Abstract

Sampling from the Greedy Mixture Posterior

Dylan Potter O'Connell

2021

Mixtures of distributions provide a flexible model for heterogeneous data, but this versatility is concomitant with computational difficulty. We study the task of generating samples from the "greedy" Gaussian mixture posterior. While it is widely known that Gibbs sampling can be slow to converge, concrete results quantifying this behavior are scarce. In this dissertation, we establish conditions under which the number of steps required by a Gibbs sampler is exponential in the separation of the data clusters.

Further, we analyze the efficacy of potential solutions. The simulated tempering algorithm uses an auxiliary temperature variable to flatten the target density (reducing the effective cluster separation). As existing implementations are poorly suited to the unusual properties of the mixture posterior, we adapt simulated tempering by flattening the individual likelihood components (referred to as *internal annealing*). However, this is no universal solution, and we characterize conditions under which the original cause of slow convergence will persist. An alluring alternative is *subsample annealing*, which instead flattens the posterior by reducing the size of the observed subsample. Still, this approach is sensitive to the selection of the data, and we prove that a single poorly chosen datum can be sufficient to preclude rapid convergence.

Sampling from the Greedy Mixture Posterior

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Dylan Potter O'Connell

Dissertation Director: Dr. Andrew R. Barron

June, 2021

Contents

List of Figures

Acknowledgments

I am deeply indebted to my advisor, Professor Andrew R. Barron, for his guidance and endless patience. Above all, he has taught me the virtues of diligence and perseverance, and their vital role in tackling any challenge which is difficult enough to truly matter. I will never forget these five rewarding years with the Yale Department of Statistics & Data Science—I am enormously grateful to all the professors who have have shaped my academic journey, and to the staff whose tireless support made it possible.

One of my life's great privileges has been the generosity of my many teachers—both in my time at Yale, and along the winding path that brought me here. In high school, Mary Kate Bluestein, Catherine Brewster, Mark Limperis, and others offered me uncommon attention and support when I must have appeared an unlikely target. In college, Rob Manning, Joshua Sabloff, and the rest of the Haverford Mathematics Department sparked my curiosity and forever rewired the workings of my brain. In my senior year, Bill Huber gave me my first tantalizing glimpse of the world of statistics through unplanned late night conversations after Problem Solving Group. I am lucky to be in the debt of many more teachers than I could list here.

My greatest blessing is my friends and family, who have given these years texture and meaning. Above all, I am grateful for my parents, Mark and Alison. In every sense, they are the source of my fortune and privilege—I am who I am because of them.

Chapter 1

Introduction

1.1 Mixture Models

Mixtures of distributions are an invaluable tool for bridging the fundamental divide between idealized statistical models and the heterogeneous world of data that lies beyond the classroom door. As an explicit model, mixtures are necessary to describe the generation of data from a finite set of distinct sources, but they are equally vital as a tool for approximation. To quote the famous aphorism of George Box [1], "All models are wrong, but some are useful." Even when frequently "wrong", the remarkable flexibility of mixtures allows for the approximate characterization of heterogeneity in observed data. From unsupervised cluster analysis to density estimation, mixture models have been successfully applied in a dizzying array of fields, spanning the alphabet from astronomy (e.g. clustering the famed galaxy dataset [2] as a computational benchmark) to zoology (e.g. Karl Pearson's [3] 19th century analysis of the ratio between the forehead and body length of shore crabs). Above all, mixtures enable the use of distributions whose theoretical properties have been deeply studied (like the Gaussian) as a building block in the modeling of complex systems.

However, the endless versatility of mixtures presents a Faustian bargain—this flexibility is concomitant with significant computational cost. The evolution of mixture models over

the past century has been inextricably linked to the computational developments that govern their use. Perhaps the seminal modern advance in mixture computation was the work of Dempster et al. [4], whose Expectation Maximization algorithm estimates the maximum likelihood using the latent variable framework. The fundamental insight is the use of the unobserved (i.e. latent) variables denoting the original "source" of each datum, which we refer to as a *label*. While the likelihood of the mixture on the *observed data* may pose a significant computational challenge, the distribution of the *complete data* (which includes both the observed data and unobserved labels) implies conditional distributions that are easy to manage. In this dissertation, we will study the Gibbs sampling algorithm, which is inspired by the same premise.

Our particular interest lies in the Bayesian setting—given observed data generated from a Gaussian mixture, we wish to draw inferences about the underlying mixture component parameters. This is typically accomplished by generating samples from the posterior distribution. The use of a Gaussian prior on the mixture component centers results in a conjugate posterior that is also a mixture of Gaussians. However, this posterior is a mixture over the *exponential* count of possible labels (where each label describes a potential assignment of data to likelihood mixture components). We will focus on the "greedy" setting, where the likelihood is a mixture between a *variable* Gaussian component (whose center parameter is the target of inference), and a pre-defined *fixed* component density (with no variable parameter). We will discuss this model in detail, but in short, it represents a single step in the iterative process of fitting additional mixture components. As this greedy construction results in a posterior that shares the same Gaussian mixture form, it can broadly be viewed as the simplest model that captures the fundamental underlying computational challenge.

For much of the 20th century, Bayesian inference was computationally infeasible for all but the simplest of models. The field was revolutionized by the advent of powerful Markov Chain Monte Carlo (MCMC) techniques, which vastly increased the range of viable applications. It is typically straightforward to construct a Markov chain with the key theoretical

guarantee that it will eventually converge to the correct stationary distribution. However, this elides the critical question of *how long that process will take*. While there are myriad potential MCMC implementations, as a simplified introduction, we can assume that they typically follow a transition mechanism that relies on *local* information. The most intuitive might be the Metropolis-Hastings Random Walk (MHRW), but other popular techniques incorporate the gradient for guidance (such as the Metropolis Adjusted Langevin Algorithm, popularized by Roberts et al. [5], or Hamiltonian Monte Carlo, typically attributed to Duane et al. [6]).

However, in multimodal settings, local information fails to provide global guidance, and convergence may be problematically slow. The Gaussian mixture density is a canonical example of multimodality—its surface is characterized by individually unimodal regions separated by deep, low-density valleys that locally-based transition mechanisms struggle to traverse. The use of more ambitious transition rules (which can push past a low-density valley to reach the next high-density region) will typically struggle in high dimensions, as such blind exploration is unlikely to stumble upon the regions of interest.

The unifying goal of this dissertation is to characterize and better understand the computational challenge of generating posterior samples through MCMC. We mirror the literature and use the language of "mixing" to describe the convergence of a Markov chain—thus, the difficulty lies in ensuring sufficient "flow" between isolated regions, or else a "bottleneck" will occur and "mixing" will be slow. Formally, the number of steps required until a Markov chain is sufficiently close to its stationary distribution to generate samples is called the *mixing time*. In particular, we wish to draw the critical distinction between *rapid* mixing, which grows polynomially in the specified input parameters, and *slow* mixing, which grows exponentially in the specified input parameters. This fundamental divide is common in the analysis of computational tractability, due to the expectation that Moore's Law will eventually render lower order factors inconsequential.

The canonical MCMC technique for mixture posteriors is the *Gibbs sampler*, popularized

by Diebolt & Robert [7] in 1994 (with important precursors including the publications by Geman & Geman [8] and Tanner & Wong [9]). The Gibbs sampler follows the same fundamental insight as the Expectation Maximization algorithm, and it constructs a Markov chain using alternating conditional draws (generating the parameters given an estimate of the latent variables, and generating the latent variables given an estimate of the parameters). While the Gibbs sampler provides a powerful tool for sampling from the mixture posterior, it is an inherently *local* process, and it faces the same familiar computational concerns. This intersection between the multimodality of the mixture and the locality of MCMC techniques poses the fundamental challenge that this dissertation will confront.

1.2 Dissertation Summary

In the remainder of Chapter 1, we establish the computational task: generating samples from the "greedy" Gaussian mixture posterior. The prior literature that is specialized to mixture posteriors has primarily addressed other concerns, and there is a relative paucity of results characterizing this computational challenge (Section 1.3). As the design and implementation of successful sampling techniques hinges on our understanding of the underlying impediments to mixing, this is a key gap in the literature.

The "greedy" form of the Gaussian mixture posterior (Section 1.4) provides a particularly appealing target for our research. First, it sidesteps the issue of "label switching" from non-identifiable components (discussed in Section 1.3), which has drawn significant attention (but is not of direct interest to us). Second, it is arguably the simplest model that still captures the fundamental structure of the mixture posterior (and its exponential component count). Third, there are strong previous results demonstrating that an iterative greedy approach can estimate complex models with high accuracy (such as the work of Barron & Li [10]). While these results are articulated in terms of estimation, there are natural and clear parallels to our task of sampling. Finally, despite these advantages, it has received little attention in the

existing literature.

Gibbs sampling is the canonical MCMC technique for the mixture posterior, and it constructs a time homogeneous Markov chain by leveraging the latent variable structure (Section 1.5). In this dissertation, we study the Markov chain generated by the *collapsed* Gibbs sampler, which takes advantage of the available closed form solution to operate directly on the discrete state space of the latent posterior labels (by integrating out the step that generates an intermediate component parameter). When the closed form is available, "collapsing" the Gibbs sampler is usually computationally beneficial, but in this particular setting it offers two powerful advantages. First, the discrete state space of the labels will facilitate the clean conductance arguments used to prove our mixing bounds. Second, we will discuss (in Section 3.2) how the theoretical analysis of Markov chains on mixture density targets hinges on the transfer of information between mixture components—thus, it is sensible to explicitly define our Markov chain so that it models these transfers.

The fact that the Gibbs sampler may mix slowly on the multimodal mixture posterior is common folklore, but literature quantifying this behavior is relatively scarce. In Chapter 2, we identify conditions on the data that prevent the Markov chain from mixing rapidly. We use a conductance argument (Section 2.1) to show that the existence of a label with a small probability of escape implies a lower bound on the mixing time (defined as the number of steps until the chain reaches a fixed total variation distance from the posterior distribution). Using this technique, Theorem 2.2.1 (Section 2.2) formalizes conditions under which a sufficiently isolated data cluster causes the mixing time to grow at a rate that is exponential in the two cluster isolation parameters—u (the distance between cluster centers) and Δ (the minimum distance from any datum outside the cluster to the cluster center). Supplemental evidence from empirical simulations suggests that despite the restrictions required for the proof, this result is illustrative of a broader relationship between cluster isolation and mixing time. While this result is fairly intuitive, there is value in quantifying the behavior. In particular, the exponential relationship between cluster isolation and mixing time is suggestive of a

potential solution—the use of algorithmic techniques that implicitly *reduce* the effective cluster isolation. This insight helps to motivate our study of "annealing" methods in the remainder of the dissertation.

In Chapter 3, we analyze simulated tempering (a natural MCMC implementation of the annealing framework), and its specialization to the greedy mixture posterior. As a simplified preview of the premise, *annealing* introduces an auxiliary temperature variable that progressively flattens (i.e. "anneals") the original posterior density—when the temperature is cold, the annealed posterior equals the original target, and when the temperature is hot, the flattened density exhibits rapid mixing (Section 3.1). The simulated tempering algorithm creates a chain on the joint space of the original target and the auxiliary temperature variable, paving new paths that circumvent the original barriers to mixing (at high temperatures, it is easy to transfer between previously separated regions).

This raises the critical question of *how* to anneal the target density, and it is useful to distinguish between two potential domains for mixture sampling. Our interest lies in the task of generating samples from a *posterior* mixture, arising from observed data and a known model. In contrast, a common task is to generate samples solely using oracle queries from an otherwise opaque mixture density, which we refer to as the *generic* mixture setting. The preexisting analyses for simulated tempering on mixtures have typically focused on *generic* mixtures, and this setting limits the available methods for annealing the density to the canonical choice—the direct exponentiation of the target density using the inverse temperature. However, this direct exponentiation proves problematic for the mixture posterior, as the assumptions which could plausibly control its behavior in the generic mixture setting are untenable, and the exponentiation erases the valuable latent variable structure (Section 3.2).

The limited preexisting literature that studies annealing in the specific context of the mixture posterior has typically mirrored this use of direct exponentiation. However, in the mixture posterior setting, we need *not* be restricted to oracle queries, and we enjoy greater

optionality in our choice of annealing implementation.

We introduce the technique of *internal annealing*, which instead flattens the posterior components individually, thus preserving the mixture structure at all temperatures (Section 3.3). This approach offers a variety of computational advantages, and critically, it facilitates theoretical mixing analysis—the simulated tempering chain can again operate directly on the discrete state space of the labels, enabling a familiar conductance argument to bound the mixing time (Section 3.4). Specifically, we recall that Theorem 2.2.1 identified conditions that cause a mixing bottleneck for the collapsed Gibbs sampler, suggesting the applicability of simulated tempering. While this *can* be effective, it is no panacea, and Theorem 3.4.4 establishes further conditions under which the mixing bottleneck will persist (despite the use of simulated tempering).

In Chapter 4, we explore the advantages and potential pitfalls of an alternative implementation of the annealing framework. Originally (in Chapter 3), we flattened the posterior through the classical choice of an auxiliary "temperature" variable, but the annealing premise can be applied through any technique which transforms the difficult target density into a rapidly mixing one. In the Bayesian setting, a natural method to connect the prior (which is rapidly mixing) and the posterior is to control the size of the observed subsample (Section 4.1), and this *subsample annealing* offers a promising alternative to the standard temperature-based approach. The use of subsample annealing can be independently motivated by its clear computational benefits (as the complexity of queries scales with the sample size, which may be large), and thus its theoretical mixing properties are of particular interest. However, because the state space of the posterior labels varies with the subsample size, we cannot directly use it to define a simulated tempering chain under the collapsed Gibbs sampler. We solve this with the introduction of *fractional annealing* (Section 4.2), which individually controls the contribution of each datum. Here, we use fractional annealing as a method to implement subsample annealing, but we note its broader potential as a flexible framework for creating specialized annealing schedules (it contains internal annealing and

subsample annealing as specific examples).

While subsample annealing does not exhibit the *same* particular bottleneck that causes slow mixing under temperature annealing, it is highly sensitive to the composition of the subsamples (Section 4.3). Theorem 4.3.2 establishes a set of conditions under which the removal of a single datum causes such a large shift in the posterior that the original bottleneck (under the collapsed Gibbs sampler) must persist in the full simulated tempering chain. Given this sensitivity to subsample composition, we propose *tempered transitions* as a natural target for further study, as it allows for regular changes to the annealing schedule (Section 4.4). In the appendices, we include relevant extensions of this work that are referenced in the text, as well as further details regarding the implementation of the empirical simulations.

1.2.1 Notation and Structure

Throughout the dissertation, we use bold letters to refer to collections across multiple data indices (thus x_i refers to a d-dimensional datum, and $\mathbf{x}$ the dataset with sample size N). Let negative indexing omit the index from a collection (thus $\mathbf{x}_{-i}$ refers to the data $\mathbf{x}$ with the datum x_i removed). Let I denote the identity matrix, let $\mathbb{P}(\cdot)$ mark the probability of a specific event, and let $\tilde{p}(\cdot)$ refer to an unnormalized form of a density $p(\cdot)$. We write $\mathcal{N}(\,\cdot\,;\theta,\Sigma)$ to denote the multivariate Gaussian density with mean vector θ and covariance matrix Σ. When we wish to describe a generic Markov chain (which does not reflect our specific posterior setting), we use $y \in \mathcal{Y}$ as the state space. Typically, capital letters denote sets—thus, for element y, set Y, and state space $\mathcal{Y}$, we have $y \in Y \subset \mathcal{Y}$.

The proofs for all theorems and lemmas are relegated to a separate section at the end of each chapter. For convenient reference, a supplementary index of important terms can be found in Appendix A.

1.3 Prior Literature

Before we explicitly introduce our chosen model (in Section 1.4), it is valuable to sketch its context within the existing literature. As a preview, the important takeaway from this section is simply that the computational challenge of generating samples from the Bayesian mixture posterior is relatively underexplored. The research that is specialized to this setting is largely directed towards other concerns, while the research which *does* share our task of interest is typically not specialized to this setting. This dissertation will draw inspiration from techniques developed in related domains, and will adapt these methods to the specific structure of the mixture posterior.

The Gibbs sampler is straightforward to implement, and it generates a Markov chain whose stationary distribution matches the Bayesian mixture posterior. Under light assumptions, the chain must converge to the correct distribution, but we lack guarantees on its rate of convergence. The focus of this dissertation is this challenge of *computation*, rather than the myriad concerns that arise in model construction and inference. In particular, as mentioned earlier, we wish to distinguish between rates of convergence that imply *rapid* (i.e. polynomial time) and *slow* (i.e. exponential time) mixing.

The critical issue of the convergence rate has been studied in a variety of different Gibbs sampling applications. Common techniques for proving *upper bounds* include coupling arguments (e.g. image restoration, by Gibbs [11]), or minorization & drift conditions (e.g. hierarchical Poisson models, by Rosenthal [12]). To prove a *lower bound*, a common style of analysis popularized by Madras & Randall [13] uses a state space partition to capture the multimodality causing the mixing bottleneck (e.g. genomic discovery, by Woodard & Rosenthal [14]). While originally framed for a general Markov chain, this state space partition provides the foundation for the simulated tempering analysis of Woodard et al. [15], which will prove influential in our study of temperature annealing in Chapter 3 (we will wait until that chapter to introduce the sources relevant to annealing).

There is a wealth of literature discussing the usage of Bayesian mixture models. The

monograph text by Frühwirth-Schnatter [16] provides a comprehensive foundation, and alternatives include the McLachlan textbook [17] or an abbreviated introduction by Marin et al. [18]. These resources address both the myriad choices in model construction, and the challenges which arise in estimation and inference. By comparison, the focus of this dissertation is relatively narrow—we wish to characterize the computational challenge of sampling from a given model, whereas the broader Bayesian mixture literature has primarily studied other adjacent topics.

The first concern that has drawn significant attention is the issue of *label switching* (e.g. Celeux et al. [19] and Stephens [20]). As a simplified summary, when our inference targets are the parameters from K exchangeable mixture components, the posterior parameter space has $K!$ symmetric regions (for the $K!$ equivalent permutations of the data labels that lead to the same index partition). This is primarily a problem for certain forms of inference (e.g. a naive posterior expectation is foiled by this symmetry), but it also has practical implications for running the Markov chain. Full exploration of the posterior space is arduous, and while it is only necessary to explore a single symmetric region, restricting the chain may be difficult in practice. Potential solutions tackle different aspects of the problem, and examples include artificial identifiability constraints, deterministic relabeling strategies, or permutation invariant loss functions. This is a key motivation for our choice of the greedy setting—with only a single variable mixture component, there is no issue of identifiability. As our concerns are strictly computational, the greedy model is a natural way to narrow our focus to the fundamental underlying challenge.

A second strand of literature studies the challenge of mixture models with an *unknown* number of mixture components (e.g. Richardson & Green [21] and Stephens [22]). The greedy framework does provide a natural way to address this issue (as we control the number of greedy steps), but again the focus of this dissertation is computational, and we will not address the choices and concerns of model construction in detail. The Frühwirth-Schnatter's monograph [16] provides an accessible introduction to both of these challenges.

