f_0 = e^{-2\ell^\star} f(h(\ell^\star), u(\ell^\star)) - \int_0^{\ell^\star} e^{-2\ell} t(\ell)^2 d\ell \tag{5.25}$$

where as usual $t(\ell^\star) = 1$. Let us first find $\ell^\star = \ell(t = 1)$ by noticing that we can solve for $\ell(w)$ from eqn. 5.24 above, and then substituting in $w(t)$ by solving dw/dt. It is again easier to work with s as defined above, finding

$$s(t) = \tilde{a} W\left(\frac{s_0}{\tilde{a}} e^{s_0/\tilde{a}} (t/t_0)^{-c^2/\tilde{a}d}\right), \tag{5.26}$$

where $\tilde{a} = ac/d - 1$. Now, $\ell^\star = w(t = 1)$, and using the identity $\exp W(x) = x/W(x)$ we find the scaling term of the free energy to be

$$\begin{aligned} e^{-2l^\star} &= t_0^2 \left[\frac{\tilde{a}}{s_0} W\left(\frac{s_0}{\tilde{a}} e^{s_0/\tilde{a}} t_0^{c^2/\tilde{a}d}\right)\right]^{2d(1-\tilde{a})/c^2}, \\ &= t_0^2 \left(\frac{s(1)}{s_0}\right)^{2\frac{ac-2d}{c^2}}. \end{aligned} \tag{5.27}$$

To proceed we recall that since $s = c/dw + 1$, when w_0 is small after coarse-graining s_0 will be large. Therefore we can understand the nature of the singularities produced by the resonance by expanding about $s_0 \to \infty$:

$$\begin{aligned} \left(\frac{W(yxe^x)}{x}\right)^b = 1 + \frac{b \log y}{x} + b\frac{(b-1)(\log y)^2 - 2\log y}{2x^2} \\ + \mathcal{O}\left(\frac{1}{x}\right)^3, \end{aligned} \tag{5.28}$$

where $y = t_0^{c^2/\tilde{a}d}$ and $x = s_0/\tilde{a}$. Expanding to higher order will produce terms proportional to $t_0^2 \log(t_0)^n$ for all $n > 0$ in the scaling part of the free energy. We can thus write the scaling part of the free energy which multiplies the scaling function as

$$e^{-2l^\star} = t_0^2 \left(\sum_{nm} y_{nm} t_0^n (\log t_0)^m\right), \tag{5.29}$$

where $y_{00} = 1$ and all other y_{nm} are explicit functions of a, c, d, and s_0, which vanish as $s_0 \to \infty$ (corresponding to a vanishing $u_0 = 0$ of the amplitude of the corresponding irrelevant perturbation).

This then modifies the prediction of the free energy to be (ignoring the integral term in eqn. 5.25)

$$f_0(t_0, h_0, u_0) = t_0^2 \left(\sum_{nm} y_{nm} t_0^n (\log t_0)^m \right) f(h(\ell^\star), u(\ell^\star)). \tag{5.30}$$

Let us compare this prediction to the exact free energy in zero field, which is $f(t) = a(t) \log t + b(t)$ with some analytic functions a and b. We see that $y_{nm} = 0$ for $n, m > 0$ must hold for the amplitudes of the resonances in the flows of eqn. 5.23 to agree with the exact solution. Examining the exponent of eqn. 5.27 we see that if $ac = 2d$, all of the log corrections disappear, and $\exp(-2l^\star) = t_0^2$ as if there were no resonances. In the variables of the flows in eqn. 5.23, $ac = 2d$ if $a_1 = a^2/2\lambda_t$. Hence either the universal terms $ac = 2d$ or the amplitudes of the irrelevant perturbations (and hence y_{mn}) must vanish in the lattice models.

We can further constrain the possibilities by analyzing the integral term of eqn. 5.25. We first change variables, inverting our previous solution $s(l)$ to find $l(s)$, and invert eqn. 5.26. We find

$$\begin{aligned} -\int_0^{\ell^\star} e^{-2\ell} t(\ell)^2 d\ell &= -\int_{s_0}^{s(t=1)} e^{2l(s)} t(s)^2 \frac{dl}{ds}\, ds, \\ &= t_0^2 \left(B_1 + (B_2 - B_3\, s(1)) \left(\frac{s(1)}{s_0} \right)^{\frac{2(2d-ac)}{c^2}} \right), \end{aligned} \tag{5.31}$$

where the constants B_i are functions of a, c, d, and s_0, and $s(1) = s(t = 1)$ as given by eqn. 5.26. Curiously, whereas the condition $ac = 2d$ was found to cancel all the powers of logarithms in eqn. 5.27, the powers of logarithms persist in eqn. 5.31 in this limit. Since

$s(1)$ has essentially the same asymptotic behavior as the scaling part of the free energy in eqn. 5.28, we see that the predicted free energy will thus have the form

$$\begin{aligned} f_0(t_0,h_0,u_0) =& t_0^2\left(\sum_{nm} y_{nm} t_0^n (\log t_0)^m\right) f(h(\ell^\star),u(\ell^\star)) \\ &+ t_0^2\left(\sum_{nm} x_{nm} t_0^n (\log t_0)^m\right). \end{aligned} \tag{5.32}$$

Therefore since even if $ac = 2d$ the irrelevant variable resonances of the form of eqn. 5.23 produce logarithmic singularities which are inconsistent with Onsager's exact solution, we conclude that such resonances cannot contribute to the flow equations of the two dimensional square lattice Ising model: either $u_0 = 0$ (and hence the $x_{mn} = 0$ for the lattice model, or the universal terms in the flow equations $a = b = d = c = 0$.

5.6 Corrections to the Susceptibility

It has been hypothesized that these resonances we have been discussing could contribute logarithmic singularities to the zero-field magnetic susceptibility $\chi = \partial^2 f/\partial h^2$. To understand the effect of resonances of the form of eqn. 5.23 on χ, we must study the new scaling variable $h(\ell^\star)$ in the scaling function of eqn. 5.25. The standard scaling variable predicted by the linearized flows is $h/t^{\lambda_h/\lambda_t}$, which is found by integrating $dh/dt = (dh/d\ell)/(dt/d\ell)$, which should be invariant under coarse-graining.

We proceed another way, starting with the ansatz $h(l) = e^{\lambda_h l} q(l)$ for some function $q(l)$, which, when combined with eqn. 5.23 suggests $q'(l) = \lambda_h b q(l) w(l)$, which can be solved by integration similar to eqn. 5.31. The result, upon coarse-graining till $t = 1$ is

$$\begin{aligned} h(l^\star) = h_0 \left(t_0^2 \frac{s(1)}{s_0}\right)^{\lambda_h \frac{(ac-2d)(bc-d)}{c^2 d}} \\ \times e^{\frac{\lambda_h b}{c}(s_0 - s(1))} \left(\frac{1-s_0}{1-s(1)}\right)^{\frac{2\lambda_h b}{c}}, \end{aligned} \tag{5.33}$$

where as before $s(1) = s(t = 1) = s(\ell^\star)$ is given by eqn. 5.26. Therefore we can see that the field scaling variable $h(\ell^\star)$ will also give rise to powers of $\log t_0$. Will any of these logarithms not found in the exact solution of the 2D Ising model persist along the critical manifold $h_0 = 0$? First, we must take two derivatives with respect to h_0 of the free energy in eqn. 5.25, and then take the limit of $h_0 \to 0$ to predict the zero field susceptibility. It does not appear possible that any of these logarithms multiplied by h_0 in $h(\ell^\star)$ can explain the powers of logs which have been observed in the susceptibility [Orrick et al., 2001]. Further, it does not appear possible to produce such powers of logarithms in the susceptibility via resonance terms in the flows *without* adding powers of logarithms to the free energy, which we know lacks these corrections.

5.7 Finite Size Scaling

One way to perform finite size scaling is to make the system size a control variable, adding it to the flow equations. Since the critical point can only occur at infinite system size, we add a flow in $dL^{-1}/d\ell = L^{-1}$. This new equation has an eigenvalue of 1, and so normal form theory leads us to suspect the presence of resonances. Are these resonances physically meaningful, or artifacts of the analysis? We can compare the predictions of NFT to the exact finite size heat capacity [McCoy and Wu, 2014] to learn more. There will be a resonance between f and L^{-1} since $0 \cdot D + 2 \cdot \lambda_{L^{-1}} = D$ for $D = 2$, and with t. One choice of the normal form is (ignoring the irrelevant resonances and analytic corrections)

$$\begin{aligned} d\tilde{f}/d\ell &= 2\tilde{f} + a\tilde{t}^2 + bL^{-2} \\ d\tilde{t}/d\ell &= \tilde{t} + cL^{-1} \\ dL^{-1}/d\ell &= L^{-1}. \end{aligned} \tag{5.34}$$

Solving eqn. 5.34 and coarse-graining till $\ell = \log L_0$ for system size L_0, we predict the finite size scaling form

$$\begin{aligned} f(t,L) =& L^{-2}(\Phi(c\log L + tL) - b\log L + \frac{ac^2}{3}(\log L)^2) \\ &+ t^2\log L + ac\frac{t}{L}(\log L)^2, \end{aligned} \tag{5.35}$$

where $\Phi(x)$ is a universal scaling function.

Taking two derivatives with respect to t we find the heat capacity as

$$c(t) = -2a\log L + \Phi''(Lt + cL^2\log L). \tag{5.36}$$

We can proceed with a scaling collapse by solving for the scaling function, which is universal, so that if we have the correct parameters a and c, all curves independent of L should lie on top of each other. Note that $a = \log^2(1+\sqrt{2})/\pi$ is known from asymptotic analysis [Ferdinand and Fisher, 1969], and so we need only search for a collapse by varying c.

Following Salas [Salas, 2001] and Ferdinand [Ferdinand and Fisher, 1969] we can calculate the exact specific heat per spin for any finite system size, omitting the analytic part of the calculation to focus on the singular behavior. The top of fig. 5.1 shows the finite size scaling collapse using the scaling form of eqn. 5.36 as predicted by normal form theory. Note that $c \approx 0$ is necessary to get a decent collapse, so if the resonance term with L^{-1} contributes to the t flows it does so with a rather small amplitude. The bottom

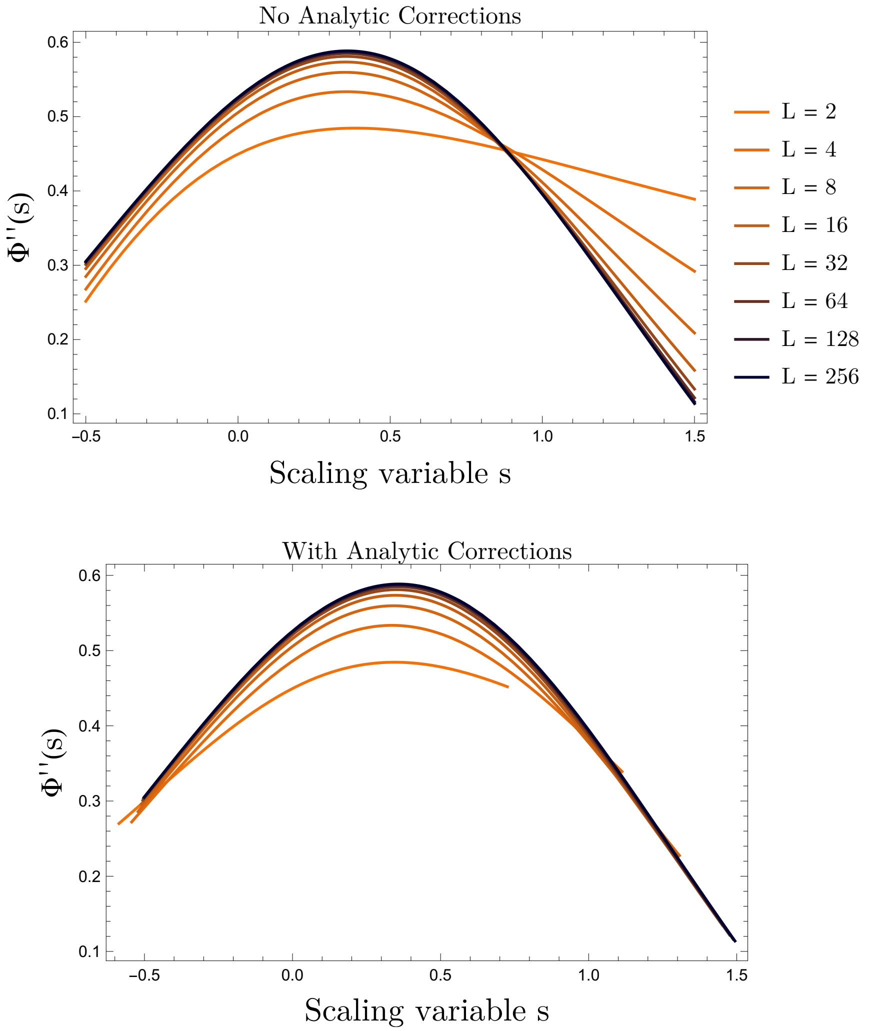


Figure 5.1: Finite size scaling collapse of the 2D Ising exact specific heat. Top: with no analytic corrections $\tilde{t}$. Bottom: with analytic corrections. The coefficient of the resonance correction in the normal form in eqn. 5.34 is consistent with $c = 0$.

of fig. 5.1 shows the same scaling collapse using the nonlinear scaling field we found in eqn. 5.14, with a modest improvement as the peaks of each finite size curve are better aligned. Note here that we did not incorporate the expected finite-size analytic corrections to scaling given by expanding $\tilde{t}(t, L^{-1})$ and $\tilde{f}(f, t, L^{-1})$ which presumably would have acted to remove the remaining finite-size effects; see Ferdinand and Fisher [Ferdinand and Fisher, 1969].

5.8 Legendre Transform of the Flows

In our previous work on normal form theory [Raju et al., 2019] we noticed that the t^2 resonance term in the free energy flow equation of the 2D Ising model led to what appeared to be a transcritical bifurcation in the flow equation of the specific heat. This necessarily led us to deviate from the history of RG, which limited its calculation to the canonical ensemble, which expresses the free energy as a function of temperature and field, for example. There is no physical origin for the choice of using the canonical ensemble; we could for example study the Ising model in the microcanonincal ensemble, where the thermodynamic potential is entropy as a function of energy instead of temperature.

The RG has historically separated the control and observation variables. There is no a priori reason to make this distinction in statistical mechanics: what would the flow equations look like in a different ensemble? The relationship between different ensembles can be understood as a Legendre transformation [Zia et al., 2009]. Here we will restrict our attention to the Legendre transformation of free energy with respect to temperature, which yields the entropy as a function of energy. Writing the free energy as $\mathcal{F} = -\log \mathcal{Z}$, with partition function $\mathcal{Z}$, we know that

$$\mathcal{F}(\beta) = tE - S, \tag{5.37}$$

where S is the entropy and E is the energy. This is just one of many thermodynamic potentials, however, and we could just as well write

$$\mathcal{F}(\beta(E)) = t(E)\frac{df}{dt} - S(E), \tag{5.38}$$

where E and t are related by the fact that $df/d\beta = E$. Solving for the entropy,

$$S(E) = Et(E) - \mathcal{F}(E), \tag{5.39}$$

we see that $S(E)$ is an equivalent expression to $\mathcal{F}(\beta)$ in that they are related by an information-preserved invertible transformation called the Legendre transform.

So if thermodynamic potentials are really not physically significant, what does RG look like as flows in S and E instead of f and t? Writing our f flows as

$$\frac{df}{d\ell} = Df + \Pi, \tag{5.40}$$

where Π is an analytic function of all the variables t, h, u, etc. Assuming the flows in all other observables θ^γ are of the form $d\theta^\gamma/d\ell = \beta^\gamma$ for some function $\beta^\gamma = \lambda_y y + g_\gamma(\{\theta^\delta\})$, then the total derivative with respect to l can be written (where repeated indices are summed) $d/d\ell = \beta^\gamma\partial_\gamma$. Therefore,

$$\frac{df}{d\ell} = \beta^\gamma\partial_\gamma f = Df + \Pi, \tag{5.41}$$

so that we can find flows of $E = \partial_t f$ for example by taking partial derivatives of the flow to find

$$\begin{aligned} \partial_\alpha(\beta^\gamma\partial_\gamma f) &= D\partial_\alpha f + \partial_\alpha\Pi \\ \partial_\alpha\beta^\gamma\partial_\gamma f + \beta^\gamma\partial_\alpha\partial_\gamma f &= D\partial_\alpha f + \partial_\alpha\Pi. \end{aligned} \tag{5.42}$$

We can then commute partial derivatives as $\beta^\gamma\partial_\alpha\partial_\gamma f = \beta^\gamma\partial_\gamma\partial_\alpha f = d(\partial_\alpha f)/d\ell$ to find

$$\frac{d(\partial_\alpha f)}{d\ell} = D(\partial_\alpha f - \partial_\alpha\beta^\gamma\partial_\gamma f) + \partial_\alpha\Pi. \tag{5.43}$$

We can now find the flows of S by direct differentiation and simplification to be

$$\begin{aligned}\frac{dS}{d\ell} &= DS - \Pi + \beta^t E \\ &\quad + t\left(\partial_t \Pi - D\partial_t \beta^\gamma \partial_\gamma f\right), \end{aligned} \tag{5.44}$$

where $t = t(E)$. If we assume that all the other flow equations are linear so that $\beta^\gamma = \lambda_{(\gamma)}\theta^{(\gamma)}$ (no summing over indices), these flows simplify to

$$\begin{aligned}\frac{dS}{d\ell} &= DS + (t\partial_t - 1)\Pi \\ \frac{dE}{d\ell} &= (D - \lambda_t)E + \partial_t \Pi, \end{aligned} \tag{5.45}$$

which, if $\Pi = 0$, causes the flows in S and E to be hyperbolic and behave as expected from standard RG theory. Since in the 2D Ising model, $\Pi = -t^2$ due to a resonance, however, we predict the normal form flows of S and E are

$$\frac{dS}{d\ell} = DS - t(E)^2 \tag{5.46}$$

$$\frac{dE}{d\ell} = (D - \lambda_t)E - 2t(E), \tag{5.47}$$

which is problematic because $E(t) = -2at\log t$ near the critical point, and does not admit a Taylor series, so that the inverse function $t(E)$ cannot have a Taylor series. We conclude that any nonlinear terms in the flows of f and t must in general produce non-analytic flow equations in S and S. This is troubling because the universal behavior of RG depends on the vicinity to the critical point being described by an analytic flow equation. Does nature prefer one ensemble over another for universal behavior in coarse-graining?