In contrast, there is a notable paucity of theoretical guarantees on the computation time for sampling from mixture posteriors, outside of cases with restrictive assumptions. Mou et al. [23] use clever analysis to prove a polynomial time bound for power posteriors, but their sampling technique is specialized to the *symmetric* two-component case. Likely the most relevant prior work is that of Tosh & Dasgupta [24], who use a conductance argument to prove exponentially slow mixing for two specific arrangements of data clusters. While their underlying model and ultimate goals differ from ours, their conductance argument inspires the strategy we use for our own mixing lower bound (Theorem 2.2.1), and thus subsequent sections will discuss their work in greater detail.

A distinct but highly relevant task is the generation of samples from *generic* mixtures, where we use "generic" to denote a setting where we are restricted to oracle value and gradient queries of the mixture density (and lack other information about the components). The absence of any latent variable structure precludes the use of Gibbs sampling, but analyzing the mixing behavior of other MCMC techniques is insightful. This literature is highly relevant to our work (particularly the twin perspectives on simulated tempering offered by Ge et al. [25] and Woodard et al. [15]), and it will be discussed at length in Chapter 3.

More broadly, it is instructive to consider the task of sampling from general non-log-concave density targets (which need not take the form of a mixture). There is a wealth of literature on the use of discretized SDEs (e.g. Langevin diffusion), but as expected, these approaches tend to mix slowly when the target is non-log-concave. One technique of note is to contain the region of non-log-concavity within a ball of radius R, in which case light regularity conditions and the existence of a smoothness parameter L are sufficient to imply a mixing time that is (at worst) exponential in LR^2 (Cheng et al. [26] or Ma et al. [27]). However, this technique is not feasible for the Bayesian mixture posterior, whose construction implies an R which scales linearly with dimension.

Taking a final step back, there is a rich field of research studying parameter estimation and clustering for Gaussian mixtures (e.g. Expectation Maximization, method of moments,

spectral clustering, and more). As there are strong parallels between sampling and estimation, these techniques help guide our study of MCMC, but they are not within the direct focus of this dissertation. We note that our mixing time bounds are *increasing* in the cluster separation, while a large cluster separation tends to make the estimation task *easier*, which is an important divergence between the two tasks.

In summary, the computational challenge of sampling from the Gaussian mixture posterior demands further attention. In Chapters 2 - 4, we will take steps towards characterizing the underlying mixing behavior, and this analysis will guide our study of potential algorithmic solutions. In particular, we will take tools developed in other settings and tailor them to the specific properties of this domain. First, in the remainder of this introductory chapter, we will make our model and computational task explicit—the greedy Gaussian mixture posterior, and the use of Gibbs sampling.

1.4 The Greedy Mixture Posterior

While there is flexibility in the construction of a Bayesian mixture model, all chapters in this dissertation analyze the same shared "greedy" Gaussian mixture posterior, whose definition we now make explicit. Our interest lies in computation, not the choices of model construction, and discussion of particularly relevant variants is relegated to the appendices.

1.4.1 Model Setting

We define a two-component mixture likelihood, comprised by a variable Gaussian (whose center parameter θ is our object of interest), and a fixed component. We refer to this as a "greedy" model—we are adding a single additional variable mixture component to an already specified fixed density. For observed data $\mathbf{x} = (x_1, \ldots, x_N)$ (where each x_i is d-dimensional), let $\mathbf{z} = (z_1, \ldots, z_N) \in \{0, 1\}^N$ denote a latent *labeling* variable, where $z_i = 1$ denotes membership for the variable Gaussian component, and $z_i = 0$ denotes membership

for the fixed component. This hypothetical construction describes which of the two likelihood components implicitly "generated" the observed datum. We define the variable Gaussian density as $p(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, \sigma^2 I)$, with θ denoting the center parameter (our variable of interest), and $\sigma^2 I$ the fixed, spherical covariance. We set a conjugate Gaussian prior $p(\theta) := \mathcal{N}(\theta; 0, (\sigma^2/\alpha)I)$ centered at the origin, for some $\alpha \in (0, 1]$. We use the generic notation $p(x_i \mid z_i = 0)$ to denote the fixed component density. For the purposes of deriving the conjugate posterior, we need not specify this term any further (as the posterior has the same structure for any fixed density), but in Section 1.5.1 we will discuss its how it is defined in practice.

The data generating distribution is the mixture between the variable component and the fixed density, $p(x_i \mid \theta) := \frac{1}{2}[p(x_i \mid z_i = 0) + p(x_i \mid \theta, z_i = 1)]$, with equal weights (we discuss alternatives at the end of the section). We assume each draw is independent, and thus we can write the mixture likelihood as *either* the product of N sums (of the two components) *or* the sum over 2^N potential labelings,

$$\begin{aligned} p(\mathbf{x} \mid \theta) &= \prod_{i=1}^{N} \frac{1}{2}[p(x_i \mid z_i = 0) + p(x_i \mid \theta, z_i = 1)] \\ &= \frac{1}{2^N} \sum_{\mathbf{z}} \underbrace{\prod_{i=1}^{N} p(x_i \mid z_i, \theta)}_{p(\mathbf{x}|\mathbf{z},\theta)} \\ &= \frac{1}{2^N} \sum_{\mathbf{z}} p(\mathbf{x} \mid \mathbf{z}, \theta). \end{aligned}$$

Thus, the conjugate posterior similarly takes the form of a sum over exponentially many potential labelings,

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x} \mid \theta)p(\theta) \tag{1.1}$$

$$\propto \sum_{\mathbf{z}} p(\mathbf{x} \mid \theta, \mathbf{z})p(\theta). \tag{1.2}$$

Before we derive the explicit form of the posterior, it is useful to pause for a moment of

context. This model reflects a *step* in a greedy procedure, where we are fitting a single additional variable Gaussian (specifically, its center parameter) given observed data and a previously computed fixed density. We have intentionally not specified the full greedy procedure, as it will vary depending on the application. The model may represent a single step in an iterative sampling process, it may be used as an initialization method for Expectation Maximization, or it may explicitly describe the setting of interest (i.e. with no other greedy steps assumed). The critical point is that this model is *broadly reflective* of the fundamental computational challenge faced when sampling from the mixture posterior. We will discuss this further in Section 1.5.1, but in short, it captures the shared structure of an exponential count of Gaussian components governed by latent variables (including the more general case where there are multiple variable components). Thus, the greedy setting narrows our focus to the key underlying local mixing behavior, without being muddied by concerns of identifiability.

A natural alternative to the use of constant equal weights is to treat them as variable component parameters with a specified prior (typically the Dirichlet, as it is conjugate to the mixture model). In Appendix B.1, we consider this choice, but as a brief summary, the use of variable weights does not fundamentally alter our theoretical analysis—given light assumptions, it simply introduces an additional polynomial factor into our bounds (and the impact of non-uniform constant weights is similar). Thus, as the definition of the weights will vary depending on the chosen application, it is sensible to use the clarifying assumption of constant equal weights to narrow our focus to the computational behavior of interest.

1.4.2 Conjugate Posterior

As the Gaussian prior is conjugate to the Gaussian mixture likelihood, the mixture posterior (Equation 1.2) is a sum of exponentially many Gaussian components whose individual

parameters we can compute. For notation, given a labeling $\mathbf{z}$, let

$$N_{\mathbf{z}} := \sum_{i=1}^{N} z_i,$$

denote the number of data points assigned to the variable component, and let

$$\bar{x}_{\mathbf{z}} := \frac{1}{N_{\mathbf{z}}} \sum_{i:z_i=1} x_i$$

denote their sample mean (these terms will be frequently cited throughout our analysis). Then, the posterior distribution is a mixture of 2^N Gaussian densities, where $\tilde{p}(\mathbf{z} \mid \mathbf{x})$ denotes the unnormalized posterior component weight,[1] and $p(\theta \mid \mathbf{z}, \mathbf{x})$ denotes the posterior component density. The explicit formula for the full posterior is given by Lemma 1.4.1 (as with the other proofs in this dissertation, the derivation is relegated to the end of the chapter, in Section 1.6).

Lemma 1.4.1. *For the Bayesian greedy mixture model described in Section 1.4.1, the full formula for the conjugate posterior is given by*

$$p(\theta \mid \mathbf{x}) \propto \sum_{\mathbf{z}} \tilde{p}(\mathbf{z} \mid \mathbf{x}) p(\theta \mid \mathbf{z}, \mathbf{x}), \tag{1.3}$$

with Gaussian component densities

$$p(\theta \mid \mathbf{z}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I),$$

1. We recall that $\tilde{p}$ is general notation used to denote that the density is unnormalized.

whose parameters are

$$\tilde{\mu}_{\mathbf{z}} := \frac{N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}} \bar{x}_{\mathbf{z}},$$
$$\tilde{\sigma}^2_{\mathbf{z}} := \frac{1}{\alpha + N_{\mathbf{z}}} \sigma^2,$$

and whose unnormalized mixture weights are

$$\tilde{p}(\mathbf{z} \mid \mathbf{x}) = \left[\prod_{i:z_i=0} p(x_i \mid z_i = 0) \right] \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{\alpha}{\alpha + N_{\mathbf{z}}} \right)^{\frac{d}{2}} \times \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 \right] \right).$$

For convenient reference, we recall that $N_{\mathbf{z}} := \sum_{i=1}^{N} z_i$ *and* $\bar{x}_{\mathbf{z}} := \frac{1}{N_{\mathbf{z}}} \sum_{i:z_i=1} x_i$ *denote the respective sample size and sample mean of the data subset assigned to the variable Gaussian component under the label* $\mathbf{z}$*.*

These Gaussian component densities follow an intuitive form—they are equivalent to a typical conjugate Gaussian posterior, if the observed data were simply the subset assigned to the variable component under $\mathbf{z}$. Thus, each center parameter $\tilde{\mu}_{\mathbf{z}}$ is a weighted average between the sample mean and the prior center, and the posterior variance $\tilde{\sigma}^2_{\mathbf{z}}$ shrinks as more data are assigned (and our confidence increases).

While the mixture posterior has an exponential component count, proportional queries can be computed in polynomial time through the product of the prior and the likelihood (Equation 1.1). An immediate corollary of this posterior mixture formulation (Equation 1.3) is that given labels drawn according to their posterior distribution $p(\mathbf{z} \mid \mathbf{x})$, it is trivial to generate samples from the original parameter posterior $p(\theta \mid \mathbf{x})$, as the conditional posterior $p(\theta \mid \mathbf{z}, \mathbf{x})$ is simply a Gaussian whose components can be computed. Of course, while the unnormalized density $\tilde{p}(\mathbf{z} \mid \mathbf{x})$ can be easily queried, there are exponentially many potential labelings, and it is difficult to generate label samples from this distribution. Still, this insight

is the crux of Section 1.5, as we can define our Markov chain directly on the state space of the labels, and then translate these label samples into the desired posterior samples of our target parameter θ.

In Chapters 2 - 4, we will characterize the computational challenge of generating samples from the posterior shown in Lemma 1.4.1. However, before we can begin, there is one final missing piece—the sampling method that underpins our analysis. Thus, in Section 1.5, we formally introduce the Markov chain generated by the Gibbs sampler, whose mixing properties will prove central to our study.

1.5 Gibbs Sampling

Gibbs sampling is the canonical MCMC technique for generating samples from the Bayesian mixture posterior. We begin with the high level intuitive premise, before describing its particular implementation for Bayesian mixtures. Consider some joint distribution $p(y_1, \ldots, y_p)$ defined on the p-fold joint space $\mathcal{Y}^p$ which is difficult to sample from (we use this generic state space, $y \in \mathcal{Y}$, to avoid any confusion with the mixture posterior setting). However, suppose that the conditional distributions for each of the p individual variables, $p(y_i \mid \mathbf{y}_{-i})$, are easy to sample from.[2] The Gibbs sampler constructs a Markov chain whose stationary distribution is the specified joint distribution, following a sequence of these conditional draws. At each step, we select the ith variable in the joint space, and update its value with a draw conditioned on the current value of the other variables, $y_i' \sim p(\cdot \mid \mathbf{y}_{-i})$. This index i may be selected via *random scan* (i.e. uniformly at random) or *systematic scan* (i.e. following a pre-defined, deterministic pattern). In this dissertation we mirror the typical literature and exclusively follow a random scan, but generally the distinction is not significant (we discuss this choice, as specialized to our mixture posterior setting, in Appendix B.2).

This premise is naturally suited to the latent variable formulation of the Bayesian mixture

2. We recall that $\mathbf{y}_{-i}$ denotes the collection of variables with the ith index omitted.

posterior. Rather than study the posterior on the parameters $p(\theta \mid \mathbf{x})$ directly, we consider the *complete data* posterior $p(\theta, \mathbf{z} \mid \mathbf{x})$, including the unobserved latent labels. This is an easy fit for the Gibbs sampling framework—the conditional distribution $p(\theta \mid \mathbf{z}, \mathbf{x})$ on the parameters is Gaussian, and the conditional distribution $p(\mathbf{z} \mid \theta, \mathbf{x})$ on the labels has independent data indices. Thus, it is straightforward to generate samples from either conditional distribution, and the marginal $p(\theta \mid \mathbf{x})$ for just the parameter will match our original target posterior. In summary, we alternate conditional draws between the parameters (θ) given the current latent labels ($\mathbf{z}$), and the latent labels given the current parameters, as formalized in the pseudocode of Algorithm 1.

Algorithm 1: The Standard Gibbs Sampler

Let T denote the total number of time steps;

Initialize parameter $\theta^{(0)}$;

for t *in* $\{1, 2, \ldots, T\}$ **do**

  Sample $\mathbf{z} \sim p(\cdot \mid \theta^{(t-1)}, \mathbf{x})$;

  Sample $\theta^{(t)} \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$;

  $t \leftarrow t + 1$;

end

return $\theta^{(T)}$;

For the purposes of this dissertation, it will prove advantageous to push this approach one step further. Let $\mathbf{z} \in \mathcal{Z}$ denote the discrete state space of the labels and let $\theta \in \Omega$ denote the state space of the parameters (under our current greedy construction, $\Omega = \mathbb{R}^d$). The Gibbs method naturally constructs a chain on the joint space $\mathcal{Z} \times \Omega$ (with both labels and parameters), but it can be equivalently framed as a chain operating exclusively on either state space. That is, it can be viewed as a chain defined on the state space of the labels $\mathcal{Z}$ whose transition rule leverages an intermediate parameter θ, or it can be viewed as a chain defined on the state space of the parameters Ω whose transition rule leverages an intermediate label $\mathbf{z}$ (as the parameter is our target for inference, this is the framing we use in Algorithm 1).

However, due to the conjugacy of our model, we could simply *integrate out* the step which generates that intermediate θ, and instead *directly* draw a new label $\mathbf{z}'$ conditioned on the current label $\mathbf{z}$. Generally, this procedure (where we integrate out a conditional draw) is referred to as "collapsing" the Gibbs sampler.

When a closed form formula exists (and is easy to compute), "collapsing" the Gibbs sampler is typically thought to be computationally beneficial (e.g. the theoretical argument of Liu [28]). In our case, it will prove particularly advantageous for the purposes of mixing analysis. The fundamental change is that collapsing the Gibbs sampler allows us to define our Markov chain directly on the discrete space of the posterior labels, $\mathcal{Z}$. This is an equally valid approach to the original sampling task, as we can trivially translate *labels* into *parameters* using the Gaussian form of the conditional posterior (Equation 1.3). But it is favorable for theoretical analysis—the discrete space of the labels will enable a cleaner form of conductance argument, and more broadly, we will observe how the key impediment to mixing lies in the transfer of information between isolated mixture components (discussed in greater detail in Chapter 3). Thus, it will clarify our analysis to capture this behavior *directly* within our Markov chain.

While the use of the collapsed Gibbs sampler is widespread, we cite the work of Tosh & Dasgupta [24] as a useful starting example (as we also draw guidance from their conductance argument, we discuss their work further in Section 2.1). The collapsed Gibbs sampler updates a single data label index with each iteration. Starting at some label $\mathbf{z}$, we transition to a new label $\mathbf{z}'$ through the following steps.

1. Sample a data index $i \in \{1, \ldots, N\}$ uniformly at random.
2. Generate a new datum label, $z_i' \sim p(\cdot \mid \mathbf{z}_{-i}, \mathbf{x})$, where $\mathbf{z}_{-i}$ omits the ith index.
3. Set $\mathbf{z}'$ to reflect this updated z_i': $\mathbf{z}' \leftarrow (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_N)$.

We briefly discuss some alternative implementations in Appendix B.2, and the process that we use throughout this dissertation is formalized in the pseudocode of Algorithm 2.

Algorithm 2: The Collapsed Gibbs Sampler

Let T denote the total number of time steps;

Initialize labeling $\mathbf{z}^{(0)}$;

for t *in* $\{1, 2, \ldots, T\}$ **do**

  Sample uniform $i \in \{1, \ldots, N\}$;

  Sample $z_i' \sim p(\cdot \mid \mathbf{z}_{-i}^{(t-1)}, \mathbf{x})$;

  Set $\mathbf{z}^{(t)} \leftarrow (z_1^{(t-1)}, \ldots, z_{i-1}^{(t-1)}, z_i', z_{i+1}^{(t-1)}, \ldots, z_N^{(t-1)})$;

  $t \leftarrow t + 1$;

end

return $\mathbf{z}^{(T)}$;

If our goal is to generate samples from the posterior, we can simply replace the object we return with a draw $\theta^{(T)} \sim p(\cdot \mid \mathbf{z}^{(T)}, \mathbf{x})$. Thus, the only missing step required to implement this algorithm is the formula for the conditional transition probabilities $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$, whose closed form solution we write in Lemma 1.5.1. We relegate the full derivation (which leverages a convolution over θ) to the end of the chapter (Section 1.6.2). While it cannot substitute for the full computation, it is perhaps instructive to first informally articulate the intuition behind the result.

Given a current label $\mathbf{z}$ and selected index i, there are two possible destinations—assigning the ith datum to the fixed component ($z_i' = 0$), or the variable component ($z_i' = 1$). Each destination is associated with a density (describing how well a datum would fit with that mixture component), and the relative probability of each destination is *weighted* by those densities evaluated at the datum x_i. Thus, the density that provides the relative weight for the *fixed* component destination is simply the fixed density itself, $p(x_i \mid z_i = 0)$. On the other hand, the density that weights the *variable* component destination is the conditional *posterior predictive* density. In short, this represents our current estimate of the variable mixture component (conditioned on the label $\mathbf{z}$), and it is worth making this intuition explicit.

Let $\mathbf{x}_\mathbf{z} := \{x_i : z_i = 1\}$ denote the subset of data assigned to the variable component

under the labeling $\mathbf{z}$. Under a typical conjugate Gaussian model, we represent what we have learned about a parameter θ given observed data $\mathbf{x_z}$ through its conjugate posterior. This exactly matches the form of our mixture posterior conditioned on the label $\mathbf{z}$, which we derived to be $p(\theta \mid \mathbf{z}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I)$ (Equation 1.3). However, we wish to estimate the variable component as defined over *data* (not the parameter), and thus we instead need to represent what we have learned about some future generated *datum* x_i given observed data $\mathbf{x_z}$. This is called the *posterior predictive* density, and it is the natural way to interpret the expression that pops out of our explicit derivation.

Thus, in Lemma 1.5.1, we see that the relative probability of transition is a comparison between the fixed component density $p(x_i \mid z_i = 0)$, and the posterior predictive component density given the labeling $\mathbf{z}_{-i}$, which we write as $\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}} \sigma^2 I)$ (with parameters defined in the lemma).