Since we can calculate f and E analytically for the 2D Ising model, we can numerically invert $t(E)$ to find E, and then plot $S(E) = t(E)E - f(t(E))$ versus E. We can then test two analytic predictions. We can solve the non-analytic flows of eqn. 5.47 to predict

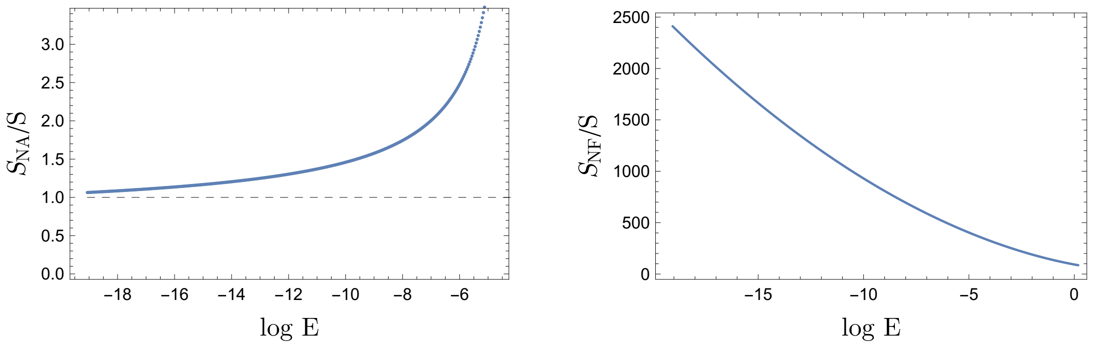


Figure 5.2: Testing the asymptotic behavior of two theories of the 2D Ising entropy $S(E)$. (left) $S_{\mathrm{NA}}(E)/S(E)$ is plotted as a function of E, where $E = 0$ is the critical point, and where $S_{\mathrm{NA}}(E)$ of eqn. 5.48 is predicted by the Legendre transform of the 2D Ising normal form flows. (right) $S_{\mathrm{NF}}(E)/S(E)$ is plotted as a function of E, where $S_{\mathrm{NF}}(E) = -E^2 \log(Ee^{-}\tau)$ is the prediction of the entropy assuming that the flows of S and E should be analytic.

that near the critical point the non-analytic S_{NA} is

$$S_{\mathrm{NA}}(E) = \frac{E^2}{W(e^{1-\tau/2}E)}\left(\frac{3}{2} + \frac{1}{4W(e^{1-\tau/2}E)}\right), \tag{5.48}$$

where e is the base of the natural logarithm and W is our old friend the Lambert product log function in its -1 branch. $\tau \approx 3.7476$ comes from Onsager's solution and is the ratio of the coefficients of t^2 and $t^2 \log t$ in eqn. 5.13. The competing predictions are assuming Normal Form Theory holds for the flows, and assuming that $t(E)$ is analytic. In the latter case, there is still a resonance between $D = 2$ and $D - \lambda_t = 1$, so that near the critical point the normal form entropy S_{NF} is

$$\frac{dS}{d\ell} = DS - E^2 \tag{5.49}$$

$$\frac{dE}{d\ell} = (D - \lambda_t)E, \tag{5.50}$$

which then produces a familiar prediction $S_{\mathrm{NF}}(E) = -E^2 \log(Ee^{-\tau})$.

$$\begin{tikzcd} f,t \arrow[r, "\mathrm{CT}"'] \arrow[d, "\mathrm{LT}"'] & \tilde{f},\tilde{t}\ \text{(nonlinear)} \arrow[d, "\mathrm{LT}"] \\ S,E & \tilde{S},\tilde{E}\ (\mathrm{non}-\mathrm{analytic}) \end{tikzcd}$$

Figure 5.3: Relationship between quantities under coordinate transforms (CT) and Legendre transforms (LT). If one begins with hyperbolic (linear) flows in f, t, these will gives rise to hyperbolic flows of S, E under LT. If one performs a CT to find $\tilde{f}$, $\tilde{t}$, the resulting flows after LT will generically be non-analytic.

Figure 5.2 shows the asymptotic behavior of these two predictions. On the left, is the ratio of eqn. 5.48 over the true Ising entropy $S_{\mathrm{NA}}(E)/S(E)$. We see that as the critical point is approached (here we set $E(t=0)=0$), the ratio asymptotically approaches unity. On the right, we see the ratio of $S_{\mathrm{NF}}(E)/S(E)$, which assumed an analytic flow equation for S and E of the form of eqn. 5.50. We see that this theory diverges from the true entropy as $E \to 0$. We must conclude that the flow equations of S and E, at least in the case of the 2D Ising model are in fact non-analytic.

Our theory of the Legendre transformation of the flows seems to suggest that if one wishes for analytic flow equations in S and E, for example, one must start with hyperbolic or linear flows in f and t, etc. Therefore any system with resonances like the 2D and 4D Ising models will inherently have non-analytic flows of the entropy. More disturbingly, even if we begin with linear flows in $\tilde{f}$ and $\tilde{t}$, as we may do in the 3D Ising model which has no resonances, analytic corrections of the form $\tilde{t} = t + u_t(t)$ generically produce non-analytic flows in S. This can be seen, as such a transformation leaves powers of t in $dS/d\ell$, which when inverted to $t(E)$ generically produces a non-analytic function of E. This relationship is indicated schematically in fig. 5.3, where basically Legendre transformation and coordinate transformation of RG flows do not commute. Therefore it would seem perhaps that nature prefers doing RG in the canonical ensemble over the microcanonical ensemble.

5.9 Discussion

Conformal field theory predicts an infinite number of irrelevant variables for the 2D ising model of the form $\lambda_u = -2n, -(2n + 1/8)$ for all integers $n > 1$. We know from the exact solution of the 2D square lattice Ising model that singular corrections from irrelevant operators with eigenvalues $\lambda_u = -(2n + 1/8)$ are constrained, as no fractional powers of t are present in the free energy. This means that for many of the terms, either $u_0 = 0$, or the free energy scaling function happens to be independent of u. We have shown in this work that resonances due to operators with eigenvalue $\lambda_u = -2n$ also contribute no corrections to scaling, concluding that either $u_0 = 0$ for the square lattice Ising model, or the coefficients of the resonances in the flows are zero. In this section, we explain a conjecture that these integer irrelevant variables are identically zero for the square lattice Ising model, by arguing that they are redundant [Cardy, 1996] or gauge irrelevant [Raju and Sethna, 2018] variables.

Redundant operators are those that only change non-universal properties like the position of the critical point, leading to a sub-manifold of universally-equivalent RG fixed points [Fisher and Randeria, 1986]. The most famous example of a redundant operator is the ϕ^3 term in the Hamiltonian density of the Ising model. Writing a Landau-like expansion for the Hamiltonian in terms of some order parameter field ϕ as

$$\mathcal{H} = \int \mathrm{d}\mathbf{x}\ (\nabla\phi)^2 + t\phi^2 + b\phi^3 + u\phi^4 + \ldots, \tag{5.51}$$

we can see that invoking the transformation $\phi \to \phi + K$, for some choice of K which depends on u, the ϕ^3 term can be eliminated. Dimensional analysis arguments will classify

ϕ^3 as relevant, but we know that a constant-offset of the field K should not change the physics, and so we conclude that the ϕ^3 term is redundant, playing no part in the RG dynamics. Note that since we started with an arbitrary coefficient in front of ϕ^3, there is at least a whole line of equivalent RG fixed points for this Hamiltonian.

Previous work has shown that in the case of the period-doubling onset of chaos, there are an infinite number of irrelevant operators which are redundant, playing no part in the RG flows [Raju and Sethna, 2018]. This choice of gauge for the functional map of the period doubling analysis led to the new name of gauge irrelevant variables, which are both redunandant and irrelevant. We believe that a compelling and possible explanation for the fact that none of the integer-irrelevant variables $\lambda_u = -2n$ contribute to the scaling of the 2D Ising model, is that they, too, are gauge-irrelevant.

Consider the transformation $\phi' = \phi + a(\partial_x^4 \phi + \partial_y^4 \phi)$. This is the lowest-order modification which will introduce anisotropy consistent with the square-lattice, which Caselle et al. [Caselle et al., 2002] say is responsible for the most dangerous irrelevant variables with eigenvalue $-2, -4$. How does this transformation perturb the scaling of the correlation function? The scaling ansatz of the correlation function is [Chen et al., 2013]

$$C(r) = \langle \phi(0)\phi(r) \rangle = r^{-\eta} \mathcal{C}(r/t^{-\nu}, h/t^{\beta\delta}, u/t^{\lambda_u/\lambda_t}), \tag{5.52}$$

where $\eta = 1/4$, $\nu = 1$, $\beta = 1/8$, and $\delta = 15$ are the 2D Ising critical exponents. Putting our transformation ϕ' into this ansatz, we can calculate the corrections to scaling due to these irrelevant perturbations. The leading correction due to $\phi' = \phi + a(\partial_x^4 \phi + \partial_y^4 \phi)$ scales as $r^{-4-\eta}$. Perturbing ϕ with a Laplacian of ϕ produces a leading correction which scales as $r^{-2-\eta}$. We could also propose the transformation $\phi' = \phi + a\phi^3$, using CFT to predict the corrections due to 3 and 4-point correlation functions. These integer singular corrections could also be caused by the integer irrelevant eigenvalues predicted by CFT.

Therefore we hypothesize that (for the two dimensional) Ising model, the reason there do not appear resonance corrections and singular scaling is because all these irrelevant operators are redundant or gauge irrelevant. We will devote future works to fleshing out this line of inquiry.