Lemma 1.5.1. *For the Bayesian mixture posterior described above, and selected data index* $i \in \{1, \ldots, N\}$*, the collapsed Gibbs conditional transition probabilities are given by*

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \begin{cases} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}} \sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}} \sigma^2 I) + p(x_i \mid z_i = 0)}, & \text{for } z_i = 1, \\ \frac{p(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}} \sigma^2 I) + p(x_i \mid z_i = 0)}, & \text{for } z_i = 0, \end{cases} \tag{1.4}$$

with $\tilde{\mu}_{\mathbf{z}_{-i}} := \frac{N_{\mathbf{z}_{-i}}}{N_{\mathbf{z}_{-i}} + \alpha} \bar{x}_{\mathbf{z}_{-i}}$ *and* $\tilde{V}_{\mathbf{z}_{-i}} := 1 + \frac{1}{N_{\mathbf{z}_{-i}} + \alpha}$.

This completes the implementation of the collapsed Gibbs sampler (Algorithm 2).

1.5.1 Idealized Fixed Component

Our formula for the collapsed Gibbs conditional transition probabilities (Lemma 1.5.1) does not yet specify the fixed density, $p(x_i \mid z_i = 0)$. This flexibility is intentional, as it shows that *any* fixed density results in a posterior that is a Gaussian mixture (only the label weights are impacted), and the construction of the model may vary depending on the application. However, in order to place concrete bounds on the mixing time, we will need to specify the

fixed density. In this section, we introduce the form of the density (Equation 1.5) which we will use in our subsequent theoretical analysis (in Chapters 2 - 4).

In particular, we will explain why this specification is the *natural* choice under the greedy framework. As a preview of the result, the fixed density will be an estimate of a mixture component, and the idealized form of this estimate is given by the posterior predictive density on a *previously identified* subset of data. While we nominally call this a "choice" (due to the model's potential flexibility), it is not arbitrary. We will first derive this form, and then discuss its clear motivation—both as an explicit step in a greedy process, and as the model that best reflects the computational challenge of sampling from general mixture posteriors.

We begin with the underlying greedy premise—the iterative addition of new density components to a mixture (the fixed density results from these prior iterative steps). If our task was density estimation, this form would be literal (the output of the previous step is itself a density). In our case of sampling, it is not so direct, but we can derive the parallel form. Each previous step adds a new mixture component density estimate, and in the idealized case, the intuition is that we identify a subset of data, and estimate the Gaussian that generated it.

We start with the explicit form of this estimate for a single step. In short, the estimate is given by the posterior predictive density conditioned on the previously identified subset of data. This mirrors the form of the density estimate we used for the collapsed Gibbs transition probabilities, but for clarity we reiterate that description here. It is convenient to refer to a subset of data using its corresponding latent label (i.e. the set $\mathbf{x_z} := \{x_i : z_i = 1\}$). Let $\mathbf{w}$ denote the label corresponding with the subset that we use to estimate this density component (so that throughout the dissertation it is distinct from the labels $\mathbf{z}$, which we treat as random variables in our model). As described above, in a conjugate Gaussian model, we represent what we have learned about the parameter θ given observed data $\mathbf{x_w}$ through its conjugate posterior. This exactly matches the form of our mixture posterior conditioned on the label $\mathbf{w}$, previously derived as $p(\theta \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{w}}, \tilde{\sigma}^2_{\mathbf{w}} I)$ (Equation 1.3). However, our

goal is to estimate a mixture component density defined over the data (not the parameter). Thus, we instead must represent what we have learned about some future generated *datum* x, given observed data $\mathbf{x}_\mathbf{w}$. This is the posterior predictive density, which we derived in Equation 1.8 as $p(x \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(x; \tilde{\mu}_\mathbf{w}, \tilde{V}_\mathbf{w}\sigma^2 I)$.

When our theoretical mixing bounds require a specified fixed density, we will define it to be the result of a *single* such estimate step,

$$p(x \mid z = 0) := p(x \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(x; \tilde{\mu}_\mathbf{w}, \tilde{V}_\mathbf{w}\sigma^2 I). \tag{1.5}$$

This "choice" is not arbitrary, and it can be viewed as the natural specification under two perspectives—it best reflects both the form of an explicit *greedy* procedure, and the broader computational challenge of sampling from a *general* mixture posterior.

As a model of a greedy procedure, there are two aspects of this specification to consider—the form of the density estimate, and the fact that it represents a *single* step. The intuition for the former is outlined above—while the precise details vary with the application, the goal of the greedy process is to estimate the mixture components that generated the observed data. Thus, in an *idealized* step, we simply estimate the Gaussian density that generated a *specific subset* of data, $\mathbf{x}_\mathbf{w}$ (which we have already identified). Regarding the latter, it is true that Equation 1.5 is nominally restricted to representing the *second step* within a greedy procedure (i.e. with a single component identified, we consider the addition of the next), but for our purposes this is not particularly restrictive. The simplest reason is due to the nature of what we actually wish to prove—in this dissertation, we establish conditions that lead to a problematic mixing bottleneck, and as the iterative construction of *any* mixture must add the crucial second component, a bottleneck in this step implies a bottleneck in the whole process.

However, the primary motivation for this fixed density specification is that it best reflects the broader computational challenge. By Lemma 1.5.1, the collapsed Gibbs transition

probabilities at the current label $\mathbf{z}$ are determined by the comparison of two densities: the fixed density $p(x_i \mid z_i = 0)$, and the posterior predictive density $\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)$. This choice (Equation 1.5) of fixed component will *mirror* the local structure of a general mixture model. We need not delve into formal detail, but as brief overview, consider the collapsed Gibbs transition probabilities for a mixture model with K variable components starting at some label $\mathbf{z}$, with index i selected for transition. The probability of transitioning that label to the kth component is *proportional* to the posterior predictive density for x_i, conditioned on the subset of data currently assigned to that kth component (under the labeling $\mathbf{z}$).[3] Crucially, this comparison between posterior predictive densities *mirrors* what occurs in the greedy setting if we follow Equation 1.5—the change is simply that we use the fixed subset of data $\mathbf{x}_\mathbf{w}$, rather than the subset of data assigned to the kth component under the current labeling $\mathbf{z}$. Returning to the case of the explicit greedy process, we note that similar logic applies—while our fixed density denotes a single mixture component, it will also mirror the general computational behavior observed in later steps. For example, consider a fixed density that is a mixture of two such Gaussian estimates. If they are well-separated (which is the computationally interesting case), then locally the behavior approximately mirrors that of just the dominant nearby component, and the comparison between posterior predictive densities will match what we observed in the original specification.

In summary, the focus of this dissertation is the fundamental computational challenge of sampling from the mixture posterior, not the details of implementing the greedy framework. Thus, we allow flexibility for the fixed density in the initial setup, but when we turn our attention to establishing concrete mixing bounds, we cite the specification of Equation 1.5, as it broadly reflects the underlying computational behavior of interest.

3. This can be easily derived by simply extending the logic used in our greedy case, or an example derivation can be found in Tosh & Dasgupta [24], although parts of their model diverge from ours.

1.6 Proofs for Chapter 1

1.6.1 Proofs for Section 1.4

Proof of Lemma 1.4.1. We consider a single conditional likelihood term in the posterior sum (Equation 1.2). For a given labeling $\mathbf{z}$, we have

$$\begin{aligned} p(\mathbf{x} \mid \theta, \mathbf{z}) &= \prod_{i=1}^{N} p(x_i \mid \theta, z_i) \\ &= \underbrace{\left[\prod_{i:z_i=0} p(x_i \mid z_i = 0)\right]}_{p^{(0)}(\mathbf{x}|\mathbf{z})} \left[\prod_{i:z_i=1} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i - \theta\|^2\right)\right]. \end{aligned}$$

We recall that $N_{\mathbf{z}} := \sum_{i=1}^{N} z_i$ denotes the count of data assigned to the variable component, and $\bar{x}_{\mathbf{z}} := \frac{1}{N_{\mathbf{z}}} \sum_{i:z_i=1} x_i$ denotes the corresponding sample mean. Let $p^{(0)}(\mathbf{x} \mid \mathbf{z})$ be slightly abusive notation for the joint density of all data that are assigned to the *fixed* component.

$$= p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + N_{\mathbf{z}}\|\bar{x}_{\mathbf{z}} - \theta\|^2\right]\right) \quad (1.6)$$

We recall our choice of conjugate normal prior

$$p(\theta) = \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2/\alpha}\|\theta\|^2\right). \quad (1.7)$$

We combine Equations 1.6 & 1.7 to compute a single component in the posterior mixture (Equation 1.2).

$$\begin{aligned} p(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta) = p^{(0)}(\mathbf{x} \mid \mathbf{z}) &\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + N_{\mathbf{z}}\|\bar{x}_{\mathbf{z}} - \theta\|^2\right]\right) \\ &\times \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2/\alpha}\|\theta\|^2\right) \end{aligned}$$

We wish to isolate the dependence on θ. We complete the square, using $\tilde{C} := \frac{1}{2\sigma^2/\alpha} + \frac{N_{\mathbf{z}}}{2\sigma^2} = \frac{1}{2\sigma^2}(\alpha + N_{\mathbf{z}})$, and observe the following factorization.

$$
\begin{aligned}
&= p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i:z_i=1}\|x_i - \bar{x}_{\mathbf{z}}\|^2 - \frac{1}{\tilde{C}}\frac{\alpha N_{\mathbf{z}}}{4\sigma^4}\|\bar{x}_{\mathbf{z}}\|^2 - \tilde{C}\left\|\theta - \frac{\frac{N_{\mathbf{z}}}{2\sigma^2}}{\tilde{C}}\bar{x}_{\mathbf{z}}\right\|^2\right) \\
&= p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i:z_i=1}\|x_i - \bar{x}_{\mathbf{z}}\|^2 - \frac{1}{2\sigma^2}\frac{1}{\left(\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}\right)}\|\bar{x}_{\mathbf{z}}\|^2\right) \\
&\qquad \times \exp\left(-\frac{1}{\frac{2\sigma^2}{\alpha+N_{\mathbf{z}}}}\left\|\theta - \frac{N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}}\bar{x}_{\mathbf{z}}\right\|^2\right)
\end{aligned}
$$

The term that depends on θ identifies the posterior component density for a labeling $\mathbf{z}$—it is Gaussian with mean $\tilde{\mu}_{\mathbf{z}} := \frac{N_{\mathbf{z}}}{\alpha+N_{\mathbf{z}}}\bar{x}_{\mathbf{z}}$ and variance $\tilde{\sigma}^2_{\mathbf{z}} := \sigma^2/(\alpha + N_{\mathbf{z}})$ (intuitively, this is a weighted average between the prior center and sample mean $\bar{x}_{\mathbf{z}}$). The remaining term (which does not depend on θ) is the posterior label weight.

$$
\begin{aligned}
&= p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i:z_i=1}\|x_i - \bar{x}_{\mathbf{z}}\|^2 - \frac{1}{2\sigma^2}\frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}}\|\bar{x}_{\mathbf{z}}\|^2\right) \\
&\qquad \times \left(2\pi\tilde{\sigma}^2_{\mathbf{z}}\right)^{\frac{d}{2}} \underbrace{\left(\frac{1}{2\pi\tilde{\sigma}^2_{\mathbf{z}}}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\tilde{\sigma}^2_{\mathbf{z}}}\|\theta - \tilde{\mu}_{\mathbf{z}}\|^2\right)}_{\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I)} \\
&= p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{\alpha}{\alpha + N_{\mathbf{z}}}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1}\|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}}\|\bar{x}_{\mathbf{z}}\|^2\right]\right) \\
&\qquad \times \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I)
\end{aligned}
$$

We recall that $\tilde{p}$ denotes the unnormalized form of a density. Thus, if we sum this result over all potential labels $\mathbf{z}$,

$$p(\theta \mid \mathbf{x}) \propto \tilde{p}(\theta \mid \mathbf{x}) = \sum_{\mathbf{z}} p^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{\alpha}{\alpha + N_{\mathbf{z}}}\right)^{\frac{d}{2}} \times \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2\right]\right) \times \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I),$$

we reach the desired formula in the statement of the lemma (Equation 1.3). □

1.6.2 Proofs for Section 1.5

Proof of Lemma 1.5.1. The conditional probabilities, $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$, for the collapsed Gibbs sampler could be computed directly using the formula for the unnormalized posterior weights (Equation 1.3). However, that messy calculation obscures the clean form of the result, and it is preferable to derive the transition probabilities directly using a convolution.

The key step is to integrate out θ from the marginal distribution of the data under a specified labeling. For temporary notation, we write $A^1(\mathbf{z})$ as shorthand to denote the marginal distribution of the data identified by $\mathbf{z}$ under the variable Gaussian (implicitly involving the observed data and the known prior), and $A^0(\mathbf{z})$ as the marginal distribution of the data assigned to the fixed component. That is,

$$A^1(\mathbf{z}) := \int p(\theta) \underbrace{\left[\prod_{i:z_i=1} p(x_i \mid z_i = 1, \theta)\right]}_{q^1_\theta(\mathbf{z})} d\theta,$$

$$A^0(\mathbf{z}) := \prod_{i:z_i=0} p(x_i \mid z_i = 0),$$

where we write $q^1_\theta(\mathbf{z})$ for the marginal distribution of the data assigned to the variable Gaussian under the labeling $\mathbf{z}$, conditioned on a given θ. We can use these formulae to write

out the conditional distribution of interest on the labels, which involves this integration by θ. This will require some additional notation. We recall that $\mathbf{z}_{-i}$ denotes the vector $\mathbf{z}$ with the ith index omitted. We also need to be able to refer to the vector when the ith index has been assigned to a specific value—to specify this unusual construction, we write $\mathbf{z}^{[i\to 1]}$ or $\mathbf{z}^{[i\to 0]}$. This denotes the vector $\mathbf{z}$ with the ith index overwritten to equal 1 or 0, respectively (when z_i might previously have been the same or different value). For clarity, we use $\mathbb{P}(z_i = \cdot \mid \mathbf{z}_{-i}, \mathbf{x})$ to denote the probability of the event that z_i takes a specific value. We examine both cases to compute the general distribution, $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$.

$$
\begin{aligned}
&\mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) \propto \mathbb{P}(z_i = 1, \mathbf{z}_{-i}, \mathbf{x}) \\
&\quad = \int p(\theta) p(x_i \mid z_i = 1, \theta) \left[\prod_{\substack{j:z_j=1,\\ j\neq i}} p(x_j \mid z_j = 1, \theta) \right] \left[\prod_{\substack{j:z_j=0,\\ j\neq i}} p(x_j \mid z_j = 0) \right] d\theta \\
&\quad = \int p(\theta) q_\theta^1(\mathbf{z}^{[i\to 1]}) A^0(\mathbf{z}_{-i}) d\theta \\
&\quad = A^1(\mathbf{z}^{[i\to 1]}) A^0(\mathbf{z}_{-i})
\end{aligned}
$$

Intuitively, this is just the product of the marginal distribution of the data assigned to each of the components (variable and fixed), under the labeling $\mathbf{z}$ when the ith datum is explicitly assigned to the variable component. We can write the result for the assignment $z_i = 0$ in similar form,

$$
\mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) \propto A^1(\mathbf{z}_{-i}) A^0(\mathbf{z}^{[i\to 0]}).
$$

As these two probabilities sum up to 1, we can normalize them. We note that the ratio $A^0(\mathbf{z}^{[i\to 0]})/A^0(\mathbf{z}_{-i}) = p(x_i \mid z_i = 0)$, as they only disagree on that single factor in the

product.

$$\begin{aligned}\mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) &= \frac{A^1(\mathbf{z}^{[i\to 1]})A^0(\mathbf{z}_{-i})}{A^1(\mathbf{z}^{[i\to 1]})A^0(\mathbf{z}_{-i}) + A^1(\mathbf{z}_{-i})A^0(\mathbf{z}^{[i\to 0]})} \\ &= \frac{A^1(\mathbf{z}^{[i\to 1]})}{A^1(\mathbf{z}^{[i\to 1]}) + p(x_i \mid z_i = 0)A^1(\mathbf{z}_{-i})} \\ &= \frac{\frac{A^1(\mathbf{z}^{[i\to 0]})}{A^1(\mathbf{z}_{-i})}}{\frac{A^1(\mathbf{z}^{[i\to 1]})}{A^1(\mathbf{z}_{-i})} + p(x_i \mid z_i = 0)}\end{aligned}$$

Similarly, to compute $\mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x})$, we simply replace the numerator with $p(x_i \mid z_i = 0)$. Thus, the final step is to compute the ratio $A^1(\mathbf{z}^{[i\to 1]})/A^1(\mathbf{z}_{-i})$, which requires a convolution.

$$\begin{aligned}\frac{A^1(\mathbf{z}^{[i\to 1]})}{A^1(\mathbf{z}_{-i})} &= \frac{\int p(\theta)p(x_i \mid z_i = 1, \theta)q^1_\theta(\mathbf{z}_{-i})d\theta}{\int p(\theta)q^1_\theta(\mathbf{z}_{-i})d\theta}, \\ &= \int p(x_i \mid z_i = 1, \theta) \underbrace{\frac{p(\theta)q^1_\theta(\mathbf{z}_{-i})}{\int p(\theta')q^1_{\theta'}(\mathbf{z}_{-i})d\theta'}}_{\mathcal{N}(\theta;\tilde{\mu}_{\mathbf{z}_{-i}},\tilde{\sigma}^2_{\mathbf{z}_{-i}}I)} d\theta.\end{aligned}$$

The bracketed term is the posterior distribution of θ under the labeling $\mathbf{z}_{-i}$, with parameters given by variance $\tilde{\sigma}^2_{\mathbf{z}_{-i}} := \sigma^2/(N_{\mathbf{z}_{-i}} + \alpha)$ and mean $\tilde{\mu}_{\mathbf{z}_{-i}} := \frac{N_{\mathbf{z}_{-i}}}{N_{\mathbf{z}_{-i}}+\alpha}\bar{x}_{\mathbf{z}_{-i}}$.

$$= \int p(x_i \mid z_i = 1, \theta)\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{\sigma}^2_{\mathbf{z}_{-i}}I)d\theta$$

This is just the formula for the convolution of the normal, and thus as $p(x_i \mid z_i = 1, \theta) = \mathcal{N}(x_i; \theta, \sigma^2 I)$, we have

$$\begin{aligned}&= \int \mathcal{N}(x_i; \theta, \sigma^2 I)\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{\sigma}^2_{\mathbf{z}_{-i}}I)d\theta \\ &= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, (\tilde{\sigma}^2_{\mathbf{z}_{-i}} + \sigma^2)I). \qquad (1.8)\end{aligned}$$

For notational simplicity, we define the scaling factor on this posterior predictive variance as $\tilde{V}_{\mathbf{z}_{-i}} := 1 + \frac{1}{N_{\mathbf{z}_{-i}}+\alpha} = (\tilde{\sigma}^2_{\mathbf{z}_{-i}} + \sigma^2)/\sigma^2$. We substitute this result into our formula for the transition probabilities, and this completes the proof. □

Chapter 2

Mixing Bounds for the Collapsed Gibbs Sampler

In Chapter 2, we introduce the use of a conductance argument to lower bound the mixing time for the collapsed Gibbs sampler (Section 2.1). We leverage this strategy to establish conditions under which the mixing time will be exponentially slow in the separation of the data clusters (Theorem 2.2.1), and provide empirical evidence which suggests that this result broadly characterizes the mixing behavior of the setting (Section 2.2).