5.10 Conclusion

We have studied the normal form of the 2D Ising model flow equations, clarifying the understanding of its corrections to scaling. Using the normal form we showed that the scaling ansatz of the free energy should not multiply the log by a scaling function, and we used the exact solution to find the nonlinear scaling field in $\tilde{t}(t)$. We found the resonances due to irrelevant operators like $\lambda_u = -3/4, -2, -4, \ldots$, showing that the powers of logarithms they would produce in the free energy contradict Onsager's solution which contains a single log. We concluded that either the irrelevant variables must be zero at the critical point, as Caselle et al. [Caselle et al., 2002] found for the irrelevant operator $\lambda_u = -2$, or the coefficients in the nonlinear terms of the flows must be zero. We concluded further that the resonances cannot predict powers of logs in the zero field susceptibility without also adding spurious powers of logarithms to the free energy. We explored the normal form for the finite-size 2D Ising model, used our nonlinear scaling field to improve the scaling collapse of the specific heat, and concluded numerically that resonance with L^{-1} either does not arise or happens with rather small amplitude. Finally we investigated the behavior of the RG flows under the Legendre transformation, finding that generically flows in the entropy and energy will be non-analytic, in stark contrast to the analytic assumptions underpinning RG.

5.11 Appendix

5.11.1 Analytic Corrections and Irrelevant Variables

Corrections to scaling from irrelevant variables are usually called singular because upon expansion of the scaling function their generically non-integral exponents lead to singularities in predictions from the RG. In the 2D Ising model the most dangerous of the irrelevant variables have integer eigenvalues[Caselle et al., 2002], and so their contributions are indistinguishable from analytic corrections. Therefore we must understand the extent to which irrelevant variables limit our ability to infer the flow equation of the 2D Ising model from coordinate transformations of the exact solution.

Here we show an example that demonstrates that our coordinate transformation will yield the coefficients of the analytic terms of the flows as functions of the irrelevant variables of the square-lattice Ising model. in the flows which depend on irrelevant variables. Define an analytic change of variables $t = \tilde{t}\exp(-Au/\lambda)$. Under this change of variables the normal form, including some irrelevant variable u with eigevalueλ transforms to

$$df/d\ell = 2f - t^2 e^{2Au/\lambda} \tag{5.53}$$

$$dt/d\ell = t(1 + Au) \tag{5.54}$$

$$du/d\ell = -\lambda u. \tag{5.55}$$

Could we infer the constant A by transforming the exact solution to the normal form solution? Applying the inverse of this coordinate transformation to the normal form solution $\tilde{f}_0 = \tilde{t}_0^2(\mathcal{F}(u_0\tilde{t}_0^2) - \log\tilde{t}_0)$ we can quickly find the free energy predicted by

Eqn. 5.55 (dropping the subscripts for brevity)

$$f(t) = t^2 e^{2Au/\lambda}(\mathcal{F}(ut^2 e^{2Au/\lambda}) - \log t - Au/\lambda) \tag{5.56}$$

where $\mathcal{F}$ a universal scaling function. Say Eqn. 5.56 is the true solution to the 2D Ising model, with true corresponding flows specified by Eqn. 5.55. If we find the coordinate transform bringing Eqn. 5.56 to the normal form solution, does the does the inverse of this allow us to infer the flows correctly? The transformation $\tilde{t} = t \exp Au_0/\lambda$ will bring Eqn. 5.56 to the normal form, where u_0 is the constant value of u for the Ising model. Since we can not understand how the exact solution varies with u_0 we infer the flows

$$\frac{d\tilde{f}}{d\ell} = 2\tilde{f} - \tilde{t}^2 e^{2Au_0/\lambda} \qquad\qquad \frac{d\tilde{t}}{d\ell} = \tilde{t}. \tag{5.57}$$

These flows are equivalent to the normal form we began with, up to a rescaling of $\tilde{t}$ by $\exp(2Au_0/\lambda)$, and we have failed to recover any of the terms which depend on u. We conclude that since we cannot vary irrelevant variables in the exact solution we will not be able to infer any of the terms with irrelevant variables in the flow equations. Conversely, we will not be able to infer the contributions of irrelevant variable to the scaling variables $\tilde{t}$, $\tilde{f}$.

Chapter 6

On the Use of ArXiv as a Dataset

6.1 Introduction

Real world datasets are typically multimodal (comprised of images, text, and time series, etc) and have complex relational structures well captured by a graph. Recently, advances have been made on models which act on graphs, allowing the rich features and relational structures of real-word data to be utilized [Hamilton et al., 2017b,a, Battaglia et al., 2018, Goyal and Ferrara, 2018, Nickel et al., 2016]. Many of these advances have been facilitated by the availability of large, benchmark datasets: for example, the ImageNet [Russakovsky et al., 2015] dataset has been widely used as a community standard for image classification. We believe the arXiv can provide a similarly useful benchmark for large scale, multimodal, relational modelling.

The work constituting this chapter was done in collaboration with Matthew Bierbaum, Kevin O'Keeffe, and Alexander A. Alemi. It was presented in an ICLR 2019 workshop and can be found at `arXiv:1905.00075`.

6.1. Introduction

The arXiv[1] is the de-facto online manuscript pre-print service for Computer Science, Mathematics, Physics, and many interdisciplinary communities. Since 1991 the arXiv has offered a place for researchers to reliably share their work as it undergoes the process of peer-review, and for many researchers it is their primary source of literature. With over 1.5 million articles, a large multigraph dataset can be built, including full-text articles, article metadata, and internal co-citations.

The arXiv has been used many times as a dataset. Liben-Nowell and Kleinberg [2007] used the topology of the arXiv co-authorship graph to study link prediction. Dempsey et al. [2019] used the authorship graph to test a hierarchically structured network model. Lopuszynski and Bolikowski [2013] used the category labels of arXiv documents to train and assess an automatic text labelling system. Dai et al. [2015] used a subset of the full text available on the arXiv to study the utility of "paragraph vectors" for capturing document similarity. Alemi and Ginsparg [2015] used the fulltext to evaluate a method for unsupervised text segmentation. Eger et al. [2019] and Liu et al. [2018] built models to predict future research topic trends in machine learning and physics respectively. The arXiv also formed the basis of the popular 2003 KDD Cup [Gehrke et al., 2003], in which researchers competed for the prize of best algorithm for citation prediction, download estimation, and data cleaning[2].

All these works used different subsets of arXiv's data, limiting their potential impact, as future researchers will be unable to directly compare their work to these existing results. The goal of this paper is to improve this situation by providing an open-source pipeline to standardize, simplify, and normalize access to the arXiv's public data, providing a benchmark to facilitate the development of models for multi-modal, relational data.

[1] https://arxiv.org

[2] The data for those challenges are available at `http://www.cs.cornell.edu/projects/kddcup/datasets.html`

6.2 Dataset

We built a freely available, open-source pipeline[3] for collecting arXiv metadata from the Open Archive Initiative [Lagoze and Van de Sompel, 2001], and bulk PDF downloading from the arXiv[4]. Further, this pipeline converts the raw PDFs to plaintext, builds the intra-arXiv co-citation network by searching the full-text for arXiv ids, and cleans and normalizes author strings.

6.2.1 Metadata

Through its participation in the Open Archives Initiative,[5] the arXiv makes all article metadata[6] available, with updates made shortly after new articles are published[7]. We provide code for utilizing these public APIs to download a full set of current arXiv metadata. As of 2019-03-01, metadata for 1,506,500 articles was available. For verification and ease of use purposes, we provide a copy of the metadata (less abstracts) on the date we accessed it. An example listing is shown in Figure 6.1. Each article includes an arXiv id (e.g. 0704.0001)[8] used to identify the article, the publicly visible name of the submitter, a list of authors, title, abstract, versions and category listings, as well as optional doi, journal-ref and report-no fields. Of particular note is the first category listed, the *primary* category, of which there are 171 at this time. Notice that the list of authors is just a single string of author names, potentially joined with commas or ‘and’s. We’ve provided a suggested normalization and splitting script for splitting these

[3] https://github.com/mattbierbaum/arxiv-public-datasets/releases/tag/v0.2.0

[4] https://arxiv.org/help/bulk_data

[5] http://www.openarchives.org/

[6] https://arxiv.org/help/prep

[7] Further details available at https://arxiv.org/help/oa

[8] There are two forms of valid arXiv IDs, delineated by the year 2007, described in https://arxiv.org/help/arxiv_identifier.

`authors` strings into a list of author names. Additional fields may be present to denote `doi`, `journal-ref` and `report-no`, although these are not validated they can potentially be used to find intersections between the arXiv dataset and other scientific literature datasets. Population counts for the optional fields are shown in Table 6.1.

```
{'id': '1905.00075',
 'submitter': 'Colin B. Clement',
 'authors': 'Colin B. Clement, Matthew Bierbaum, Kevin P. O\'Keeffe, and Alexander A. Alemi',
 'title': 'On the Use of ArXiv as a Dataset',
 'comments': '7 pages, 3 figures, 2 tables',
 'journal-ref': '',
 'doi': '',
 'abstract': 'The arXiv has collected 1.5 million pre-prints over 28 years,
 hosting literature from physics, mathematics, computer science, biology,
 finance, statistics, electrical engineering, and economics. Each pre-print
 features text, figures, author lists, citation lists, categories, and other
 metadata. These rich, multi-modal features, combined with the natural
 relational graph structure created by citation, affiliation, and co-authorship
 makes the arXiv an exciting candidate for benchmarking next-generation models.
 Here we take the first necessary steps toward this goal, by providing a
 pipeline which standardizes and simplifies access to the arXiv's publicly available data. We
      use this pipeline to extract and analyze a 6.7 million edge citation graph, with an 11
     billion word corpus of full-text research articles. We present some baseline
     classification results, and motivate application of more exciting relational neural
     network models.'
 'categories': ['cs.IR'],
 'versions': ['v1']}
```

Figure 6.1: An example of what the metadata for this very article may look like if it were submitted to the arXiv.

Count	1,506,500	1,491,303	1,229,138	810,209	608,286	154,922
Field	`id`	`submitter`	`comments`	`doi`	`journal-ref`	`report-no`

Table 6.1: Number of articles with the corresponding field populated. Note that the fields `id`, `abstract`, `authors`, `versions`, and `categories` are always populated.