2.1 Conductance Analysis

While the collapsed Gibbs sampler defined by Algorithm 2 exhibits the correct stationary distribution, the central question for any MCMC technique is the rate of convergence, which might be infeasibly slow. It is critical to our practical and theoretical understanding of this task that we can identify which specific data settings lead to a mixing bottleneck. In this section, we introduce a simple *conductance* argument that translates an upper bound on the probability of escaping a given label into a lower bound on the mixing time. In Section 2.2, we will use this technique (Lemma 2.1.1) to establish conditions on the data that guarantee exponentially slow mixing.

While this style of conductance argument is widespread, we cite the work of Tosh & Dasgupta [24] as an inspiration for its use in our setting, albeit with different ends. The target of their analysis uses a general K-component mixture likelihood, with Dirichlet priors on the weights, whereas we consider a greedy approach (with a flexible fixed component). Thus, their analysis devotes significant attention to the issue of non-identifiability (i.e. label switching), which is not a concern in the greedy setting. They study a pair of examples (providing mixing time lower bounds for a certain well-specified arrangement of clusters, and a certain misspecified arrangement of clusters), whereas our goal is to identify general conditions within the greedy setting (which builds the foundation for our annealing analysis in Chapter 3). Defining the Gibbs sampler on the space of the labels is natural (not novel), and their goals diverge from ours, but it is important to note this inspiration for our work. We begin by formalizing our definition of the mixing time, and its relation to the conductance.

2.1.1 Preliminaries

The rapidity of mixing for a Markov chain can be defined in a variety of ways, and we will mirror the literature with an intuitive and common criterion (with definitions drawn from the Levin et al. [29] textbook). For two probability measures μ and ν defined on state space $\mathcal{Y}$, let $\|\mu - \nu\|_{\mathrm{TV}} := \sup_{Y \subset \mathcal{Y}} |\, \mu(Y) - \nu(Y)|$ denote their *total variation distance* (over Borel subsets $Y \subset \mathcal{Y}$). Consider some Markov chain with stationary distribution p and transition kernel $T(\cdot \mid \cdot)$.[1] Let $T^t(\cdot \mid y)$ denote the distribution of the Markov chain after t time steps, initialized at state $y \in \mathcal{Y}$. Then, we define the *maximal distance to stationarity* at time step t (given initial position y) as

$$d(t) := \sup_{y \in \mathcal{Y}} \|T^t(\cdot \mid y) - p\|_{\mathrm{TV}},$$

1. This is often written as $P(\cdot, \cdot)$, but writing it as a conditional transition probability is more clarifying within our work.

and define the *mixing time* for some $\epsilon > 0$ as

$$\tau(\epsilon) := \min\{t : d(t) < \epsilon\}.$$

It is common to set a *fixed* ϵ to determine mixing, typically $\epsilon = 1/4$, which we abbreviate as

$$\tau_{\text{mix}} := \tau(1/4).$$

This mixing time is our primary object of analysis.

One technique to establish bounds on τ_{mix} is to leverage the *conductance* of the chain. Intuitively, this provides a measure of the flow out of a subset (i.e. from Y to its complement Y^C), relative to the total weight of that subset at stationarity (i.e. $p(Y)$). Our definition of the mixing time covers both discrete and continuous $\mathcal{Y}$, but it will be convenient to specialize our definition of conductance to the discrete case of interest (of course, generalizing it is straightforward). If there exists a subset with low conductance, then our chain will be slow to mix (following the descriptive language of "mixing" and "flow", we refer to this as a *bottleneck*).

Definition 2.1.1. For a Markov chain with transition kernel $T(\cdot \mid \cdot)$, and stationary distribution p, the *conductance of a set* Y is defined as

$$\Phi(Y) := \frac{1}{p(Y)} \sum_{\substack{y \in Y \\ y' \in Y^C}} p(y) T(y' \mid y), \tag{2.1}$$

and the *conductance of the Markov chain* is the minimum possible conductance of any set Y such that $p(Y) \leq 1/2$,

$$\Phi^* := \min_{Y : p(Y) < 1/2} \Phi(Y).$$

This definition is powerful because it establishes a straightforward lower bound on τ_{mix}, specifically

$$\tau_{\text{mix}} \geq \frac{1}{4\Phi^*}, \tag{2.2}$$

as proved in 1989 by Jerrum & Sinclair [30] (with an accessible introductory summary to the broader topic of conductance provided by Levin & Peres [29]).

2.1.2 Mixing Time Bound

Our strategy is to use such a conductance argument to prove that certain conditions on the observed data guarantee slow mixing. For any labeling with less than half the total posterior probability mass, we can show that a small probability of transitioning away from that label (i.e. "escape") implies a large mixing time. The premise of Lemma 2.1.1 follows naturally from the definitions above, but it is clarifying to concretely define it in the terms of our specific mixture setting.

We consider the Markov chain that arises from the collapsed Gibbs sampler (Algorithm 2) on the greedy Gaussian mixture posterior (Section 1.4). Let $T(\cdot \mid \cdot)$ denote the collapsed Gibbs transition kernel, which combines the random selection of a transition index i (with uniform probability $1/N$), with the collapsed Gibbs conditional probability of accepting that move (given by Lemma 1.5.1). To write out its explicit form, let $\mathbf{z}'$ denote a destination label that solely differs from the current label $\mathbf{z}$ on the ith index (i.e. $z_i' = 1 - z_i$, and $z_j = z_j'$ for $j \neq i$). Then, the collapsed Gibbs transition kernel can be written as

$$T(\mathbf{z}' \mid \mathbf{z}) = \frac{1}{N} p(z_i' \mid \mathbf{z}_{-i}, \mathbf{x}). \tag{2.3}$$

Further, let $T^*_{\mathbf{z}}$ denote the maximal probability of "escape" from the label $\mathbf{z}$, given by

$$T^*_{\mathbf{z}} := \max_i \{p(1-z_i \mid \mathbf{z}_{-i}, \mathbf{x})\} = \max_{\mathbf{z}' \neq \mathbf{z}} \{NT(\mathbf{z}' \mid \mathbf{z})\}. \tag{2.4}$$

We note that $T^*_{\mathbf{z}}$ is *not* actually increasing with the sample size N, rather the factor of N shown in the final equality of Equation 2.4 simply cancels with the factor of $1/N$ inherent to $T(\mathbf{z}' \mid \mathbf{z})$ (Equation 2.3). To put it simply, $T(\mathbf{z}' \mid \mathbf{z})$ is a transition probability that *includes* the step of randomly selecting an index for transition, while $T^*_{\mathbf{z}}$ is *solely* the maximal probability of accepting any such transition. Then, we use $T^*_{\mathbf{z}}$ to upper bound the probability that a single collapsed Gibbs update moves us away from $\mathbf{z}$ (to any destination).

Thus, Lemma 2.1.1 translates an upper bound on the probability of escaping a given label into a lower bound on the mixing time.

Lemma 2.1.1. *Consider the greedy mixture model described in Section 1.4, and the Markov chain that results from the collapsed Gibbs sampler (Algorithm 2). Then, given any label* $\mathbf{z}$ *with posterior weight* $p(\mathbf{z} \mid \mathbf{x}) \leq 1/2$*, the Markov chain mixing time is lower bounded by*

$$\tau_{mix} \geq \frac{1}{4T^*_{\mathbf{z}}},$$

where $T^*_{\mathbf{z}}$ *is the maximal probability of transitioning away from* $\mathbf{z}$ *under the collapsed Gibbs transition rule (Equation 2.4).*

The weakness of this bound is that it can be crude. There will be situations where a *subset* of labelings form a conductance bottleneck, where transitioning out of the subset is difficult, but transitioning within the subset is easy. However, our bound should still be quite illustrative of the broader structure, and we will provide further evidence for this point through empirical experimentation.

2.2 Characterizing the Mixing Bottleneck

In this section, we will state and prove Theorem 2.2.1, which establishes general conditions on the observed data that guarantee a mixing bottleneck. The premise of the argument is that we will identify an instance of *cluster separation* within the data, implying severe multimodality in the discrete space of the posterior labels. Intuitively, there will be a low probability of escaping the label corresponding with that cluster, even when preferable labels exist outside of its local region (causing a bottleneck). Thus, we can apply the conductance argument of Lemma 2.1.1, and prove that the mixing time must be slow.

2.2.1 Setting

We use a limited set of parameters to characterize the key properties of the data that determine our mixing time bound. This allows us to move beyond simply analyzing the behavior of a given example, to instead characterize the underlying impediment to mixing. In this section, we introduce the data and its descriptive parameters, and in Section 2.2.2 we formally state the theoretical result.

For observed data $\mathbf{x}$, let $\mathbf{z}$ denote the key label whose properties we will analyze. In particular, we will prove that the probability of transitioning away from this label under the collapsed Gibbs sampler is small. Throughout this discussion, we again find it convenient to refer to a subset of data through the *label* vector that "identifies" it (i.e. the data subset $\mathbf{x}_\mathbf{z} := \{x_i : z_i = 1\}$). For our analysis, we assume that $\mathbf{z}$ identifies a cluster of data contained by a reasonably small maximal radius, given by

$$\delta := \max_{i:z_i=1} \|\bar{x}_\mathbf{z} - x_i\|.$$

Then, let the labeling $\mathbf{w}$ identify the subset of data that provides the basis for our previously constructed fixed component density (described in Section 1.5.1). Thus,

$$p(x_i \mid z_i = 0) := \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, \tilde{V}_{\mathbf{w}} \sigma^2 I), \tag{2.5}$$

where we recall $\tilde{\mu}_{\mathbf{w}} := \frac{N_{\mathbf{w}}}{N_{\mathbf{w}}+\alpha} \bar{x}_{\mathbf{w}}$, $\tilde{V}_{\mathbf{w}} := 1 + \frac{1}{N_{\mathbf{w}}+\alpha}$, $N_{\mathbf{w}}$ denotes the sample size of the data subset identified by $\mathbf{w}$, and $\bar{x}_{\mathbf{w}}$ denotes the corresponding sample mean. We do not further specify $\mathbf{w}$ (unlike with $\mathbf{z}$, we do not require that the data subset forms a tight cluster).

We characterize the cluster separation through two key parameters. First, we define

$$u := \|\bar{x}_{\mathbf{w}} - \bar{x}_{\mathbf{z}}\|,$$

where u measures the distance between the two identified sample means. Intuitively, a larger value of u implies that the two densities compared by the collapsed Gibbs transition probabilities will be more divergent. The second of our two key separation parameters, Δ, measures the distance to the closest datum that could be *added* to the variable Gaussian under the labeling $\mathbf{z}$, written as

$$\Delta := \min_{i:z_i=0} \|\bar{x}_{\mathbf{z}} - x_i\|.$$

These twin separation parameters allow us to characterize the difficulty of transitioning away from the labeling $\mathbf{z}$. Critically, we expect these distances to scale with the dimension d. Thus, exponential scaling of the mixing time in the separation parameters implies exponential scaling with *dimension*, which is problematic for computation. A visual representation of these parameters is shown in Figure 2.1.

The final pieces of our construction are the ratios between these distances. The difficulty of transition away from $\mathbf{z}$ is premised on the fact that u is large relative to δ, and our bound

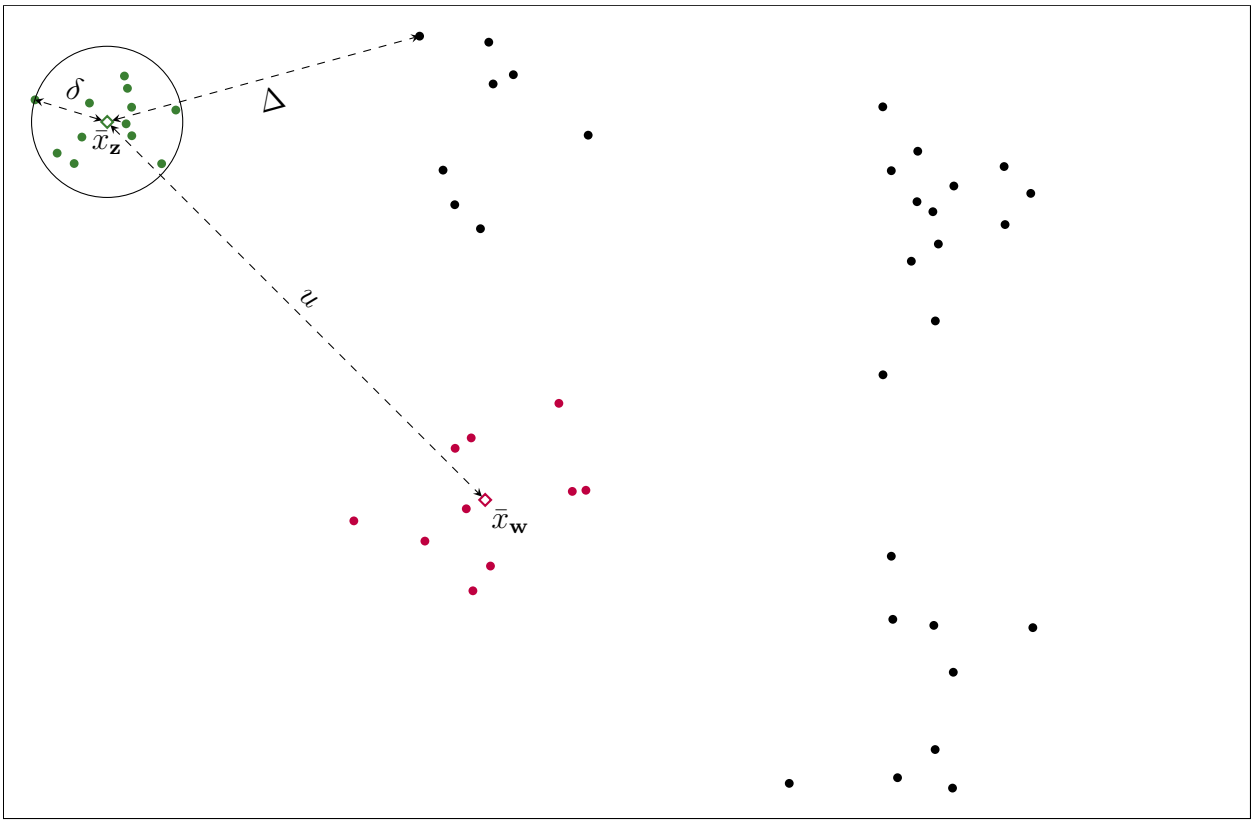


Figure 2.1: Illustration of the cluster separation parameters. Let • denote the cluster identified by the label $\mathbf{z}$, • the cluster identified by $\mathbf{w}$ (for the fixed density), and • the remainder of the data. Let ◇ and ◇ denote the respective sample means for the two specified labels, $\bar{x}_{\mathbf{z}}$ and $\bar{x}_{\mathbf{w}}$. The maximal radius δ, and the cluster separation parameters u and Δ, are shown.

will require that the ratio

$$r_\delta := \frac{\delta}{u}$$

be reasonably small. Finally, to ensure that a transition away from $\mathbf{z}$ is unlikely, we need to control the relative distances from any datum not contained by $\mathbf{z}$ to the twin centers $\bar{x}_\mathbf{z}$ and $\bar{x}_\mathbf{w}$. Thus, we require the ratio

$$R := \max_{i:z_i=0} \frac{\|\bar{x}_\mathbf{w} - x_i\|}{\|\bar{x}_\mathbf{z} - x_i\|}$$

also be reasonably small.

2.2.2 Conditions for Slow Mixing

With these parameters established, we can provide the technical requirements for proving the mixing time bound. In short, if r_δ and R are reasonably small, and the sample size (for the labels we identify) is sufficiently large, then the mixing rate is exponentially slow in our two key separation parameters—with u measuring the distance between the sample means of $\mathbf{z}$ and $\mathbf{w}$, and Δ measuring the minimum distance from the sample mean of $\mathbf{z}$ to any new data point. The requirements we state are not necessarily fundamental barriers, rather, they are chosen for technical convenience to fit with our proof at the end of the chapter.

We require $R < 1/2$ and $r_\delta < 9/40$, to ensure that the identified cluster $\mathbf{z}$ is sufficiently isolated. Then, we place requirements on the sample sizes $N_\mathbf{z}$ (the data cluster we analyze) and $N_\mathbf{w}$ (the data subset used to build the fixed density). The precise sample size requirements are stated in the following box (Equations 2.6 & 2.7). In short summary, these basic assumptions ensure that the removal of any single datum does not have too significant an impact on the relevant parameters (e.g. the sample sizes must be lower bounded by dimension, and they must scale with the distance of the data to the origin).

Sample Size Requirement:

For $N^* := \min\{N_{\mathbf{z}}, N_{\mathbf{w}}\}$, we require

$$N^* \geq \max\{d, 9\}, \tag{2.6}$$

and for any index i, we require

$$N^* \geq \begin{cases} \frac{1}{\delta}\|x_i\| + 1 - \alpha & \text{if } z_i = 1, \\ \frac{10\alpha}{R}\frac{\|x_i\|}{\|\bar{x}_{\mathbf{z}} - x_i\|} - \alpha & \text{if } z_i = 0. \end{cases} \tag{2.7}$$

With these building blocks established, we can state our mixing time bound.

Theorem 2.2.1. *Consider a greedy Gaussian mixture posterior (described in Section 1.4), and the corresponding Markov chain generated by the collapsed Gibbs sampler (Algorithm 2). Let τ_{mix} denote the number of steps required so that the total variation distance to stationarity is at most $1/4$. For observed data $\mathbf{x}$, let $\mathbf{z}$ and $\mathbf{w}$ denote labels such that $R < \frac{1}{2}$, $r_\delta < \frac{9}{40}$, and whose sample sizes satisfy Equations 2.6 & 2.7.*

Then, the mixing time of the resulting Markov chain is exponentially slow in our separation parameters u and Δ, with a lower bound

$$\tau_{mix} \geq \frac{1}{8}\min\left\{\exp\left(\left[\frac{7 - 14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right), \exp\left(\left[\frac{9 - 40r_\delta}{20}\right]\frac{u^2}{\sigma^2}\right)\right\}. \tag{2.8}$$

The proof of this theorem hinges on bounding the conditional transition probabilities for the collapsed Gibbs sampler at the label $\mathbf{z}$, shown in Lemmas 2.2.2 & 2.2.3.

Lemma 2.2.2. *Under the conditions stated in Theorem 2.2.1, the maximal probability of transition away from the labeling $\mathbf{z}$ for any data index i such that $z_i = 0$ is*

$$\max_{i:z_i=0} \mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) \leq 2\exp\left(-\left[\frac{7 - 14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right).$$

Lemma 2.2.3. *Under the conditions stated in Theorem 2.2.1, the maximal probability of transition away from the labeling* $\mathbf{z}$ *for any data index* i *such that* $z_i = 1$ *is*

$$\max_{i:z_i=1} \mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) \leq 2 \exp\left(- \left[\frac{9 - 40 r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right).$$

In summary, when we can identify an isolated cluster of data, the mixing time scales exponentially with respect to the degree of isolation—as measured by u^2/σ^2 (denoting the separation between the fixed density and the cluster center), and Δ^2/σ^2 (denoting the minimum separation from the cluster to any other datum). Crucially, we note that in a typical model setting, we expect both of these terms to scale linearly with dimension (e.g. for centers drawn from our prior $p(\theta) := \mathcal{N}(\theta; 0, (\sigma^2/\alpha)I)$, then $\mathbb{E}[\|\theta\|^2/\sigma^2] = d/\alpha$), and thus, the mixing time will grow as $O(e^d)$. As this is often intractable for interesting applications, we refer to this behavior as a *mixing bottleneck*. It is notable that the conditions we have identified are relatively local—if we examine some isolated cluster, the requirements we place on the layout of the rest of the data are fairly broad.