6.2.2 Full Text

One advantage the arXiv has over other graph datasets is that it provides a very rich attribute at each `id` node: the full raw text and figures of a research article. To extract the raw text from PDFs, we provide a pipeline with two parts. A helper script downloads the full set of PDFs available through the arXiv's bulk download service[9]. Since arXiv hosts their data in a requester-pay AWS S3 buckets, this constitutes $\sim$ 1.1 TB and $\sim$ \$100 to fully download. For posterity, we have provided MD5 hashes of the PDFs at the state of the frozen metagraph extraction. Raw TEX source is also available for the subset of articles that provide it. Second, we provide a standard PDF-to-text converter – powered by `pdftotext`[10] – to convert the PDFs to plaintext.

Using this pipeline, it is currently possible to extract a corpus of 1.37 million raw text documents. Figure 6.2 shows an example of the text extracted from a PDF. Though the extracted text isn't perfectly clean, we believe it will still be useful for many tasks, and hope future contributions to our repository will provide better data cleaning procedures.

The extracted raw-text dataset is $\sim$ 64 GB in size, totaling $\sim$ 11 billion words. An order of magnitude larger than the common billion word corpus [Chelba et al., 2013], this large size makes the arXiv raw-text a competitive alternative to other full text datasets. Moreover, the technical nature of the arXiv distinguishes it from other full text datasets. For example, the TEX data contained in the arXiv presents an opportunity to study mathematical formulae in bulk, as is done in the NTCIR-11 Task: Math-2 [Aizawa et al., 2014].

6.2.3 Co-Citations

While the arXiv does not currently publicly provide an API to access co-citations, our pipeline allows a simple but large co-citation network to be extracted. We extracted this network by searching the text of each article for valid arXiv ids, thereby finding which nodes should be linked to a given node in the co-citation network. We provide

[9] `https://arxiv.org/help/bulk_data`

[10] Version 0.61.1, available on most Debian systems from the `apt` package `poppler-utils`

```
Published as a conference paper at ICLR 2019

O N THE U SE OF A R X IV AS A DATASET
Colin B. Clement
Cornell University, Department of Physics
Ithaca, New York 14853-2501, USA
cc2285@cornell.edu

Matthew Bierbaum
Cornell University, Department of Information Science
Ithaca, New York 14853-2501, USA
mkb72@cornell.edu

Kevin O'Keeffe
Senseable City Lab, Massachusetts Institute of Technology
Cambridge, MA 02139
kokeeffe@mit.edu

Alexander A. Alemi
Google Research
Mountain View, CA
alemi@google.com

A BSTRACT
The arXiv has collected 1.5 million pre-print articles over 28 years, hosting literature from
      scientific fields including Physics, Mathematics, and Computer Science. Each pre-print
     features text, figures, authors, citations, categories, and other
metadata. These rich, multi-modal features, combined with the natural graph
structure---created by citation, affiliation, and co-authorship---makes the arXiv
an exciting candidate for benchmarking next-generation models. Here we take the
first necessary steps toward this goal, by providing a pipeline which standardizes
and simplifies access to the arXiv's publicly available data. We use this pipeline to
extract and analyze a 6.7 million edge citation graph, with an 11 billion word corpus of full
     -text research articles. We present some baseline classification results,
and motivate application of more exciting generative graph models.
```

Figure 6.2: Example text extracted from this PDF.

a compressed binary of the resulting network at the repository[11], so that researchers can study it directly, and avoid the difficulty of constructing it themselves. Table 6.2 summarizes the size and statistical structure of our co-citation network, compared with other popular citation networks. Šubelj et al. [2014] also studied data from the arXiv, but as indicated in the bottom row of Table 6.2, it used only the 34,546 articles from the 2003 KDD Cup challenge.

Table 6.2 reports standard statistics for the co-citation network. Our arXiv co-citation network contains $O(10^6)$ nodes, an order of magnitude larger than the $O(10^5)$ nodes in

[11] As part of one of the tagged releases: https://github.com/mattbierbaum/arxiv-public-datasets/releases

Table 6.2: **Graph statistics for popular citation networks**. All but the data for this work (first row) were taken from Table 1 and 2 in [Šubelj et al., 2014]. $\langle k \rangle$ is the average degree, and α_{in} and α_{out} are power law exponents of best fit for the degree distribution. WCC refers to the largest weakly connected components, computed using the python package 'networkx'. The power law exponents $\alpha_{\text{in}}, \alpha_{\text{out}}$ were found using the python module `powerlaw`. When fitting data to a powerlaw, the package discards all data below an automatically computed threshold x_{min}. These thresholds for k_{in} and k_{out} were $x_{\text{min}} = 73$ and $x_{\text{min}} = 59$ respectively.

Dataset	N_{nodes}	N_{edges}	$\langle k \rangle$	α_{in}	α_{out}	% WCC
arXiv	1.35×10^6	6.72×10^6	9.933	2.93	3.93	62
WoS	1.40×10^5	6.4×10^5	9.11	2.39	3.88	97
CiteSeer	3.84×10^5	1.74×10^6	9.08	2.28	3.82	95
KDD2003	3.34×10^4	4.21×10^5	24.50	2.54	3.45	99.6

the other citation networks. The exponents of best fit for the degree distributions α_{in} and α_{out} are consistent with the existing citation networks Šubelj et al. [2014], as it the the degree $\langle k \rangle$. 62% of the nodes are contained in the largest weakly connected component, while 31% of the nodes are fully isolated – meaning their in-degree k_{in} and out-degree k_{out} are zero. Recall that our arXiv co-citation network only contains publications which have been posted on the arXiv; a given paper which cites papers published elsewhere – and not on the arXiv – will have $k_{out} = 0$ in this set, which is an explanation the large number of isolated nodes.

Beyond constructing and analyzing a co-citation network, the arXiv dataset can be used for many tasks, such as relationally powered classification, author attribution, segmentation, clustering, structured prediction, language modeling, link prediction and automatic summary generation. As a basic demonstration, in Table 6.3 we show some baseline category classification results. These were obtained by training logistic regression on 1.2 million arXiv articles to predict in which category (e.g. `cs.Lg`, `stat.ML`) a given article resides. See Appendix .1 for a detailed explanation of the experimental setup. Titles and abstracts were represented by vectors from a pre-trained instance[12] of the Universal Sentence Encoder of Cer et al. [2018]. We see that including more aspects of each document (titles, abstracts, fulltext) and exposing their relations via co-citation leads to better predictive power. This is only scratching the surface of possible tasks and models applied to this rich dataset.

Table 6.3: Baseline classification performance on a holdout set of 390k articles. Titles and abstracts were embedded in a 512 dimensional subspace using the Universal Sentence Encoder, and trained on 1.2 million articles with logistic regression. 'All' refers to the concatenation of titles, abstract, fulltext, and co-citation features. 'All - X' refers to the ablation of feature X from 'All.' Top n is the classification accuracy testing when the correct class is in the top n most confident predictions. Detailed explanation of the features and methods can be found in Appendix .1.

Features	Top 1	Top 3	Top 5	Perplexity
Titles (T)	36.6%	59.3%	68.8%	12.7
Abstracts (A)	46.0%	70.7%	79.5%	7.5
Fulltext (F)	64.2%	79.4%	85.9%	4.6
Co-citation (C)	37.8%	49.4%	53.8%	18.5
All = T + A + F + C	**78.4%**	**91.4%**	**94.5%**	**2.3**
All - T	77.0%	90.7%	94.0%	2.5
All - A	74.7%	88.3%	91.9%	2.8
All - F	59.0%	79.8%	86.2%	4.6
All - C	75.5%	89.9%	93.6%	2.6

[12] From `https://tfhub.dev/google/universal-sentence-encoder/2`

6.3 Conclusion

As research moves increasingly towards structured relational modelling [Hamilton et al., 2017b,a, Battaglia et al., 2018], there is a growing need for large-scale, relational datasets with rich annotations. With its authorship, categories, abstracts, co-citations, and full text, the arXiv presents an exciting opportunity to promote progress in relational modelling. We have provided an open-source repository of tools that make it easy to download and standardize the data available from the arXiv. Our preliminary classification baselines support the claim that each mode of the arXiv's feature set allows for greatly improved category inference. More sophisticated models that include relational inductive biases—encoding the graph structures of the arXiv—will improve these results. Further, this new benchmark dataset will allow more rapid progress in tasks such as link prediction, automatic summary generation, text segmentation, and time-varying topic modeling of scientific disciplines.

.1 Logistic Regression Article Classification Baseline

ArXiv articles are assigned primary categories (e.g. `cs.AI` is Artifical Intelligene and `cs.CC` is computational complexity) by the article submitter, which is then confirmed by the ArXiv moderation system. This label can be obtained for each article from the OAI metadata described in the main article, and is the first element of a space-delimited string in the `categories` attribute. There are, at the time of writing, $L = 175$ possible categories. Since more categories can be added in the future and the metadata can be modified, please consult the frozen metadata file in the github repository release[13] for these 175 categories. This appendix explains how we developed the article classification baselines using features from the titles, abstracts, full-text, and co-citation network. The code for performing this task can be found in the `git` repository[14].

.1.1 Building Features

The title, abstract, and full-text of each article is a variable-length string, and each article has both a title and abstract from the OAI metadata, but not all articles have a full-text PDF. In our frozen dataset there are $N = 1,506,500$ articles with metadata, but only 1,357,536 have full-text in the ArXiv. We vectorized each string into 512 dimensions using the pretrained Universal Sentence Encoder,[15] substituting zeros for missing full-text.