The main limitation with this theoretical approach is that our conductance argument identifies a *single* label that is difficult to escape from. In practice, the barrier to mixing might arise from a local subset of labels, where transfer *between* the subset elements is easy, but escaping the *whole* subset is difficult. Thus, our definition of cluster "isolation" can be violated by the placement of a single datum. While the idealized theoretical results of Theorem 2.2.1 would still be insightful in their own right (as explicitly stated), we also expect them to be broadly reflective of the typical relationship between cluster separation and mixing time. We will use empirical simulations (Section 2.2.3) to provide supplemental evidence for this behavior in settings that fit the *spirit* of cluster separation, even if they do not satisfy the precise requirements of the stated theorem.

The fact that mixing time scales exponentially in the cluster separation does not clash with our prior intuition, but the value of this analysis lies in quantifying that informal

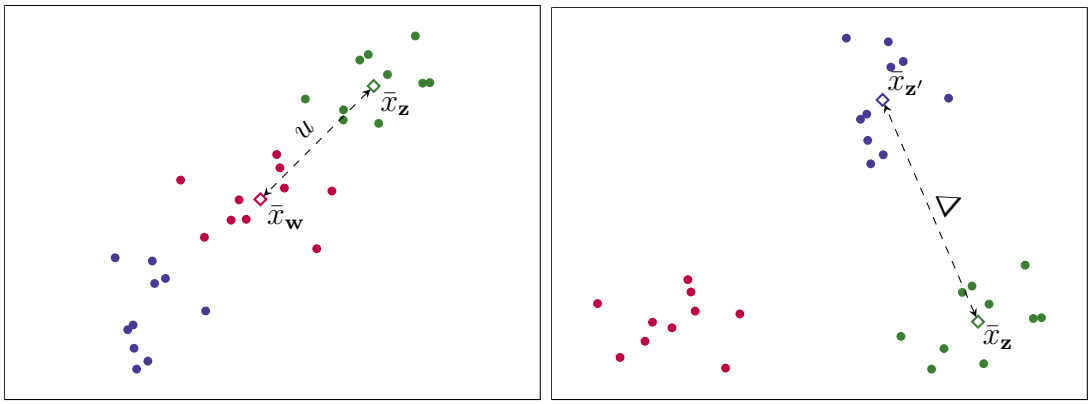
characterization of the mixing behavior. The flexibility we allow in the layout of the data narrows our focus to the key parameters creating the impediment to mixing, and this insight will guide our search for potential solutions in Chapters 3 & 4.

2.2.3 Empirical Simulations

The explicit statement of Theorem 2.2.1 guarantees that when certain conditions are met, there exists a lower bound on the mixing time that grows exponentially in the cluster separation. In this section, we examine empirical simulations that suggest that the central insight of this theorem (the exponential relationship between cluster separation and mixing time) is broadly illustrative of the underlying computational challenge.

In particular, our simulations consider two key points. First, we examine whether this relationship *generalizes* beyond the specific requirements of the theorem statement. That is, we expect the approximately exponential relationship between mixing time and cluster separation to be robust to slight violations of the stated assumptions (e.g. singular data points that violate the cluster spread). Second, we consider whether the exponential scaling on the mixing time *lower bound* actually reflects the mixing time in practice (we could imagine a lower bound that is technically true, but so crude that it offers little insight into the typical behavior).

The primary challenge when using empirical simulations to quantify the mixing time lies in assessing the convergence of the Markov chain. In Theorem 2.2.1, we define the mixing time using the total variation distance, but in practice we cannot easily compute this quantity. However, there are a variety of choices of convergence criteria that similarly characterize the same underlying mixing properties—the key is to follow a *consistent* benchmark when making comparisons. We draw inspiration from the literature, and will use the *potential scale reduction factor* (PSRF) of Gelman & Rubin [31] to assess convergence. As it can be difficult to determine convergence based on the observed behavior of a single chain, the PSRF is computed using a comparison between multiple independent chains. We relegate



(a) Varying u^2.

(b) Varying Δ^2.

Figure 2.2: Illustration of the three-cluster data arrangements, for varying cluster separation (a) u^2, and (b) Δ^2. We associate the three clusters •, •, and •, with the labels $\mathbf{w}$ (the basis for the fixed density), $\mathbf{z}$, and $\mathbf{z}'$, respectively. The colored diamonds (◇, ◇, and ◇) denote their cluster centers. Experiment (a) varies the distance between the fixed cluster center and the variable clusters (i.e. $u := \|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{w}}\|$), and experiment (b) varies the distance between the twin variable cluster centers (i.e. $\Delta := \|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{z}'}\|$). In both cases, the variable cluster centers are equidistant from the fixed cluster center (i.e. $\|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{w}}\| = \|\bar{x}_{\mathbf{z}'} - \bar{x}_{\mathbf{w}}\|$), but these distances are omitted to highlight the separation parameter that varies in the experiment.

the full introduction of this methodology (and discussion of its implementation) to Appendix C.1, and we note that this tool is widespread in the literature (e.g. the van de Meent et al. [32] study that we discuss in Chapter 4).

The first empirical experiment measures the relationship between the mixing time and the u^2 separation parameter, under a natural three-cluster data setting. The first cluster is centered at the origin, and it is used as the basis for our fixed density (i.e. it corresponds with the label $\mathbf{w}$). The second cluster center is placed at distance u from the origin, and the third cluster center is its reflection about the origin (i.e. we multiply the second cluster center vector by -1). This cluster arrangement is illustrated in Figure 2.2a. Each data cluster has an equal sample size of 10, generated from a multivariate Gaussian, and re-centered to have the specified sample mean. This is a natural interpretation of the greedy setting—there are three clusters to learn, we have previously identified the cluster at the origin, and we consider the addition of a new variable component.

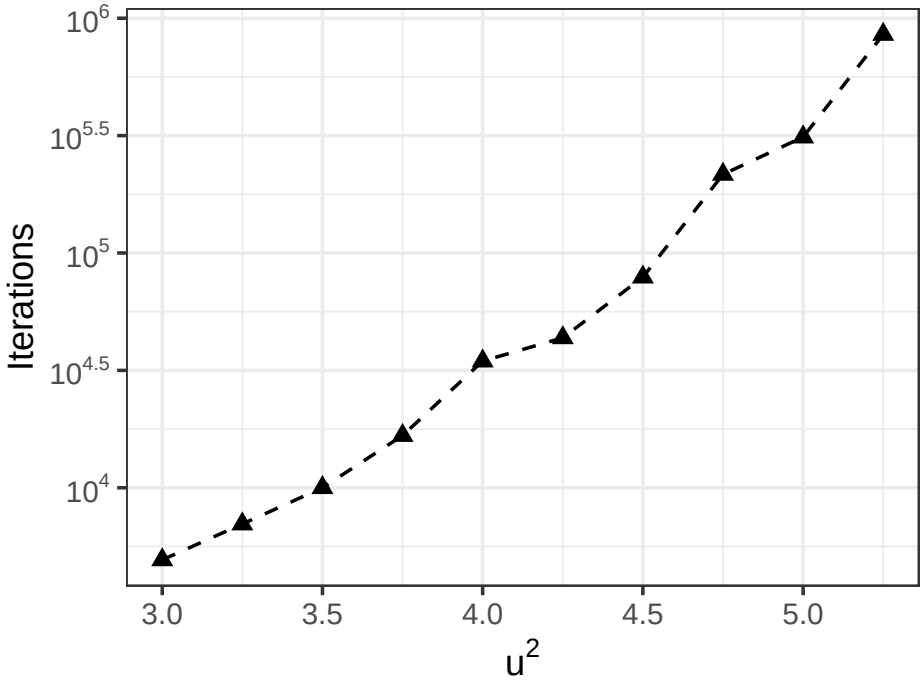
We generate 50 such datasets for each level of u^2, and record the number of Markov

chain iterations needed until our convergence criterion is satisfied. The full experiment specification and methodology is described in Appendix C.2.2. In Figure 2.3a, we plot the mean count of the iterations until convergence is reached on the log scale, and observe that it grows at an approximately linear rate with u^2, matching the exponential rate predicted by the lower bound in Theorem 2.2.1. This provides affirmative evidence for our two primary questions above—the experimental setting does not *exactly* match the theorem assumptions (e.g. as the data are Gaussian, we do not precisely control the cluster radius δ), and it is the observed mixing time that grows exponentially, not just the theoretical lower bound.

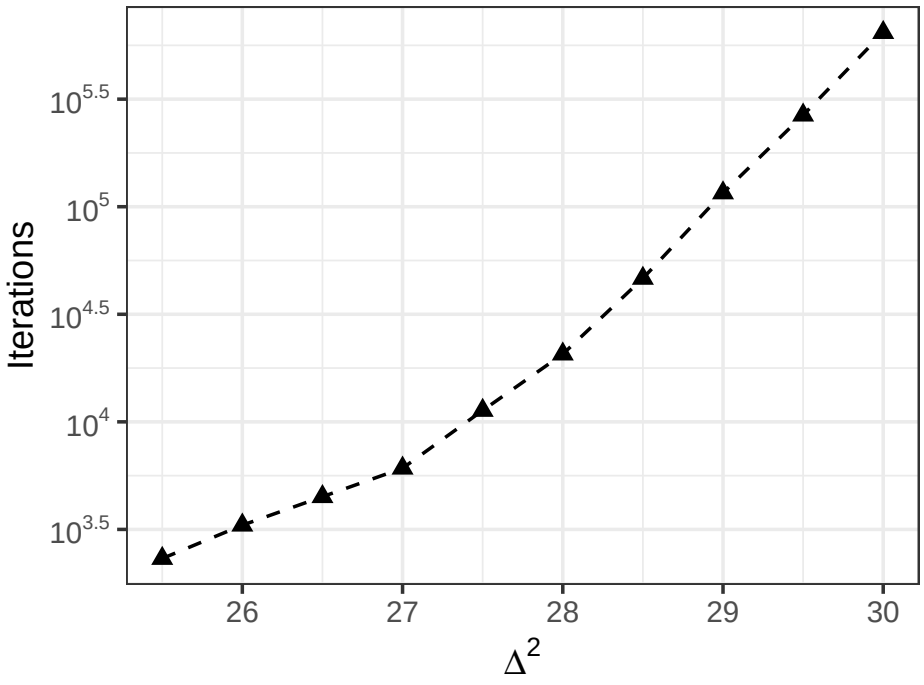
The second experiment follows a similar structure, except we rearrange the data clusters so instead it is the Δ^2 parameter that varies. We leave the first cluster center at the origin, and place the second and third cluster centers so that two conditions are satisfied—they must be distance Δ apart, and they must be equidistant from the fixed cluster center. An example of this data arrangement is illustrated in Figure 2.2b. We note that this is a *slight* departure from the definition of Δ cited in the theorem statement, but it better captures the spirit of cluster separation. As our intention is to characterize the broader relationship between separation and mixing time, this definition is more natural for the setting of the experiment.[2]

The full specification for the experiment is described in Appendix C.2.2. We generate 50 such datasets for each level of Δ^2, and in Figure 2.3b, we again show that the relationship between the mean mixing time on a log scale and Δ^2 is approximately linear. In summary, Theorem 2.2.1 describes an exponential relationship between the mixing time of the chain and a specific definition of the isolation of the data clusters. Both experiments suggest that this relationship is *broadly reflective* of the fundamental computational challenge, beyond the strict statement of the theorem itself.

2. To be precise, the definition of Δ in the theorem statement measures the distance to the nearest outside *datum*, rather than the nearest *cluster center*. The theorem definition is chosen to be *general* (i.e. it does not require that the rest of the data form tight clusters), and it is convenient for technical reasons. In this experiment, the difference is minimal, but as we construct the dataset by specifying the locations of cluster centers, this is the natural choice.



(a) Varying u^2.



(b) Varying Δ^2.

Figure 2.3: The mean number of iterations until convergence is reached (the vertical axis is defined on a log scale), for varying choices of (a) u^2, and (b) Δ^2. See Appendix C.2 for details on methodology.

2.3 Proofs for Chapter 2

2.3.1 Proofs for Section 2.1

Proof of Lemma 2.1.1. We cite the definition of conductance (Equation 2.1), and choose the singleton $\mathbf{z}$ as our subset of the state space. By the the definition of $T^*_{\mathbf{z}}$ (Equation 2.4), we have

$$\Phi^* \leq \Phi(\mathbf{z}) = \frac{1}{p(\mathbf{z})} \sum_{\mathbf{z} \neq \mathbf{z}'} p(\mathbf{z}) T(\mathbf{z}' \mid \mathbf{z}) \leq T^*_{\mathbf{z}},$$

and by the Jerrum & Sinclair [30] mixing time bound (Equation 2.2), we have

$$\tau_{\text{mix}} \geq \frac{1}{4\Phi^*} \geq \frac{1}{4T^*_{\mathbf{z}}}.$$

□

2.3.2 Proofs for Section 2.2

Preliminaries

Before we begin the central proofs, we note a simple bound that will prove useful.

Lemma 2.3.1. *For any $d > 0$,*

$$\left(\frac{d+1}{d}\right)^{\frac{d}{2}} < 2.$$

Proof of Lemma 2.3.1.

$$\begin{aligned}
\log\left(\frac{d+1}{d}\right) &= \log(d+1) - \log(d) \\
&< \log(d) + 1/d - \log(d) \\
&= 1/d
\end{aligned}$$

As $2\log 2 > 1$,

$$< \frac{2\log 2}{d}.$$

If we exponentiate both sides, we have

$$\frac{d+1}{d} < 2^{2/d}.$$

And thus, in total,

$$\left(\frac{d+1}{d}\right)^{\frac{d}{2}} < 2.$$

□

Bounding the Transition Probabilities

Proof of Lemma 2.2.2. We first consider the case where $z_i = 0$, and we escape by switching the label to $z_i = 1$. We want to place an upper bound on the maximal probability of transition away from the current label $\mathbf{z}$. By Lemma 1.5.1, we have

$$\begin{aligned}\max_{i:z_i=0} \mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) &= \max_{i:z_i=0} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I) + p(x_i \mid z_i = 0)} \\ &\leq \max_{i:z_i=0} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{p(x_i \mid z_i = 0)}. \end{aligned} \tag{2.9}$$

By our definition of the fixed density (Equation 2.5), for any i such that $z_i = 0$, we have

$$\begin{aligned}\frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{p(x_i \mid z_i = 0)} &= \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, \tilde{V}_{\mathbf{w}}\sigma^2 I)} \\ &= \underbrace{\left(\frac{\tilde{V}_{\mathbf{w}}}{\tilde{V}_{\mathbf{z}}}\right)^{\frac{d}{2}}}_{A_1} \exp\left(-\frac{1}{2\sigma^2}\left[\underbrace{\frac{\|\tilde{\mu}_{\mathbf{z}} - x_i\|^2}{\tilde{V}_{\mathbf{z}}} - \frac{\|\tilde{\mu}_{\mathbf{w}} - x_i\|^2}{\tilde{V}_{\mathbf{w}}}}_{A_2}\right]\right). \end{aligned} \tag{2.10}$$

We first consider the ratio A_1.

$$\begin{aligned} A_1 &= \left(\frac{\tilde{V}_{\mathbf{w}}}{\tilde{V}_{\mathbf{z}}}\right)^{\frac{d}{2}} \\ &= \left(\frac{1+1/(N_{\mathbf{w}}+\alpha)}{1+1/(N_{\mathbf{z}}+\alpha)}\right)^{\frac{d}{2}} \\ &= \left(\frac{(N_{\mathbf{w}}+\alpha+1)/(N_{\mathbf{w}}+\alpha)}{(N_{\mathbf{z}}+\alpha+1)/(N_{\mathbf{z}}+\alpha)}\right)^{\frac{d}{2}} \end{aligned}$$

We lower bound the denominator by 1.

$$\leq \left(\frac{N_{\mathbf{w}}+\alpha+1}{N_{\mathbf{w}}+\alpha}\right)^{\frac{d}{2}}$$

By the sample size requirement of Equation 2.6, we have $N_{\mathbf{w}}+\alpha \geq d$. As the ratio is increasing in $N_{\mathbf{w}}+\alpha$, we substitute in d to create an upper bound

$$\leq \left(\frac{d+1}{d}\right)^{\frac{d}{2}},$$

and by Lemma 2.3.1, this is

$$\leq 2. \tag{2.11}$$

For the A_2 term in the exponent,

$$A_2 = \frac{\|\tilde{\mu}_{\mathbf{z}} - x_i\|^2}{\tilde{V}_{\mathbf{z}}} - \frac{\|\tilde{\mu}_{\mathbf{w}} - x_i\|^2}{\tilde{V}_{\mathbf{w}}},$$

we cite the sample size condition of Equation 2.6, which guarantees $\tilde{V}_{\mathbf{z}} = (N_{\mathbf{z}}+\alpha+1)/(N_{\mathbf{z}}+\alpha) \leq 10/9$, ensuring

$$\geq \frac{9}{10}\|\tilde{\mu}_{\mathbf{z}} - x_i\|^2 - \|\tilde{\mu}_{\mathbf{w}} - x_i\|^2. \tag{2.12}$$

For the first distance term,

$$\|\tilde{\mu}_{\mathbf{z}} - x_i\| = \left\| \tfrac{N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}} \bar{x}_{\mathbf{z}} - x_i \right\|$$

we temporarily define $a := \frac{N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}}$, for notational simplicity.

$$\begin{aligned} &= \|a(\bar{x}_{\mathbf{z}} - x_i) + (1-a)x_i\| \\ &\geq a\|\bar{x}_{\mathbf{z}} - x_i\| - (1-a)\|x_i\| \end{aligned}$$

The sample size requirement of Equation 2.7 implies that for i such that $z_i = 0$, we have $(1-a)\|x_i\| = \frac{\alpha}{N_{\mathbf{z}}+\alpha}\|x_i\| \leq R\|\bar{x}_{\mathbf{z}} - x_i\|/10$.

$$\begin{aligned} &\geq a\|\bar{x}_{\mathbf{z}} - x_i\| - R\|\bar{x}_{\mathbf{z}} - x_i\|/10 \\ &= (a - R/10)\|\bar{x}_{\mathbf{z}} - x_i\| \end{aligned} \tag{2.13}$$

We follow similar logic for the second distance term in Equation 2.12,

$$\|\tilde{\mu}_{\mathbf{w}} - x_i\| = \left\| \tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} \bar{x}_{\mathbf{w}} - x_i \right\|,$$

and write $b := \frac{N_{\mathbf{w}}}{N_{\mathbf{w}}+\alpha}$ to avoid notational clutter.

$$\begin{aligned} &= \|b(\bar{x}_{\mathbf{w}} - x_i) + (1-b)x_i\| \\ &\leq b\|\bar{x}_{\mathbf{w}} - x_i\| + (1-b)\|x_i\| \end{aligned}$$

By the definition of R, for i such that $z_i = 0$, we have $\|\bar{x}_{\mathbf{w}} - x_i\| \leq \|\bar{x}_{\mathbf{z}} - x_i\|R$. The sample size requirement of Equation 2.7 implies that for i such that $z_i = 0$, we have $(1-b)\|x_i\| = \frac{\alpha}{N_{\mathbf{w}}+\alpha}\|x_i\| \leq R\|\bar{x}_{\mathbf{z}} - x_i\|/10$.