The intra-ArXiv citation graph can be used via the $N \times N$ co-citation matrix, which

[13] `https://github.com/mattbierbaum/arxiv-public-datasets/releases`

[14] `https://github.com/mattbierbaum/arxiv-public-datasets/blob/v0.2.0/analysis/classification.py`

[15] `https://tfhub.dev/google/universal-sentence-encoder/2`

is defined as

$$M_{ij} = \begin{cases} 1 \text{ if article } i \text{ cites article } j \text{ or vice-versa} \\ 0 \text{ else.} \end{cases} \tag{1}$$

In order to prevent a leaking of the test set into the training set, using the train/test partition defined below, we omitted citations in M from articles in the training set which connect to the test set, but retained citations in the test set which connect to the training set.

We can also define the $N \times L$ category matrix in the standard one-hot fashion

$$C_{jl} = \begin{cases} 1 \text{ if article j is in category } l \\ 0 \text{ else.} \end{cases} \tag{2}$$

Then the co-citation feature matrix is the $N \times L$ matrix product MC. Note that this feature uses only nearest-neighbor citation graph relationships. We could include next-nearest neighbor relationships and so on by calculating $MC + aM^2C + bM^3C + \ldots$ for some constants a and b. In this paper we only used first order connections via MC as the co-citation feature vectors.

.1.2 Training

Using vector embeddings from titles, abstracts, and full-text, and co-citation features as described above, we fed several combinations of these vectors concatenated in the obvious way into the `scikit-learn` SGD classifier `sklearn.linear_model.SGDClassifier`. We used the keyword arguments `loss='log'`, `tol=1e-6`, `max_iter=50`, and `alpha=1e-7` to define the model, which uses 50 epochs, and very small quadratic regularization `alpha` on the weights and biases.

With the features and model defined, we performed a train/test split by shuffling the data in place randomly, and selecting the first $N_{\text{train}} = 1,200,000$ for training. The remaining $N_{\text{test}} = 306,500$ articles were used to evaluate the accuracy of the trained classification, and the model perplexity as reported in table in the main text.

Bibliography

Cecilia Aguerrebere, Mauricio Delbracio, Alberto Bartesaghi, and Guillermo Sapiro. Fundamental limits in multi-image alignment. *IEEE Transactions on Signal Processing*, 64(21):5707–5722, 2016.

Amnon Aharony and Michael E Fisher. Nonlinear scaling fields and corrections to scaling near criticality. *Physical Review B*, 27(7):4394, 1983.

Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. Ntcir-11 math-2 task overview. In *NTCIR*, volume 11, pages 88–98. Citeseer, 2014.

Alexander A. Alemi and Paul Ginsparg. Text segmentation based on semantic word embeddings. *CoRR*, abs/1503.05543, 2015. URL `http://arxiv.org/abs/1503.05543`.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985a.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14): 1530, 1985b.

Bibliography

P W Anderson, BI Halperin, and C M Varma. Anomalous low-temperature thermal properties of glasses and spin glasses. *Philosophical Magazine*, 25(1):1–9, 1972.

Louis-Pierre Arguin, Michael Damron, Charles M Newman, and Daniel L Stein. Uniqueness of ground states for short-range spin glasses in the half-plane. *Communications in Mathematical Physics*, 300(3):641–657, 2010.

Constantin P Bachas. Computer-intractability of the frustration model of a spin glass. *Journal of Physics A: Mathematical and General*, 17(13):L709, 1984.

Donald G Bailey, Andrew Gilman, and Roger Browne. Bias characteristics of bilinear interpolation based registration. In *TENCON 2005 2005 IEEE Region 10*, pages 1–6. IEEE, 2005.

Vijay Balasubramanian. Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions. *Neural computation*, 9(2):349–368, 1997.

Francisco Barahona. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

Mustansir Barma and Michael E. Fisher. Two-dimensional ising-like systems: Corrections to scaling in the klauder and double-gaussian models. *Phys. Rev. B*, 31:5954–5975, May 1985. doi: 10.1103/PhysRevB.31.5954. URL `http://link.aps.org/doi/10.1103/PhysRevB.31.5954`.

Alberto Bartesaghi, Doreen Matthies, Soojay Banerjee, Alan Merk, and Sriram Subramaniam. Structure of β-galactosidase at 3.2-å resolution obtained by cryo-electron microscopy. *Proceedings of the National Academy of Sciences*, 111(32):11709–11714, 2014.

Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline LS Milne, and Sriram Subramaniam. 2.2 å resolution cryo-em structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. URL `https://arxiv.org/abs/1806.01261`.

Matthew Bierbaum, Brian D Leahy, Alexander A Alemi, Itai Cohen, and James P Sethna. Light microscopy at maximal precision. *Physical Review X*, 7(4):041007, 2017.

S Boukraa, AJ Guttmann, S Hassani, I Jensen, JM Maillard, B Nickel, and N Zenine. Experimental mathematics on the magnetic susceptibility of the square lattice ising model. *Journal of Physics A: Mathematical and Theoretical*, 41(45):455202, 2008.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

A. J. Bray and M. A. Moore. Chaotic nature of the spin-glass phase. *Phys. Rev. Lett.*, 58:57–60, Jan 1987a. doi: 10.1103/PhysRevLett.58.57. URL `http://link.aps.org/doi/10.1103/PhysRevLett.58.57`.

AJ Bray and MA Moore. Lower critical dimension of ising spin glasses: a numerical study. *Journal of Physics C: Solid State Physics*, 17(18):L463, 1984.

AJ Bray and MA Moore. Heidelberg colloquium on glassy dynamics. *Lecture Notes in Physics*, 275:121–153, 1987b.

Kevin S Brown and James P Sethna. Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E*, 68(2):021904, 2003.

Pasquale Calabrese, Michele Caselle, Alessio Celi, Andrea Pelissetto, and Ettore Vicari. Non-analyticity of the callan-symanzik β-function of two-dimensional o (n) models. *Journal of Physics A: Mathematical and General*, 33(46):8155, 2000.

John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

Michele Caselle, Martin Hasenbusch, Andrea Pelissetto, and Ettore Vicari. Irrelevant operators in the two-dimensional ising model. *Journal of Physics A: Mathematical and General*, 35(23):4861, 2002.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-CÃľspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. In *In submission to: EMNLP demonstration*, Brussels, Belgium, 2018. URL `https://arxiv.org/abs/1803.11175`. In submission.

K. Chang, A. Eichler, J. Rhensius, L. Lorenzelli, and C. L. Degen. Nanoscale Imaging of Current Density with a Single-Spin Magnetometer. *Nano Letters*, 17(4):2367–2373, 2017. doi: 10.1021/acs.nanolett.6b05304.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013. URL `https://arxiv.org/abs/1312.3005`.

Yan-Jiun Chen, Natalie M Paquette, Benjamin B Machta, and James P Sethna. Universal scaling function for the two-dimensional ising model in an external field: A pragmatic approach. *arXiv preprint arXiv:1307.6899*, 2013.

Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760, 2015.

Colin Clement. Information theoretic clustering. `https://github.com/colinclement/info-clust`, 2016.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Predrag Cvitanović. *Group theory: birdtracks, Lie's, and exceptional groups*. Princeton University Press, 2008.

Andrew M. Dai, Christopher Olah, and Quoc V. Le. Document embedding with paragraph vectors. *CoRR*, abs/1507.07998, 2015. URL `http://arxiv.org/abs/1507.07998`.

Nicolaas Govert De Bruijn. *Asymptotic methods in analysis*, volume 4. Courier Corporation, 1970.

Misganu Debella-Gilo and Andreas Kääb. Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation. *Remote Sensing of Environment*, 115(1):130–142, 2011.

Walter Dempsey, Brandon Oselio, and Alfred Hero. Hierarchical network models for structured exchangeable interaction processes. *arXiv preprint arXiv:1901.09982*, 2019.

Rafael B. Dinner, Kathryn A. Moler, M. R. Beasley, and D. Matthew Feldmann. Enhanced current flow through meandering grain boundaries in y Ba2 Cu3 O7-δ films. *Applied Physics Letters*, 90(21):212501, may 2007. doi: 10.1063/1.2740610. URL `http://aip.scitation.org/doi/10.1063/1.2740610`.

Efi Efrati, Zhe Wang, Amy Kolan, and Leo P Kadanoff. Real-space renormalization in statistical mechanics. *Reviews of Modern Physics*, 86(2):647, 2014.

Steffen Eger, Chao Li, Florian Netzer, and Iryna Gurevych. Predicting research trends from arxiv. *arXiv preprint arXiv:1903.02831*, 2019.

Ismail El Baggari, Benjamin H Savitzky, Alemayehu S Admasu, Jaewook Kim, Sang-Wook Cheong, Robert Hovden, and Lena F Kourkoutis. Nature and evolution of incommensurate charge order in manganites visualized with cryogenic scanning transmission electron microscopy. *Proceedings of the National Academy of Sciences*, 115(7): 1445–1450, 2018.

D. M. Feldmann. Resolution of two-dimensional currents in superconductors from a two-dimensional magnetic field measurement by the method of regularization. *Physical Review B - Condensed Matter and Materials Physics*, 69(14):1–14, 2004. ISSN 01631829. doi: 10.1103/PhysRevB.69.144515.

Arthur E. Ferdinand and Michael E. Fisher. Bounded and inhomogeneous ising models. i. specific-heat anomaly of a finite lattice. *Phys. Rev.*, 185:832–846, Sep 1969. doi: 10.1103/PhysRev.185.832. URL `https://link.aps.org/doi/10.1103/PhysRev.185.832`.

Daniel S Fisher and David A Huse. Ordered phase of short-range ising spin-glasses. *Physical review letters*, 56(15):1601, 1986.

Daniel S Fisher and David A Huse. Equilibrium behavior of the spin-glass ordered phase. *Physical Review B*, 38(1):386, 1988.

Michael E Fisher and Mohit Randeria. Location of renormalization-group fixed points. *Physical review letters*, 56(21):2332, 1986.

Yaotian Fu and Philip W Anderson. Application of statistical mechanics to np-complete problems in combinatorial optimisation. *Journal of Physics A: Mathematical and General*, 19(9):1605, 1986.

Nikolas P Galatsanos and Aggelos K Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Transactions on image processing*, 1(3):322–336, 1992.

Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

David Gottlieb and Chi-Wang Shu. On the gibbs phenomenon and its resolution. *SIAM review*, 39(4):644–668, 1997.

Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018. URL `https://arxiv.org/abs/1705.02801`.

Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup. Efficient subpixel image registration algorithms. *Optics letters*, 33(2):156–158, 2008.

Sidharth Gupta, Konik Kothari, Maarten V de Hoop, and Ivan Dokmanić. Random mesh projectors for inverse problems. *arXiv preprint arXiv:1805.11718*, 2018.

Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):e189, 2007.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017a. URL `https://arxiv.org/abs/1706.02216`.

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. 2017b. URL `https://arxiv.org/abs/1709.05584`.

Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, 34(4):561–580, 1992.

AK Hartmann and MA Moore. Corrections to scaling are large for droplets in two-dimensional spin glasses. *Physical review letters*, 90(12):127201, 2003.

AK Hartmann and MA Moore. Generating droplets in two-dimensional ising spin glasses using matching algorithms. *Physical Review B*, 69(10):104409, 2004.

AK Hartmann and AP Young. Large-scale low-energy excitations in the two-dimensional ising spin glass. *Physical Review B*, 66(9):094419, 2002.

Alexander K Hartmann. Ground states of two-dimensional ising spin glasses: fast algorithms, recent developments and a ferromagnet-spin glass mixture. *Journal of Statistical Physics*, 144(3):519–540, 2011.

John Hertz, Anders Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

David A Huse and Daniel S Fisher. Dynamics of droplet fluctuations in pure and random ising systems. *Physical Review B*, 35(13):6841, 1987.

Brian F Hutton and Michael Braun. Software for image registration: algorithms, accuracy, efficacy. In *Seminars in nuclear medicine*, volume 33, pages 180–192. Elsevier, 2003.

Jordi Inglada, Vincent Muron, Damien Pichard, and Thomas Feuvrier. Analysis of artifacts in subpixel remote sensing image registration. *IEEE transactions on Geoscience and Remote Sensing*, 45(1):254–264, 2007.

Giovanni Jacovitti and Gaetano Scarano. Discrete time techniques for time delay estimation. *IEEE Transactions on signal processing*, 41(2):525–533, 1993.

Edwin Thompson Jaynes. How does the brain do plausible reasoning? In *Maximum-entropy and Bayesian methods in science and engineering*, pages 1–24. Springer, 1988.

LP Kadanoff. Scaling laws for ising models near critical points. In *pp 24-36 of Proceedings of the 1966 Midwest Conference on Theoretical Physics, Indiana University, May 20–21, 1966. Lichtenberg, DB (ed.). Bloomington, Ind., Indiana University, 1966.* Univ. of Illinois, Urbana, 1967.

Beena Kalisky, Eric M Spanton, Hilary Noad, John R Kirtley, Katja C Nowack, Christopher Bell, Hiroki K Sato, Masayuki Hosoda, Yanwu Xie, Yasuyuki Hikita, et al.

Locally enhanced conductivity due to the tetragonal domain structure in laalo3/srtio3 heterointerfaces. *Nature materials*, 12(12):1091–1095, 2013.

Adam D Kammers and Samantha Daly. Digital image correlation under scanning electron microscopy: methodology and validation. *Experimental Mechanics*, 53(9):1743–1761, 2013.

Naoki Kawashima. Fractal droplets in two-dimensional spin glass. *Journal of the Physical Society of Japan*, 69(4):987–990, 2000.

Naoki Kawashima and Takayuki Aoki. Zero-temperature critical phenomena in two-dimensional spin glasses. *arXiv preprint cond-mat/9911120*, 1999.

Scott Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, 34(5-6):975–986, 1984.

Mark J H Ku, Tony X Zhou, Qing Li, Young J Shin, Jing K Shi, Claire Burch, Huiliang Zhang, Francesco Casola, Takashi Taniguchi, Kenji Watanabe, Philip Kim, Amir Yacoby, and Ronald L Walsworth. Imaging Viscous Flow of the Dirac Fluid in Graphene Us-ing a Quantum Spin Magnetometer. *arXiv:1905.10791*, 2019. URL `https://arxiv.org/pdf/1905.10791.pdf`.

Carl Lagoze and Herbert Van de Sompel. The open archives initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 54–62. ACM, 2001.

Edward D Lee, Chase P Broedersz, and William Bialek. Statistical mechanics of the us supreme court. *Journal of Statistical Physics*, 160(2):275–301, 2015.

Michael E Leventon and W Eric L Grimson. Multi-modal volume registration using joint intensity distributions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1057–1066. Springer, 1998.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Wenyuan Liu, Stanisław Saganowski, Przemysław Kazienko, and Siew Ann Cheong. Using machine learning to predict the evolution of physics research. *arXiv preprint arXiv:1810.12116*, 2018.

Michal Lopuszynski and Lukasz Bolikowski. Tagging scientific publications using wikipedia and natural language processing tools. comparison on the arxiv dataset. *CoRR*, abs/1309.0326, 2013. URL `http://arxiv.org/abs/1309.0326`.

Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342 (6158):604–607, 2013.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

David JC MacKay. Hyperparameters: Optimize, or integrate out? In *Maximum entropy and bayesian methods*, pages 43–59. Springer, 1996.

David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

Barry M McCoy and Tai Tsun Wu. *The two-dimensional Ising model*. Courier Corporation, 2014.

W. L. McMillan. Domain-wall renormalization-group study of the three-dimensional random ising model. *Phys. Rev. B*, 30:476–477, Jul 1984. doi: 10.1103/PhysRevB.30.476. URL `http://link.aps.org/doi/10.1103/PhysRevB.30.476`.

Alexander Y Meltzer, Eitan Levin, and Eli Zeldov. Direct reconstruction of two-dimensional currents in thin films from magnetic-field measurements. *Physical Review Applied*, 8:064030, 2017.

Marc Mézard, Giorgio Parisi, Nicolas Sourlas, Gérard Toulouse, and Miguel Virasoro. Replica symmetry breaking and the nature of the spin glass phase. *Journal de Physique*, 45(5):843–854, 1984.

Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

A Alan Middleton. Computational complexity of determining the barriers to interface motion in random systems. *Physical Review E*, 59(3):2571, 1999.

A Alan Middleton. Energetics and geometry of excitations in random systems. *Physical Review B*, 63(6):060202, 2001.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Ch M Newman and DL Stein. Multiple states and thermodynamic limits in short-ranged ising spin-glass models. *Physical Review B*, 46(2):973, 1992.

Ch M Newman and DL Stein. Non-mean-field behavior of realistic spin glasses. *Physical review letters*, 76(3):515, 1996.

CM Newman and DL Stein. Simplicity of state and overlap structure in finite-volume realistic spin glasses. *Physical Review E*, 57(2):1356, 1998.

David Nicholson and Alberto Vecchio. Bayesian bounds on parameter estimation accuracy for compact coalescing binary gravitational wave signals. *Physical Review D*, 57(8): 4588, 1998.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1): 11–33, 2016. URL `https://arxiv.org/abs/1503.00759`.

B Nienhuis. Analytical calculation of two leading exponents of the dilute potts model. *Journal of Physics A: Mathematical and General*, 15(1):199, 1982.

Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press, 2001.

Katja C Nowack, Eric M Spanton, Matthias Baenninger, Markus König, John R Kirtley, Beena Kalisky, Christopher Ames, Philipp Leubner, Christoph Brüne, Hartmut Buhmann, et al. Imaging currents in hgte quantum wells in the quantum spin hall regime. *Nature materials*, 12(9):787–791, 2013.

WP Orrick, B Nickel, AJ Guttmann, and JHH Perk. The susceptibility of the square lattice ising model: new developments. *Journal of Statistical Physics*, 102(3):795–841, 2001.

Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

A. E. Pashitski, A. Gurevich, A. A. Polyanskii, D. C. Larbalestier, A. Goyal, E. D. Specht, D. M. Kroeger, J. A. DeLuca, and J. E. Tkaczyk. Reconstruction of current flow and imaging of current- limiting defects in polycrystalline superconducting films. *Science*, 275(5298):367–369, jan 1997. doi: 10.1126/science.275.5298.367. URL `http://www.ncbi.nlm.nih.gov/pubmed/8994028`.

Thomas C Pekin, Christoph Gammer, Jim Ciston, Andrew M Minor, and Colin Ophus. Optimizing disk registration algorithms for nanobeam electron diffraction strain mapping. *Ultramicroscopy*, 176:170–176, 2017.

Tuan Q Pham, Marijn Bezuijen, Lucas J Van Vliet, CL Luengo Hendriks, and K Schutte. Performance of optimal registration estimators. *Proceedings of SPIE, 2005 vol. 5817*, 2005.

W A Phillips. Two-level states in glasses. *Reports on Progress in Physics*, 50(12):1657, 1987. URL `http://stacks.iop.org/0034-4885/50/i=12/a=003`.

William H Press, Brian P Flannery, Saul A Teukolsky, William T Vetterling, et al. *Numerical recipes*, volume 3. cambridge University Press, cambridge, 1989.

William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing.* Cambridge university press, 2007.