$$\begin{aligned} &= R\|\bar{x}_{\mathbf{z}} - x_i\| + R\|\bar{x}_{\mathbf{z}} - x_i\|/10 \\ &= (11/10)R\|\bar{x}_{\mathbf{z}} - x_i\| \end{aligned} \tag{2.14}$$

We substitute Equations 2.13 & 2.14 into Equation 2.12.

$$\begin{aligned} A_2 &\geq \tfrac{9}{10}\|\tilde{\mu}_{\mathbf{z}} - x_i\|^2 - \|\tilde{\mu}_{\mathbf{w}} - x_i\|^2 \\ &\geq \tfrac{9}{10}\left[(a - R/10)\|\bar{x}_{\mathbf{z}} - x_i\|\right]^2 - \left[11R\|\bar{x}_{\mathbf{z}} - x_i\|/10\right]^2 \\ &= \left[\tfrac{9}{10}(a - R/10)^2 - (11R/10)^2\right]\|\bar{x}_{\mathbf{z}} - x_i\|^2 \end{aligned} \tag{2.15}$$

We expand the squares, and simplify unneeded terms to produce a convenient lower bound (in part, leveraging the fact that $R, a \leq 1$).

$$\geq \left[\frac{9}{10}a^2 - \frac{7}{5}R\right]\|\bar{x}_{\mathbf{z}} - x_i\|^2$$

By the sample size requirement of Equation 2.6, we have $a = N_{\mathbf{z}}/(N_{\mathbf{z}} + \alpha) \geq 9/10$, and thus we can bound $(9/10)a^2 \geq 7/10$.

$$\geq \left[\frac{7}{10} - \frac{7}{5}R\right]\|\bar{x}_{\mathbf{z}} - x_i\|^2$$

This informs our requirement that $R < 1/2$, as we must ensure the positivity of this term. By our definition of Δ, for i such that $z_i = 0$, we have $\Delta \leq \|\bar{x}_{\mathbf{z}} - x_i\|$, which introduces our cluster separation parameter into the bound.

$$\geq \left[\frac{7 - 14R}{10}\right]\Delta^2 \tag{2.16}$$

Finally, we substitute Equations 2.11 & 2.16 into Equation 2.10, and combine this with Equation 2.9 to reach the desired bound,

$$\begin{aligned}\max_{i:z_i=0} \mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) &\leq \max_{i:z_i=0} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{p(x_i \mid z_i = 0)} \\ &\leq A_1 \exp\left(-\frac{A_2}{2\sigma}\right) \\ &\leq 2\exp\left(-\left[\frac{7-14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right). \end{aligned} \tag{2.17}$$

□

Proof of Lemma 2.2.3. We consider a bound that mirrors Lemma 2.2.2, but in the second case—where $z_i = 1$ (and we consider the probability of transitioning to $z_i = 1$). We wish to upper bound

$$\begin{aligned}\max_{i:z_i=1} \mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) &= \max_{i:z_i=1} \frac{p(x_i \mid z_i = 0)}{p(x_i \mid z_i = 0) + \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \\ &\leq \max_{i:z_i=1} \frac{p(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)}. \end{aligned} \tag{2.18}$$

For any i such that $z_i = 1$, we have

$$\frac{p(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \tag{2.19}$$

$$\begin{aligned} &= \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, \tilde{V}_{\mathbf{w}}\sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \\ &= \underbrace{\left(\frac{\tilde{V}_{\mathbf{z}_{-i}}}{\tilde{V}_{\mathbf{w}}}\right)^{\frac{d}{2}}}_{A_3} \exp\left(-\frac{1}{2\sigma^2}\left[\underbrace{\frac{\|\tilde{\mu}_{\mathbf{w}} - x_i\|^2}{\tilde{V}_{\mathbf{w}}} - \frac{\|\tilde{\mu}_{\mathbf{z}_{-i}} - x_i\|^2}{\tilde{V}_{\mathbf{z}_{-i}}}}_{A_4}\right]\right). \end{aligned} \tag{2.20}$$

Our analysis of the A_3 term mirrors our analysis of A_1 in the proof of Lemma 2.2.2, with

$$\begin{aligned} A_3 &= \left(\frac{\tilde{V}_{\mathbf{z}_{-i}}}{\tilde{V}_{\mathbf{w}}} \right)^{\frac{d}{2}} \\ &= \left(\frac{1 + 1/(N_{\mathbf{z}} - 1 + \alpha)}{1 + 1/(N_{\mathbf{w}} + \alpha)} \right)^{\frac{d}{2}} \\ &= \left(\frac{(N_{\mathbf{z}} + \alpha)/(N_{\mathbf{z}} + \alpha - 1)}{(N_{\mathbf{w}} + \alpha + 1)/(N_{\mathbf{w}} + \alpha)} \right)^{\frac{d}{2}} \end{aligned}$$

and as the denominator is at least 1, we have

$$\leq \left(\frac{N_{\mathbf{z}} + \alpha}{N_{\mathbf{z}} + \alpha - 1} \right)^{\frac{d}{2}}.$$

By the sample size requirement of Equation 2.6, we have $N_{\mathbf{z}} + \alpha - 1 \geq d$. As the ratio is increasing in $N_{\mathbf{z}} + \alpha$, we substitute in d,

$$\leq \left(\frac{d+1}{d} \right)^{\frac{d}{2}}$$

and cite Lemma 2.3.1

$$\leq 2. \tag{2.21}$$

For the term in the exponential,

$$A_4 = \frac{\|\tilde{\mu}_{\mathbf{w}} - x_i\|^2}{\tilde{V}_{\mathbf{w}}} - \frac{\|\tilde{\mu}_{\mathbf{z}_{-i}} - x_i\|^2}{\tilde{V}_{\mathbf{z}_{-i}}},$$

our sample size requirement (Equation 2.6) implies $\tilde{V}_{\mathbf{z}_{-i}} = \frac{N_{\mathbf{z}} + \alpha}{N_{\mathbf{z}} + \alpha - 1} \leq 10/9$. As $\tilde{V}_{\mathbf{w}} \geq 1$,

$$\geq \tfrac{9}{10} \|\tilde{\mu}_{\mathbf{w}} - x_i\|^2 - \|\tilde{\mu}_{\mathbf{z}_{-i}} - x_i\|^2. \tag{2.22}$$

For the first of the two distances in Equation 2.22,

$$\|\tilde{\mu}_{\mathbf{w}} - x_i\| = \left\| \tfrac{N_{\mathbf{w}}}{N_{\mathbf{w}}+\alpha} \bar{x}_{\mathbf{w}} - x_i \right\|,$$

we again temporarily define $b := \frac{N_{\mathbf{w}}}{N_{\mathbf{w}}+\alpha}$, mirroring our earlier work.

$$\begin{aligned} &= \|b(\bar{x}_{\mathbf{w}} - x_i) + (1-b)x_i\| \\ &\geq b\|\bar{x}_{\mathbf{w}} - x_i\| - (1-b)\|x_i\| \end{aligned}$$

By the sample size requirement of Equation 2.7, for any i such that $z_i = 1$, we have $(1-b)\|x_i\| = \frac{\alpha}{N_{\mathbf{w}}+\alpha}\|x_i\| \leq \alpha\delta$. Further, by construction, $\|\bar{x}_{\mathbf{w}} - \bar{x}_{\mathbf{z}}\| = u$, and $\|\bar{x}_{\mathbf{z}} - x_i\| \leq \delta$, implying that $\|\bar{x}_{\mathbf{w}} - x_i\| \leq u - \delta$.

$$\geq (u - \delta) - \alpha\delta$$

As $\delta = r_\delta u$,

$$= (1 - r_\delta - \alpha r_\delta)u. \tag{2.23}$$

Before tackling the second distance in Equation 2.22, we note an identity that captures the effect on the sample mean from removing x_i. Taking the full sum over the sample, we have

$$(N_{\mathbf{z}} - 1)\bar{x}_{\mathbf{z}_{-i}} = N_{\mathbf{z}}\bar{x}_{\mathbf{z}} - x_i,$$

and thus we can convert between the sample means using

$$\bar{x}_{\mathbf{z}_{-i}} = \frac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}} - x_i}{N_{\mathbf{z}} - 1}. \tag{2.24}$$

Then, the remaining distance term in Equation 2.22 can be written as

$$\|\tilde{\mu}_{\mathbf{z}_{-i}} - x_i\| = \left\| \tfrac{N_{\mathbf{z}}-1}{N_{\mathbf{z}}-1+\alpha} \bar{x}_{\mathbf{z}_{-i}} - x_i \right\|,$$

and we substitute for $\bar{x}_{\mathbf{z}_{-i}}$ using Equation 2.24,

$$\begin{aligned}
&= \left\| \tfrac{N_{\mathbf{z}}-1}{N_{\mathbf{z}}-1+\alpha} \tfrac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}}-x_i}{N_{\mathbf{z}}-1} - x_i \right\| \\
&= \left\| \tfrac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}}-x_i}{N_{\mathbf{z}}-1+\alpha} - x_i \right\| \\
&= \left\| \tfrac{N_{\mathbf{z}}}{N_{\mathbf{z}}-1+\alpha} \bar{x}_{\mathbf{z}} - \tfrac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha} x_i \right\| \\
&\leq \tfrac{N_{\mathbf{z}}}{N_{\mathbf{z}}-1+\alpha} \|\bar{x}_{\mathbf{z}} - x_i\| + \tfrac{\alpha}{N_{\mathbf{z}}-1+\alpha} \|x_i\|.
\end{aligned}$$

The sample size requirement of Equation 2.7 implies $\frac{\|x_i\|}{N_{\mathbf{z}}-1+\alpha} \leq \delta$. As $\|\bar{x}_{\mathbf{z}} - x_i\| \leq \delta$, we have

$$\begin{aligned}
&\leq \delta + \alpha\delta \\
&= r_\delta(1+\alpha)u. \qquad (2.25)
\end{aligned}$$

We substitute Equations 2.23 & 2.25 into Equation 2.22

$$\begin{aligned}
A_4 &\geq \tfrac{9}{10}\|\tilde{\mu}_{\mathbf{w}} - x_i\|^2 - \|\tilde{\mu}_{\mathbf{z}_{-i}} - x_i\|^2 \\
&\geq \tfrac{9}{10}\left[(1 - r_\delta - \alpha r_\delta)u\right]^2 - \left[r_\delta(1+\alpha)u\right]^2 \\
&= \left[\tfrac{9}{10}(1 - r_\delta - \alpha r_\delta)^2 - r_\delta^2(1+\alpha)^2\right] u^2
\end{aligned}$$

We expand the squares, and simplify the expression (citing $R, a \leq 1$) to produce a convenient lower bound,

$$\leq \left[\frac{9 - 40r_\delta}{10}\right] u^2. \qquad (2.26)$$

This motivates our stated requirement on r_δ, as this scaling factor will be positive as long

as $r_\delta < 9/40$.

Finally, we can substitute Equations 2.21 & 2.26 into Equation 2.20. When combined with Equation 2.18, we observe our desired bound on the density ratio

$$\begin{aligned}\max_{i:z_i=1} \mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) &\leq \max_{i:z_i=1} \frac{p(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \\ &= A_3 \exp\left(-\frac{A_4}{2\sigma^2}\right) \\ &\leq 2\exp\left(-\left[\frac{9 - 40r_\delta}{20}\right]\frac{u^2}{\sigma^2}\right). \end{aligned} \tag{2.27}$$

□

Proof of Theorem 2.2.1. We leverage the upper bound on $T^*_{\mathbf{z}}$ provided by Lemmas 2.2.2 & 2.2.3. Specifically, by Equations 2.17 & 2.27, the maximal probability of transition (for any index i) is bounded by

$$T^*_{\mathbf{z}} \leq 2\max\left\{\exp\left(-\left[\frac{7 - 14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right), \exp\left(-\left[\frac{9 - 40r_\delta}{20}\right]\frac{u^2}{\sigma^2}\right)\right\}, \tag{2.28}$$

and we cite Lemma 2.1.1 to complete the mixing time bound. □

Chapter 3

Temperature Annealing for the Mixture Posterior

In Chapter 3, we analyze the use of temperature annealing and the simulated tempering algorithm to address the slow mixing of the collapsed Gibbs sampler. The goal of simulated tempering is to implicitly reduce the problematic cluster separation (Section 3.1). However, common implementations are poorly suited to the unusual properties of the mixture posterior (Section 3.2). We specialize simulated tempering to our setting through the introduction of *internal annealing* (Section 3.3). While empirical simulations demonstrate its straightforward and effective implementation on a toy example, we show that this is no universal panacea (Section 3.4), and establish conditions under which the original mixing bottleneck (previously identified by Theorem 2.2.1) will persist (Theorem 3.4.4).

3.1 The Annealing Framework

3.1.1 Introduction

The broad strokes of the mixing bottleneck identified in Theorem 2.2.1 come as no surprise—locally-based techniques (such as the Gibbs sampler) struggle to traverse the low-density

valleys separating isolated unimodal regions. The *annealing framework* is a natural way to ameliorate these multimodal impediments by implicitly reducing the effective separation of these regions. While this annealing structure is used in both sampling and optimization, we specialize our language to the task of sampling.

The fundamental premise of annealing is straightforward—when traversal between isolated modes in a multimodal state space is difficult, we instead rely on the transfer of information through auxiliary distributions that are constructed to enable easy exploration. To make this concrete, given a *target* distribution p which is difficult to sample from, we build a bridge between this challenging target and some "easier" (typically flattened) version of the distribution. This bridge is formed by a sequence of *interpolating distributions*, which we write as $p_1, \ldots, p_L$. Under the annealing framework, the interpolating distributions are constructed such that they satisfy the following properties:

1. The final interpolating distribution in the sequence must match our *target* distribution, $p_L = p$.

2. The first interpolating distribution, p_1, must be sufficiently easy to sample from (we call this the *base* distribution).

3. For ℓ and ℓ' which are "close", their corresponding annealed distributions, p_ℓ and $p_{\ell'}$, are also "sufficiently close" based on some chosen criteria (determined by the application of interest). We refer to this as the *spacing* of the interpolating distributions.

This premise is the basis for a wide range of powerful computational techniques, including simulated tempering [33] (the focus of our analysis), simulated annealing [34], parallel tempering [35], tempered transitions [36], and more.[1]

1. It is important to note that the terms "tempering" and "annealing" do *not* have universal definitions in the literature. In this section, we establish the terminology used throughout the dissertation, but this will inevitably clash with some of the works we cite (in particular, we do *not* assume that either term prescribes the precise method for flattening the target).

In the overwhelming majority of cases, the interpolating distributions are constructed using a form of *temperature annealing*, which follows an inverse temperature schedule $0 \leq \beta_1 < \beta_2 < \ldots < \beta_L = 1$.[2] This is the standard approach in the literature, and it is often introduced as a fundamental part of the annealing premise. We prefer the broader definition of annealing provided above, as in Chapter 4 we will explore a potential alternative to the use of temperature. However, as this chapter exclusively studies temperature annealing, it will prove convenient to adopt the terminology into our analysis (i.e. the base distribution at β_1 is "high temperature", and the target distribution at β_L is "low temperature").

The standard density construction for temperature annealing is *direct exponentiation*, written as

$$p_\ell \propto p^{\beta_\ell}. \tag{3.1}$$

This is the canonical choice, not just because it is mathematically intuitive, but because it has a natural physical interpretation. If $p \propto e^{-f}$, then the distribution for the thermodynamic energy equilibrium takes the form $p_\ell \propto e^{-\beta_\ell f}$, where f is the energy function, and β_ℓ the inverse temperature. Thus, the base distribution is a high energy state where movement is easy, and the target distribution is a cold stable state where movement is difficult. The connections between annealing and physical simulations run deep, but they are not of direct interest to our study beyond this initial motivation for the term.

There are a number of variants on temperature annealing that are similar to direct exponentiation,[3] but for this illustration we stick to the most common form. There are several reasons for its omnipresence. First, it is mathematically convenient, as it can be

2. This is referred to as the "annealing schedule", although that term also refers to the resulting sequence of interpolating distributions. While β describes an inverse temperature, it is often informally referred to as simply the "temperature parameter".

3. E.g. in the Bayesian context, we might only apply the exponent to the likelihood, leading to a base distribution that equals the prior. More broadly, we might consider a geometric mixture with some reference density p_0, given by $p_\ell \propto p^{\beta_\ell} p_0^{1-\beta_\ell}$, but this tends to be roughly equivalent for the purposes of our analysis.

readily computed even when we only have access to proportional oracle queries of the target density (this will prove crucial in Section 3.2). Second, we often measure the spacing between interpolating distributions using Metropolis-Hastings acceptance probabilities, which follow a natural form involving the term $e^{-(\beta_\ell - \beta_{\ell'})f}$.

However, in general practice, there is no reason to assume that the spacing of the interpolating distributions provided by Equation 3.1 must be *optimal*. In 1995, Geyer & Thompson [37] were already pushing against the omnipresence of this formulation in the annealing literature, providing examples where it is outperformed by alternative choices. It seems that at least some part of the overwhelming focus on this narrow construction can be traced to the demands of modeling physical systems, where temperature annealing has a natural interpretation. In Chapter 4, we will study a flexible annealing framework that offers potential advantages over this classical temperature-based approach. For the remainder of this chapter, we stick with temperature annealing, but will explore the need for specialized constructions beyond direct exponentiation.

3.1.2 Simulated Tempering

Perhaps the most intuitive implementation of the annealing framework for MCMC sampling is the *simulated tempering* algorithm, which dates back to the work of Marinari & Parisi [33]. As a brief preview, simulated tempering defines a Markov chain on the *joint* space of the original target $y \in \mathcal{Y}$ and the annealing index $\ell \in [L] := \{1, \dots, L\}$ by alternating between transitions that update each of these variables separately. Then, the output of the algorithm is simply the sequence of states for which the annealing index matches the target ($\ell = L$). While this requires the generation of many states that will ultimately be discarded, the use of the auxiliary random variable opens new mixing paths between previously isolated regions, which can circumvent problematic bottlenecks.

The detailed implementation of the algorithm follows from this basic premise. We consider some sequence of interpolating distributions $p_1, \dots, p_L$, with target distribution

p defined on $\mathcal{Y}$. We define a joint distribution $\pi(\ell, y)$ on the state space $[L] \times \mathcal{Y}$, such that the conditional distributions of the joint chain match the interpolating distributions, $\pi(y \mid \ell) = p_\ell(y)$.[4] Thus, if we select all states $\{(\ell, y) : \ell = L\}$, the resulting Markov chain on $\mathcal{Y}$ has the correct stationary distribution for our target, $p_L = p$. For this to be viable, we must ensure that a sufficient proportion of the joint states satisfy $\ell = L$, and thus we define the marginal distribution on the annealing indices to be uniform, with $\pi(\ell) = 1/L$ (at the end of the section, we will discuss how this is the idealized form, and in practice we need not achieve precise uniformity).

To construct the simulated tempering chain on the joint space, we alternate two types of transitions. *State space transitions* hold ℓ fixed, and apply a transition kernel T_ℓ to update the y variable. We simply require that the transition kernels $T_1, \ldots, T_L$ preserve the invariant distribution for the interpolating distribution p_ℓ (a simple choice might be the Metropolis-Hastings random walk). *Annealing index transitions* instead hold y fixed, and update the ℓ index (typically by proposing an adjacent index $\ell' = \ell \pm 1$ at random, and accepting or rejecting the transition with a Metropolis-Hastings probability). It is often convenient to apply multiple state space transitions between each annealing index transition.

The allure of this framework is that while T_L might be unable to escape from a local region (as the target $p = p_L$ is difficult to sample from), we assume that T_1 (corresponding with the base distribution p_1) is rapidly mixing, and thus it must be able to easily explore the full state space $\mathcal{Y}$. For example, consider two points y and y', which reside in two regions separated by a deep valley of low-density space that T_L cannot traverse. Simulated tempering opens up a new path between them—from (L, y), we march to $(1, y)$, then to $(1, y')$ (which is possible because of the rapidly mixing base distribution), and then to (L, y'). Of course, this implies its own set of challenges, and ensuring that the temperature index transitions are viable will be a focus of subsequent analysis.

Before we write the explicit form of the algorithm, there is one remaining step required

4. In this dissertation, we strictly use $\pi(\cdot, \cdot)$ to refer to the simulated tempering joint distribution.

to bridge this theoretical definition with its implementation. In practice, we typically only have access to *proportional* queries of our interpolating distributions (i.e. we can compute $\tilde{p}_\ell$, but not p_ℓ). For example, we recall that the canonical form of temperature annealing was *direct exponentiation*, given by $p_\ell(y) := p(y)^{\beta_\ell}/C_\ell$, where $C_\ell := \int p(y)^{\beta_\ell} dy$ is the relevant normalizing constant. We typically cannot assume knowledge of C_ℓ, and must query $\tilde{p}_\ell(y) := p(y)^{\beta_\ell} \propto p_\ell(y)$. This generally poses no barrier to our ability to define suitable transition kernels T_ℓ for the state space updates. However, the Metropolis-Hastings acceptance probabilities used for the *annealing index* transitions require a ratio of normalizing constants.[5] The use of imprecise normalizing constants for this ratio does not impact the conditional distribution of the joint chain $\pi(y \mid \ell)$ (and thus the distribution of the y samples we return will be correct), but it impacts the *marginal distribution* $\pi(\ell)$. Thus, without reasonable estimates of the normalizing constants, we may not have sufficient representation for each annealing index ℓ in our chain, which can lead to a mixing bottleneck.

In this dissertation, we study the behavior of the simulated tempering joint chain whose stationary distribution satisfies $\pi(y \mid \ell) = p_\ell(y)$ and $\pi(\ell) = 1/L$. While this is an idealized form (whose implementation would technically require exact normalizing constants), it captures the fundamental mixing behavior of interest. To actually construct such a chain, there is the additional challenge of *estimating* the normalizing constants. This process has its own error, which can impede the rate of convergence, but this is best viewed as a *separate* concern (and it is not within the focus of the dissertation). We again emphasize that the use of imprecise normalizing constants does *not* impact the distribution of the output sample—it simply leads to misrepresentation among the annealing indices. Thus, our interest lies solely in the fundamental mixing bottleneck that may arise in the simulated tempering chain, while practitioners need to address the additional challenge of estimating normalizing constants with reasonable accuracy.

5. In the literature, and in this dissertation, these are often informally referred to as "normalizing constants", but they are actually *relative* normalizing constants—the constants that we will use need only satisfy the correct ratios between different indices.

However, there is one exception—our *empirical simulations* still require normalizing constant estimates. We relegate the full discussion of our choice of implementation to Appendix C.3. As a brief preview, we implement the versatile "outer loop" approach for our experiments (where we iteratively estimate the subsequent normalizing constant through ratio importance sampling), and we stick to settings where we can ensure that the simulations reflect the fundamental mixing behavior of interest.

We formalize the simulated tempering algorithm described above in the pseudocode of Algorithm 3. We assume that normalizing constant estimates $\hat{C}_1, \ldots, \hat{C}_L$ have already been computed, and use M to denote the number of times we apply the state space transition kernel between each annealing index update.

Algorithm 3: Simulated Tempering

For T total time steps;

For M state space transitions per annealing index transition;

Let $\hat{C}_1, \ldots, \hat{C}_L$ denote the normalizing constant estimates;

Let $T_1(\cdot \mid \cdot), \ldots, T_L(\cdot \mid \cdot)$ denote the state space transition kernels;

Initialize starting state $(\ell^{(0)}, y^{(0)})$;

Function `StateSpaceTransition(`ℓ, y`):`

- **for** m *in* $\{1, 2, \ldots, M\}$ **do**
  - Generate $y' \sim T_\ell(\cdot \mid y)$;
  - **if** $m < M$ **then**
    - Reset $y \leftarrow y'$;
- **end**
- **return** y' ;

Function `IndexTransition(`ℓ, y`):`

- Sample uniform $\ell' \in \{\ell - 1, \ell + 1\}$;
- Set $Q \leftarrow \min\left(1, \frac{\tilde{p}_{\ell'}(y)/\hat{C}_{\ell'}}{\tilde{p}_\ell(y)/\hat{C}_\ell}\right)$;
- **if** $U \leq Q$ **then**
  - **return** ℓ' ;
- **else**
  - **return** ℓ ;

for t *in* $\{1, 2, \ldots, T\}$ **do**

- $y^{(t)} \leftarrow$ `StateSpaceTransition(`$\ell^{(t-1)}$, $y^{(t-1)}$`)`;
- $\ell^{(t)} \leftarrow$ `IndexTransition(`$\ell^{(t-1)}$, $y^{(t)}$`)`;

end

return $\{y^{(t)} : \ell^{(t)} = L\}$;

We note that simulated tempering is not the only MCMC implementation of the annealing framework. Two closely related algorithms are *parallel tempering* (often referred

to as "replica exchange MCMC" in the physics literature) and *tempered transitions*. We will introduce tempered transitions in Section 4.4, but neither algorithm is the focus of this study. These algorithms share the same underlying annealing framework, and thus generally share the same theoretical mixing analysis, with only slight adjustments required (e.g. the Woodard et al. [15] paper discussed in Section 3.2 covers both simulated and parallel tempering). In this dissertation, our interests lie in the fundamental impediments to mixing that they all share, and we focus on simulated tempering largely because its analysis is the most intuitive. In practice, the optimal choice will vary depending on the application of interest.

3.2 Simulated Tempering for Mixtures

Before we proceed, it is instructive to pause, take a step back, and clarify the plan for our analysis. In Chapter 2, we identified that the mixing time may grow exponentially in the cluster separation. This motivates the use of simulated tempering (and the broader annealing framework) to address the issue, which we introduced in Section 3.1. In Section 3.3, we will introduce our implementation of simulated tempering (specialized to the mixture posterior), and in Section 3.4, we will analyze its mixing properties. But first, the purpose of this section (Section 3.2) is to establish the context for that work within the research literature.

Simulated tempering is a popular sampling technique for mixture targets, and this existing literature will help guide our study. In Section 3.2.1 we introduce the more commonly studied mixture setting (which we call the "generic" mixture setting), and explain how it diverges from our own mixture posterior domain. In Section 3.2.2, we explain the intuition behind these preexisting methods of analysis (in particular the work of Ge et al. [25] & Woodard et al. [40]). Then, in Section 3.2.3, we examine why we cannot simply apply these existing analyses *directly* to the setting of the mixture posterior—both formally in terms of the assumptions we would violate, and informally using the intuition behind their analysis.

Thus, Section 3.2 should not be viewed as a strictly necessary building block for the

implementation and analysis of simulated tempering that follows (in Sections 3.3 & 3.4). Instead, it provides motivation and context for that work. Rather than delve into rigorous proof, we build our intuition for the underlying challenge, which will help to clarify both the form of our simulated tempering implementation, and the relevance of the resulting analysis.

3.2.1 Generic Mixtures

In this dissertation, we study the task of generating samples from the mixture distribution arising as the posterior for a known model given observed data. However, the methods for analyzing simulated tempering on mixtures which we will discuss (in Section 3.2.2) address a different mixture sampling task, one that is more common in the literature. In this section, we introduce the setting, and highlight the ways it differs from our own.

Specifically, we imagine that we wish to generate samples from some general target mixture density

$$p(y) := \sum_{k=1}^{K} w_k f_k(y), \tag{3.2}$$

with (typically log-concave) mixture component densities $f_1, \dots, f_K$ and nonnegative weights such that $\sum_{k=1}^{K} w_k = 1$. In this setting, we assume we must generate these samples solely using oracle value and gradient queries of the density p (or often, just an unnormalized form $\tilde{p}$). We refer to this as the *generic mixture* setting, to distinguish it from our *posterior mixture* setting of interest. While our mixture posterior distribution could be written in this same form, we can further use the known latent variable structure to compute *any* given posterior label weight or component density parameters—the challenge is that there are exponentially many such components. In the generic mixture setting, we may have a small number of mixture components K, and thus access to these individual mixture component densities and weights would make the task of sampling trivial. We may make other assumptions on the properties of the mixture (e.g. requiring each component to have some minimum

weight, or for the covariance matrices to follow a certain form), but we do not assume we can compute the components directly. Thus, our restricted access to solely oracle value and gradient queries, combined with the highly multimodal mixture surface, makes this task problematic for many common sampling techniques.

One implication of this restriction to oracle queries is that implementations of the annealing framework typically must construct the interpolating distributions through *direct exponentiation* (Equation 3.1). In the case of a mixture target, we refer to this as *external annealing*, as this draws a useful contrast with *internal annealing*, which anneals the individual mixture components. That is, in both cases, we anneal the target through an inverse temperature parameter β. External annealing directly exponentiates the target, and we write it as p_β^{Ext},[6]

$$p_\beta^{\text{Ext}}(y) \propto p(y)^\beta = \left(\sum_{k=1}^{K} w_k f_k(y) \right)^\beta . \tag{3.3}$$

Crucially, this can be computed even under a restrriction to proportional oracle queries of the target density (i.e. for generic mixtures). In contrast, under internal annealing we individually modify each component density,

$$p_\beta^{\text{Int}}(y) \propto \sum_{k=1}^{K} w_k f_{k,\beta}(y), \tag{3.4}$$

where $f_{k,\beta}$ is some suitably flattened version of f_k. A common approach might be to divide the covariance of f_k by β, and thus for sufficiently small β there are no more regions of low density between the mixture components (implicitly reducing the effective component separation). The implementation of simulated tempering we introduce in Section 3.3 follows the form of internal annealing, and this distinction will prove important in the analysis of Section 3.2.2.

6. Earlier, we assumed a discrete annealing schedule p_ℓ, but when convenient we analyze the interchangeable version that is annealed using some continuous parameter β.

3.2.2 Mixing Analysis

We consider the task of drawing samples from some generic mixture target (Equation 3.2), where the individual mixture components f_k are log-concave. This typically leads to a target distribution with individually unimodal regions. As the transfer between these regions may require the traversal of a deep low-density valley, this poses an obstacle to mixing (for common Markov chain techniques), and it represents a natural use case for simulated tempering.

The premise of simulated tempering is that the introduction of the auxiliary temperature variable creates a new potential "path" between any two such isolated regions, avoiding the bottleneck. From the starting region, we climb in *temperature* to the flattened base distribution, traverse the state space to the destination region (at high temperature), and descend in temperature until we reach our target. This is the shared underlying premise behind the two primary analysis frameworks that we will discuss—the *state space partition* (which follows from a broader field of research, but we focus on the work of Woodard et al. [15] specialized to simulated tempering), and the *projected chain decomposition* (Ge et al. [25]). While the technical details of these approaches diverge, the underlying intuitive premise is shared. These techniques will prove poorly suited to the posterior mixture setting of interest, but they provide an insightful foundation for our subsequent analysis.

The state space partition technique can be traced back to the seminal work of Madras & Randall (2002) [13], Madras & Zhang (2003) [38], and Bhatnagar & Randall (2004) [39]. However, the most relevant treatment is provided by the later work of Woodard et al., who establish conditions for rapid [15] and slow [40] mixing, and we focus on their analysis. We note that their work is not explicitly focused on mixtures—they simply require that the multimodal target decompose into a partition of unimodal regions. However, the most natural application (as studied in the examples they provide) is that of a mixture density.

This analysis framework uses a state space partition to decompose the mixing process into the following three properties:

1. The rapidity of mixing *within* each partition region, at *any* temperature.

2. The rapidity of mixing *among* different partition regions, at the *highest* temperature.

3. The rapidity of mixing *between* different temperatures, for the *same* partition region.

The rapidity of mixing for the overall joint chain hinges on the rapidity of mixing for these three parts, which we refer to as Requirements 1 - 3. This framing is insightful—as we cannot rely on the transfer of information between isolated regions in the target distribution, we can only reliably assume that this transfer is possible at high temperatures (although we will see they are not necessarily *sufficient* conditions). This decomposition is the natural way to view the premise of simulated tempering, when the state space can be partitioned into unimodal (individually rapidly mixing) regions.

It is illustrative to make this construction explicit. Imagine we have some partition $\mathcal{P}$, which satisfies Requirement 1 above (its regions are individually rapidly mixing). The simulated tempering chain is defined on the joint space $[L] \times \mathcal{Y}$. We can imagine a *projected chain* defined on the discrete joint space $[L] \times \mathcal{P}$, which associates each point $y \in \mathcal{Y}$ with its corresponding partition region in $\mathcal{P}$.[7] This translates Requirements 2 & 3 into a characterization of the mixing properties of the projected chain, and thus we can measure the mixing of the simulated tempering chain through an analysis of the projected chain. This is a rough and intuitive introduction to the argument (omitting technical details), but we can discuss its implications.

This approach is a natural fit for our mixture setting, as we can associate unimodal *mixture components* with unimodal *partition regions*. This provides the foundation for the mixture analysis of Woodard et al. [15]. In the case of our projected chain, if $\mathcal{Z}$ represents the discrete set of mixture component labels,[8] we can now similarly define our projected

7. The "projected" terminology is drawn from Ge et al. [25]—the analysis framework in this section combines their technical argument using a mixture decomposition with the state space partition used by other sources.

8. We use "label" to refer to any mixture component, although its interpretation as a proper "label" variable only truly fits in the case of a mixture posterior—otherwise we simply have a discrete set of components.

chain on the joint space $[L] \times \mathcal{Z}$, which measures the ease of transferring between different mixture components and temperature levels. Throughout this intuitive analysis, we cite this close correspondence, and may informally conflate partition regions and mixture component labels. Thus, it is sensible to use the language of "transfer between mixture components", even when our chain is defined in the θ parameter space.

Unsurprisingly, this "projected chain" mirrors the object of analysis in the *projected chain decomposition* approach of Ge et al. [25]. The underlying technical arguments used to reach this point sharply diverge—rather than use a state space partition, Ge et al. [25] combine simulated tempering with Langevin diffusion to prove a general Markov chain decomposition theorem that relates the spectral gap of the simulated tempering chain to that of the projected chain. For our purposes, the critical point is that their proof of rapid mixing hinges on the analysis of a hypothetical projected chain defined in the joint state space of the mixture component labels and the annealing indices.

Broadly, we will refer to this shared approach for studying mixtures under simulated tempering as *graph-based analysis*. It recognizes that the key impediment to mixing is the transfer of information between the individually unimodal mixture components, and it models this flow using a projected chain on a weighted graph. The fundamental structure underpinning the analysis of Ge et al. [25] and Woodard et al. [15] is the graph shown in Figure 3.1. Each node represents a duple of temperature index and mixture component (with $\ell = 1$ denoting the warmest state). The graph neatly encodes the premise of simulated tempering—the only paths we can *rely on* are formed by the vertical edges (Requirement 3), and the horizontal edges at just the highest temperature (Requirement 2). The node weights correspond with the posterior distribution of the labels (i.e. $(\ell, \mathbf{z})$ has weight $p_\ell(\mathbf{z} \mid \mathbf{x})$), reflecting the volume of flow required.

The graph in Figure 3.1 is a tool for modeling the flow of the chain, not a comprehensive description. That is, we would expect *some* nominal trickle of flow between any two mixture components (corresponding with a horizontal edge) under the state space transition kernel,

even at cold temperatures. However, as the components can be well-separated, we cannot assume that the volume of flow that crosses the low-density valley between them is *sufficient*, unless the temperature is high. In the case of generic mixtures, a successful proof of rapid mixing will trace the flow between isolated components using the paths at the highest temperature, as that is the only interpolating distribution for which can safely assume this exploration *must* be viable.

Thus, the state space partition of Woodard et al. [15] and the projected chain decomposition of Ge et al. [25] can be viewed as technical arguments that connect a hypothetical chain defined on this graph to the actual mixing properties of the original simulated tempering chain. Ge et al. [25] prove that their hypothetical projected chain is rapidly mixing using the method of canonical paths, and Woodard et al. [15] use these paths to encode the properties needed for their state space partition argument, but in both cases the final punchline is similar. Both establish conditions on the generic mixture target under which the simulated tempering chain will be rapidly mixing (we discuss these conditions further in Section 3.2.3).

The key takeaway is that the fundamental challenge for successful mixing lies in the transfer *between mixture components*. The hypothetical projected chain used by Ge et al. [25] is no mere trick to handle technical details, it is the natural way to characterize the mixing properties of the setting. Broadly, this helps to motivate our use of the collapsed Gibbs sampler (Algorithm 2) over the standard Gibbs sampler (Algorithm 1) for the purposes of analysis. Further, in Section 3.3, we will make this concrete by extending the approach to simulated tempering. We will define a joint chain that uses the collapsed Gibbs sampler as its transition kernel (operating on the state space of the labels), allowing us to make the graph-based analysis *explicit*, with no further technical argument required. Before we introduce this technique, it is instructive to first examine the crucial differences that make the mixture posterior setting distinctive from the typical generic mixture setting.

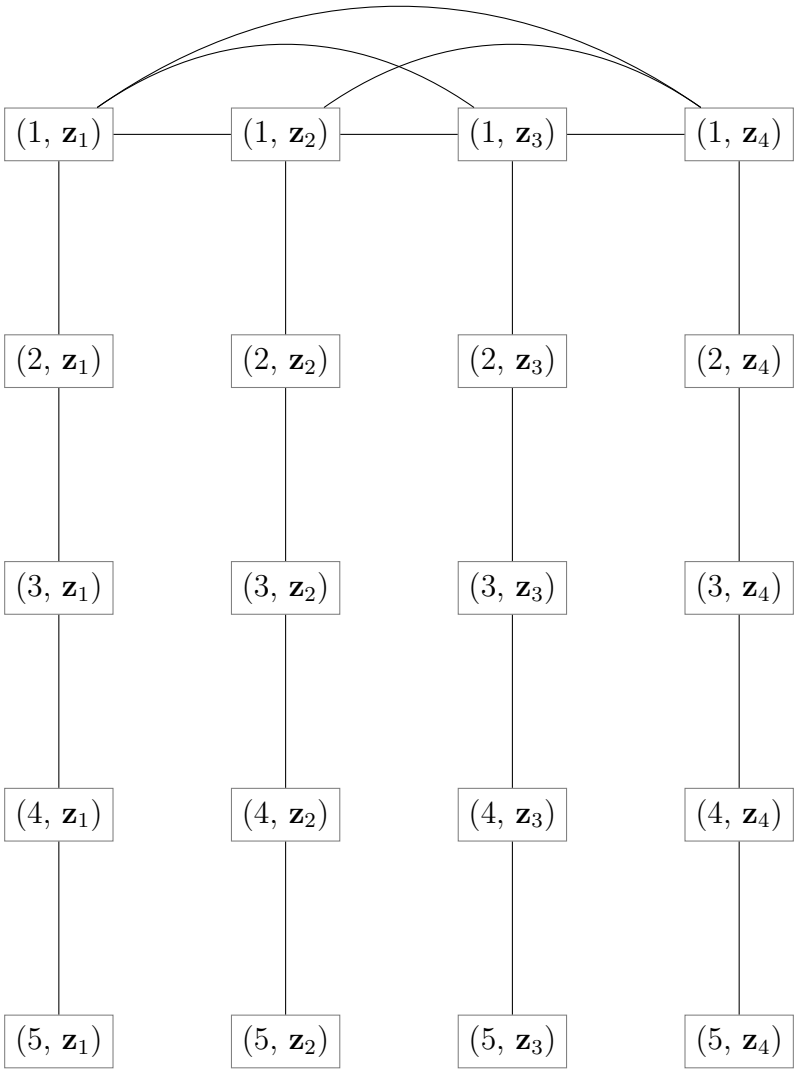


Figure 3.1: The simulated tempering premise (for generic mixtures), encoded as a graph. Each $(\ell, \mathbf{z})$ node represents a duple of temperature index and mixture component (with $L = 5$ and $\mathcal{Z} := \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$). The set of edges models the flow in the simulated tempering chain that we can reliably use in our analysis. As the mixture components may be well-separated, we cannot rely on sufficient flow between mixture components outside of the highest temperature level (i.e. we omit those horizontal edges).