Archishman Raju and James P Sethna. Reexamining the renormalization group: Period doubling onset of chaos. *arXiv preprint arXiv:1807.09517*, 2018.

Archishman Raju, Colin B. Clement, Lorien X. Hayden, Jaron P. Kent-Dobias, Danilo B. Liarte, D. Zeb Rocklin, and James P. Sethna. Normal form for renormalization groups. *Phys. Rev. X*, 9:021014, Apr 2019. doi: 10.1103/PhysRevX.9.021014. URL `https://link.aps.org/doi/10.1103/PhysRevX.9.021014`.

Dirk Robinson and Peyman Milanfar. Fundamental performance limits in image registration. *IEEE Transactions on Image Processing*, 13(9):1185–1199, 2004.

Gustavo K Rohde, Akram Aldroubi, and Dennis M Healy. Interpolation artifacts in sub-pixel image registration. *IEEE transactions on image processing*, 18(2):333–345, 2009.

Bradley J Roth, Nestor G Sepulveda, and John P Wikswo Jr. Using a magnetometer to image a two-dimensional current distribution. *Journal of applied physics*, 65(1): 361–372, 1989.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jesús Salas. Exact finite-size-scaling corrections to the critical two-dimensional ising model on a torus. *Journal of Physics A: Mathematical and General*, 34(7):1311, 2001.

Benjamin H. Savitzky, Ismail El Baggari, Colin B. Clement, Emily Waite, Berit H. Goodge, David J. Baek, John P. Sheckelton, Christopher Pasco, Hari Nair, Nathaniel J. Schreiber, Jason Hoffman, Alemayehu S. Admasu, Jaewook Kim, Sang-Wook Cheong, Anand Bhattacharya, Darrell G. Schlom, Tyrel M. McQueen, Robert Hovden, and Lena F. Kourkoutis. Image registration of low signal-to-noise cryo-stem data. *Ultramicroscopy*, 191:56 – 65, 2018. ISSN 0304-3991. doi: https://doi.org/10.1016/j.ultramic.2018.04.008. URL `http://www.sciencedirect.com/science/article/pii/S0304399117304369`.

Hubert W Schreier, Joachim R Braasch, and Michael A Sutton. Systematic errors in digital image correlation caused by intensity interpolation. *Optical engineering*, 39(11): 2915–2922, 2000.

James P. Sethna. Why is science possible?, 2015. URL `http://sethna.lassp.cornell.edu/research/why_is_science_possible`.

Karthik Shekhar, Mehran Kardar, and Arup K Chakraborty. Faithful models of viral fitness can be inferred from mutation patterns in viral dna sequences sampled from a population. *Biophysical Journal*, 104(2):495a–496a, 2013a.

Karthik Shekhar, Claire F. Ruberman, Andrew L. Ferguson, John P. Barton, Mehran Kardar, and Arup K. Chakraborty. Spin models inferred from patient-derived viral sequence data faithfully describe hiv fitness landscapes. *Phys. Rev. E*, 88:062705, Dec 2013b. doi: 10.1103/PhysRevE.88.062705. URL `http://link.aps.org/doi/10.1103/PhysRevE.88.062705`.

Noam Slonim, Gurinder Singh Atwal, Gašper Tkačik, and William Bialek. Information-based clustering. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18297–18302, 2005.

Jascha Sohl-Dickstein, Peter B Battaglino, and Michael R DeWeese. New method for parameter estimation in probabilistic models: minimum probability flow. *Physical review letters*, 107(22):220601, 2011.

DL Stein. A model of protein conformational substates. *Proceedings of the National Academy of Sciences*, 82(11):3670–3672, 1985.

Lovro Šubelj, Dalibor Fiala, and Marko Bajec. Network-based statistical comparison of citation topology of bibliographic databases. *Scientific reports*, 4:6496, 2014.

Jean-Philippe Tetienne, Nikolai Dontschuk, David A. Broadway, Alastair Stacey, David A. Simpson, and Lloyd C. L. Hollenberg. Quantum imaging of current flow in graphene. *Science Advances*, 3(4):e1602429, 2017. doi: 10.1126/sciadv.1602429. URL `http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1602429`.

Creighton K Thomas and A Alan Middleton. Exact algorithm for sampling the two-dimensional ising spin glass. *Physical Review E*, 80(4):046708, 2009.

Mark K. Transtrum and Peng Qiu. Model reduction by manifold boundaries. *Phys. Rev. Lett.*, 113:098701, Aug 2014. doi: 10.1103/PhysRevLett.113.098701. URL `http://link.aps.org/doi/10.1103/PhysRevLett.113.098701`.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.*, 104:060201, 2010.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys. Rev. E*, 83: 036701, 2011. doi: http://dx.doi.org/10.1103/PhysRevE.83.036701.

Mark K. Transtrum, Benjamin Machta, Kevin Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Sloppiness and emergent theories in physics, biology, and beyond. *Journal of Chemical Physics*, 143, 07/2015 2015. doi: 10.1063/1.4923066. URL `http://arxiv.org/abs/1501.07668v1`.

David W. Tyler. Intrinsic bias in fisher information calculations for multi-mode image registration. *Opt. Lett.*, 43(10):2292–2295, May 2018. doi: 10.1364/OL.43.002292. URL `http://ol.osa.org/abstract.cfm?URI=ol-43-10-2292`.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *arXiv preprint arXiv:1711.10925*, 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

Mikhail L Uss, Benoit Vozel, Vitaliy A Dushepa, Vladimir A Komjak, and Kacem Chehdi. A precise lower bound on image subpixel registration accuracy. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3333–3345, 2014.

Denis Vasyukov, Yonathan Anahory, Lior Embon, Dorri Halbertal, Jo Cuppens, Lior Neeman, Amit Finkler, Yehonathan Segev, Yuri Myasoedov, Michael L Rappaport, Martin E Huber, and Eli Zeldov. A scanning superconducting quantum interference device with single electron spin sensitivity. *Nature Nanotechnology*, 8(9):639–644, 2013. doi: 10.1038/nnano.2013.169.

Felipe Vico, Leslie Greengard, and Miguel Ferrando. Fast convolution with free-space green's functions. *Journal of Computational Physics*, 323:191 – 203, 2016. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2016.07.028. URL `http://www.sciencedirect.com/science/article/pii/S0021999116303230`.

Curtis R Vogel and Mary E Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE transactions on image processing*, 7(6):813–824, 1998.

Joshua J Waterfall, Fergal P Casey, Ryan N Gutenkunst, Kevin S Brown, Christopher R Myers, Piet W Brouwer, Veit Elser, and James P Sethna. Sloppy-model universality class and the vandermonde matrix. *Physical review letters*, 97(15):150601, 2006.

FJ Wegner. Some invariance properties of the renormalization group. *Journal of Physics C: Solid State Physics*, 7(12):2098, 1974.

Franz J Wegner. Corrections to scaling laws. *Physical Review B*, 5(11):4529, 1972.

Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

Rinke J Wijngaarden, HJW Spoelder, R Surdeanu, and R Griessen. Determination of two-dimensional current patterns in flat superconductors from magneto-optical measurements: An efficient inversion scheme. *Physical Review B*, 54(9):6742, 1996.

Rinke J Wijngaarden, K Heeck, HJW Spoelder, R Surdeanu, and R Griessen. Fast determination of 2d current patterns in flat conductors from measurement of their magnetic field. *Physica C: Superconductivity*, 295(3):177–185, 1998.

Kenneth G. Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Phys. Rev. B*, 4:3174–3183, Nov 1971. doi: 10.1103/PhysRevB.4.3174. URL `http://link.aps.org/doi/10.1103/PhysRevB.4.3174`.

Ryan W Wolcott and Ryan M Eustice. Visual localization within lidar maps for automated urban driving. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 176–183. IEEE, 2014.

Min Xu, Hao Chen, and Pramod K Varshney. Ziv–zakai bounds on image registration. *IEEE Transactions on Signal Processing*, 57(5):1745–1755, 2009.

Andrew B Yankovich, Benjamin Berkels, Wolfgang Dahmen, Peter Binev, Sergio I Sanchez, Steven A Bradley, Ao Li, Izabela Szlufarska, and Paul M Voyles. Picometre-precision analysis of scanning transmission electron microscopy images of platinum nanocatalysts. *Nature communications*, 5:4155, 2014.

Imam Samil Yetik and Arye Nehorai. Performance bounds on image registration. *IEEE Transactions on Signal Processing*, 54(5):1737–1749, 2006.

Pei Yu and AYT Leung. The simplest normal form and its application to bifurcation control. *Chaos, Solitons & Fractals*, 33(3):845–863, 2007.

Daliang Zhang, Yihan Zhu, Lingmei Liu, Xiangrong Ying, Chia-En Hsiung, Rachid Sougrat, Kun Li, and Yu Han. Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials. *Science*, 359(6376):675–679, 2018.

Yihan Zhu, Jim Ciston, Bin Zheng, Xiaohe Miao, Cory Czarnik, Yichang Pan, Rachid Sougrat, Zhiping Lai, Chia-En Hsiung, Kexin Yao, et al. Unravelling surface and interfacial structures of a metal–organic framework by transmission electron microscopy. *Nature materials*, 16(5):532, 2017.

Royce KP Zia, Edward F Redish, and Susan R McKay. Making sense of the legendre transform. *American Journal of Physics*, 77(7):614–622, 2009.

Jacob Ziv and Moshe Zakai. Some lower bounds on signal parameter estimation. *IEEE transactions on Information Theory*, 15(3):386–391, 1969.

Lilla Zöllei, John W Fisher III, and William M Wells III. A unified statistical and information theoretic framework for multi-modal image registration. In *IPMI*, pages 366–377. Springer, 2003.