3.2.3 Mixture Posteriors

Our mixture posterior distribution could be written and queried as a generic mixture, and thus it is instructive to explore why we cannot simply *directly* apply these preexisting results analyzing simulated tempering for generic mixtures to our setting. The short answer is simply that Ge et al. [25] assume a polynomial number of components with identical covariance, which is trivially violated by the mixture posterior, and the approach of Woodard et al. [15] fails for similar reasons. However, it is instructive to examine precisely *why* these assumptions are necessary in the first place. The properties of the mixture posterior that prove problematic for this theoretical analysis will suggest the importance of *specializing* our implementation of simulated tempering to the specific structure of the setting (which is the task of Section 3.3).

We begin by returning to the state space partition analysis of Woodard et al. [15]. In Section 3.2.2, we stated three necessary requirements for mixing, but this simplified form glosses over a subtle complication. A general requirement for simulated tempering is that the joint chain spends a sufficient amount of time at each temperature index, which is satisfied as long as our normalizing constant estimates are reasonably accurate. However, under this framework, we face a more stringent restriction—we need to ensure that the amount of time spent in *each partition region* at high temperatures is sufficiently representative of the probability mass for those partition regions at lower temperatures. This property is often called *regional mass preservation*.[9]

This issue can be cleanly articulated using the graph-based analysis introduced in Section 3.2.2. Once again, for the purposes of informal analysis, we cite the natural correspondence between partition regions and mixture components—the technical arguments would differ, but for our purposes they are equivalent decompositions of the multimodal target into uni-

9. This language is common in the literature, but we again note the close correspondence between "regions" and "components" in mixture analysis—thus depending on the context, this might refer to the probability mass of a region, or the weight of a mixture component, as the challenges posed are similar.

modal pieces. Consider two labels $\mathbf{z}_1, \mathbf{z}_2$ which have high weight at low temperature (i.e. $p_L(\mathbf{z}_1 \mid \mathbf{x})$ and $p_L(\mathbf{z}_2 \mid \mathbf{x})$ are large), and assume that their weight shrinks as the temperature increases (i.e. $p_1(\mathbf{z}_1 \mid \mathbf{x})$ and $p_1(\mathbf{z}_2 \mid \mathbf{x})$ are small). In our graph, any path connecting the high weight nodes $(L, \mathbf{z}_1)$ and $(L, \mathbf{z}_2)$ *must* pass through the low weight nodes $(1, \mathbf{z}_1)$ and $(1, \mathbf{z}_2)$, which trivially implies a bottleneck. Thus, this impediment to mixing is encoded in the properties of our weighted graph.

Regional mass preservation is arguably the *central* challenge faced by simulated tempering (and comparable annealing techniques) in high dimensions. In short, annealing a mixture whose components have unequal covariance leads to vanishing probability mass for certain regions (and thus their corresponding components) at high temperatures. Woodard et al. [40] analyze the simple case of a two-component Gaussian, and show that the probability mass assigned to the smaller variance component at high temperatures is exponentially shrinking with dimension. More broadly, Ge et al. [25] use the property of regional mass preservation as the basis for their proof of a two-component mixture where generating samples using solely oracle queries *must* require exponential time (we note that in the posterior mixture setting, we do not suffer from this restriction).

Thus, the issue of regional mass preservation provides a useful perspective for understanding the restrictive assumptions required by Ge et al. [25] and Woodard et al. [15]. Our target is comprised by exponentially many components with differing covariance, a structure that leaves us with little hope that regional mass can be preserved when the density is taken to its β exponent. For example, the recent work of Tawn et al. [41] aims to tackle the issue of regional mass preservation directly, proposing the use of a Hessian approximation to reweight the components (thus preserving the original mass). However, such a technique relies on a small number of well-separated components, which is untenable in the mixture posterior setting.

While this is typically framed as a "regional" effect, it is also illustrative to view it through the lens of our projected chain. We recall from Section 3.2.1 that there are two methods for

constructing our interpolating distributions—*external annealing* anneals the entire sum at once (Equation 3.3), and *internal annealing* anneals each mixture component individually (Equation 3.4). Critically, the projected chain that underpins the analysis of Ge et al. [25] is built using internal annealing. In particular, their proof of its rapid mixing (through a canonical paths argument) requires a sequence of interpolating distributions that preserve the form of a mixture with unchanging weights. However, in reality they are restricted to oracle queries of the target density, and they can only *compute* the externally annealed form (Equation 3.3). Thus, their proof requires that the externally annealed distribution (which they can *query*) is a sufficiently close approximation of the internally annealed distribution (whose mixing properties they can *analyze*). For unequal component covariance, or mixture weights that are too small, this approximation fails, and the internally annealed projected chain cannot be successfully linked with the original simulated tempering chain (which reflects the same underlying issue of regional mass preservation).

The distinction between these annealing forms helps clarify the distinction between the generic and posterior mixture setting. In the case of generic mixtures, external annealing (Equation 3.3) is essentially the only plausible computable choice, but for mixture posteriors it poses some notable disadvantages. First, it precludes the use of Gibbs sampling as the transition kernel. While we have seen that Gibbs sampling is not always rapidly mixing, it is the canonical technique for mixture posteriors with good reason, and it would be unfortunate to instead turn to an unspecialized kernel that neglects to leverage the latent variable structure. Second, when external annealing has been successfully applied to mixtures, regional mass preservation has only been maintained through strong assumptions on the weight and covariance structure of the components, and the mixture posterior is in stark violation of both requirements. However, in the posterior mixture setting, we do *not* face the same restrictions in the construction of our interpolating distributions, and thus external annealing is not the only available choice. In Section 3.3, we introduce a form of *internal annealing* adapted to the posterior mixture setting, which will allow us to perform simulated

tempering with the collapsed Gibbs sampler (enjoying its advantages for both computation and analysis).

3.3 Internal Annealing

Implementations of annealing in MCMC have been popular for decades, as both a target of theoretical study and a tool for practical applications. However, there is a notable paucity of literature that *specializes* its analysis to the posterior mixture setting. This is unfortunate, because the latent variable structure (and resulting exponential component count) makes the mixture posterior a highly unusual target. It could be treated as a generic mixture, but in Section 3.2 we showed that while the preexisting analysis for simulated tempering on generic mixtures was insightful, the properties of the mixture posterior made any direct application problematic. In this section, we introduce an implementation of simulated tempering specialized to the mixture posterior, whose properties we can analyze in Section 3.4.

The most prominent prior work that analyzes annealing techniques *specifically* in the mixture posterior setting is the research of Celeux et al. [19] and Jasra et al. [42]. Their primary concern is label switching (which does not apply in our greedy case), but they mirror our interest in the slow mixing of the Gibbs sampler. As discussed earlier, there are a variety of potential MCMC implementations of annealing, but they share the same theoretical foundation. Both studies choose to implement *tempered transitions* (which we introduce in Section 4.4) for their empirical experiments, but for our purposes, the underlying theoretical mixing analysis is equivalent.

The annealing schedule they study follows the *direct exponentiation* of the posterior, defined by $p_\beta(\theta \mid \mathbf{x}) \propto p(\theta \mid \mathbf{x})^\beta$ (i.e. external annealing). This choice necessitates the use of a transition kernel that ignores the latent structure (typically a form of Metropolis-Hastings). Both papers provide empirical evidence of computational speed-up on real and synthetic datasets. However, while both note the potential weaknesses of annealing in high

dimensions (hinting at the regional mass preservation concerns discussed in Section 3.2), there is minimal further theoretical analysis. This is unsurprising, not just because their primary interest lies in the concern of label switching, but because the opaque form of the externally annealed Gaussian mixture posterior is highly resistant to theoretical analysis. In the years since, there has been minimal research which specializes the implementation of annealing to Bayesian mixtures.

There are notable downsides to the use of external annealing for mixture posteriors, which we briefly reiterate. Computationally, removing the latent variable structure precludes the use of the powerful Gibbs sampler. The analysis of Section 3.2 identifies properties that allow for the preservation of regional mass under external annealing, but these are badly violated by the mixture posterior, which bodes poorly for its use in interesting high-dimensional applications. Finally, the lack of clear structure for the externally annealed target complicates theoretical analysis. The study of the mixing behavior in Chapter 2 was premised on the clean, well-understood mixture structure of the original posterior, whereas the externally annealed posterior is difficult to characterize.

In many common applications, we are restricted to oracle queries (e.g. generic mixtures), and external annealing is essentially the only available option. Crucially, when we specialize to the mixture posterior, we face no such restriction, and the latent variable framework offers us greater optionality in our choice of annealing schedule. In this section, we introduce the natural form of *internal annealing* for the mixture posterior, constructed by individually annealing each component in the likelihood mixture. The resulting posterior preserves the form of a mixture for all temperatures, enabling the use of the collapsed Gibbs sampler within simulated tempering. While this approach is natural, due to the sparsity of the literature, we are unaware of any prior work that formally explores the use of internal annealing for simulated tempering on the mixture posterior. In addition to the potential computational advantages we have suggested, the use of the collapsed Gibbs transition kernel on the discrete label space again enables clean conductance arguments. In Section 3.4 we will use this

to establish conditions that lead to slow mixing (such analysis would prove difficult when grappling with an externally annealed posterior).

First, we must explicitly derive the form of the internally annealed posterior. Previously, we defined our original mixture likelihood as

$$p(\mathbf{x}|\theta) := \prod_{i=1}^{N} \frac{1}{2} \left[p(x_i|z_i = 0) + p(x_i \mid \theta, z_i = 1) \right],$$

with our variable Gaussian defined as $p(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, \sigma^2 I)$. Instead, we write the *annealed* mixture likelihood as

$$p_\beta(\mathbf{x}|\theta) := \prod_{i=1}^{N} \frac{1}{2} \left[p_\beta(x_i|z_i = 0) + p_\beta(x_i \mid \theta, z_i = 1) \right],$$

where we anneal the variable Gaussian by dividing its variance by the temperature parameter

$$p_\beta(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, (\sigma^2/\beta) I).$$

Before we derive the annealed posterior, we clarify two implicit parts of this annealed likelihood. First, we note that throughout this chapter, we allow $\beta = 0$ and explicitly define the resulting likelihood to be the improper uniform distribution. Our interest lies in the posterior, and the $\beta = 0$ case simply sets the posterior to equal the prior (thus, it is still proper). To avoid clutter, we need not explicitly state this trivial $\beta = 0$ case in our derivation. Second, we note that while the fixed component $p_\beta(x_i \mid z_i = 0)$ must also be annealed, we do not specify its definition here, as again the posterior mixture structure is the same for any choice of fixed density (it simply determines the weights). In practice, it is natural for the annealing of the fixed component to mirror that of the variable Gaussian, but the precise details may vary.

The annealed posterior follows the same structure as our original posterior. We can write

the annealed likelihood as a sum over all 2^N possible labelings $\mathbf{z}$, given by

$$\begin{aligned} p_\beta(\mathbf{x}|\theta) &\propto \prod_{i=1}^{N} \frac{1}{2} \left[p_\beta(x_i|z_i = 0) + p_\beta(x_i \mid \theta, z_i = 1) \right] \\ &\propto \sum_{\mathbf{z}} p_\beta(\mathbf{x}|\theta, \mathbf{z}). \end{aligned}$$

Given observed data $\mathbf{x}$ and our (original) prior $p(\theta) := \mathcal{N}(\theta, 0, (\sigma^2/\alpha)I)$, the resulting posterior is proportional to

$$p_\beta(\theta|\mathbf{x}) \propto p_\beta(\mathbf{x}|\theta)p(\theta) \propto \sum_{\mathbf{z}} p_\beta(\mathbf{x}|\theta, \mathbf{z})p(\theta). \tag{3.5}$$

This form clarifies why we call it the *internally annealed* posterior—it is a mixture of component densities that individually depend on the annealing parameter β, rather than applying this annealing to the entire sum (i.e. external annealing). We follow a familiar derivation and compute the explicit formula for the conjugate Gaussian mixture posterior.

Lemma 3.3.1. *For the internally annealed greedy mixture model described in Section 3.3, the full formula for the conjugate posterior is given by*

$$p_\beta(\theta|\mathbf{x}) \propto \sum_{\mathbf{z}} \tilde{p}_\beta(\mathbf{z} \mid \mathbf{x}) p_\beta(\theta \mid \mathbf{z}, \mathbf{x}) \tag{3.6}$$

where,

$$\begin{aligned} \tilde{p}_\beta(\mathbf{z} \mid \mathbf{x}) = \left[\prod_{i:z_i=0} p_\beta(x_i \mid z_i = 0) \right] (\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2/\beta} \right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{\alpha}{\alpha + \beta N_{\mathbf{z}}} \right)^{\frac{d}{2}} \\ \times \exp\left(-\frac{\beta}{2\sigma^2} \left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 \right] \right), \end{aligned}$$

$$p_\beta(\theta \mid \mathbf{z}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\beta}, \tilde{\sigma}^2_{\mathbf{z},\beta} I),$$

and,

$$\tilde{\mu}_{\mathbf{z},\beta} := \frac{\beta N_{\mathbf{z}}}{\alpha + \beta N_{\mathbf{z}}} \bar{x}_{\mathbf{z}},$$
$$\tilde{\sigma}^2_{\mathbf{z},\beta} := \frac{1}{\alpha + \beta N_{\mathbf{z}}} \sigma^2.$$

Thus, internal annealing preserves the structure of the mixture posterior, which allows us to define the collapsed Gibbs sampler for any inverse temperature β. The intuition behind the collapsed Gibbs transition probabilities (under internal annealing) is the same as for the original posterior—the only difference is in the densities that provide the relative weight for the destinations. In particular, the posterior predictive density for the variable component must now reflect the β scaling (and the fixed density is also annealed, though we need not yet specify its form). The explicit formula is shown in Lemma 3.3.2, and the derivation mirrors our earlier work.

Lemma 3.3.2. *For the annealed Bayesian mixture posterior described above (Equation 3.6), and data index $i \in \{1, \ldots, N\}$, the collapsed Gibbs conditional transition probabilities are given by*

$$p_\beta(z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \begin{cases} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{V}_{\mathbf{z}_{-i},\beta}\sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{V}_{\mathbf{z}_{-i},\beta}\sigma^2 I) + p_\beta(x_i | z_i = 0)}, & \text{for } z_i = 1, \\ \frac{p_\beta(x_i | z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{V}_{\mathbf{z}_{-i},\beta}\sigma^2 I) + p_\beta(x_i | z_i = 0)}, & \text{for } z_i = 0, \end{cases} \tag{3.7}$$

for $\tilde{\mu}_{\mathbf{z}_{-i},\beta} := \frac{\beta N_{\mathbf{z}_{-i}}}{\alpha + \beta N_{\mathbf{z}_{-i}}} \bar{x}_{\mathbf{z}_{-i}}$ *and* $\tilde{V}_{\mathbf{z}_{-i},\beta} := \frac{1}{\beta} + \frac{1}{\beta N_{\mathbf{z}_{-i}} + \alpha}$.

In summary, the construction of internal annealing allows for the definition of a simulated tempering chain (Algorithm 3) directly on the state space of the labels $\mathbf{z} \in \mathcal{Z}$, using the collapsed Gibbs transition kernel (Algorithm 2).

3.4 The Persistent Bottleneck

Theorem 2.2.1 states that the mixing time grows exponentially with the separation between data clusters. This result hinges on identifying an exponentially small transition probability for the collapsed Gibbs sampler. Under internal annealing, it is straightforward to see that for a suitably small β, the transition probability (Lemma 3.3.2) to escape any label can be made adequately large (in fact, for $\beta = 0$, this probability is always $1/2$). However, this property is not sufficient to guarantee the rapid mixing of the simulated tempering chain. In this section, we establish conditions under which the change in the posterior label weights causes the original mixing bottleneck to persist, no matter the choice of temperature schedule.

Before we delve into the technical details, it is instructive to outline the intuition behind the argument. The proof of Theorem 2.2.1 hinges on the existence of a label ($\mathbf{z}$) whose maximal probability of escape ($T^*_{\mathbf{z}}$) is exponentially small. For convenience, we can define $\mathbf{z}^* := \arg\max_{\mathbf{z}' \neq \mathbf{z}}\{T(\mathbf{z}' \mid \mathbf{z})\}$ as the destination label that maximizes the probability of transition. This small probability of "escape" implies that the normalized weight $p(\mathbf{z} \mid \mathbf{x})$ is large relative to that of its neighbor labels (including $p(\mathbf{z}^* \mid \mathbf{x})$).

As the temperature increases (and β decreases), the internally annealed posterior tends to push the normalized weights of the labels towards uniformity (at $\beta = 0$, they are exactly equal), which is mirrored by a corresponding increase in the escape probability. To make this explicit, we mirror our notation in Section 2.2, but now include an internal annealing index ℓ. The collapsed Gibbs transition kernel, $T_\ell(\cdot \mid \cdot)$, combines the random selection of a transition index i (with uniform probability $1/N$) with the (annealed) collapsed Gibbs conditional probability of accepting that move (Lemma 3.3.2). Reproducing the earlier notation, if $\mathbf{z}'$ denotes a destination label that differs from the current label $\mathbf{z}$ on solely the ith index (i.e. $z'_i = 1 - z_i$, and $z_j = z'_j$ for $j \neq i$), then the collapsed Gibbs transition kernel

under internal annealing is given by

$$T_\ell(\mathbf{z}' \mid \mathbf{z}) = \frac{1}{N} p_{\beta_\ell}(z_i' \mid \mathbf{z}_{-i}, \mathbf{x}). \tag{3.8}$$

Thus, the maximal probability of escape at temperature index ℓ is given by

$$T^*_{\ell,\mathbf{z}} := \max_i p_{\beta_\ell}(1 - z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \max_{\mathbf{z}' \neq \mathbf{z}} N T_\ell(\mathbf{z}' \mid \mathbf{z}). \tag{3.9}$$

Again, we note that this is *not* increasing in the sample size—the factor of N (in the second equality of Equation 3.9) just cancels with the factor of $1/N$ within the transition kernel (Equation 3.8) arising from the random selection of the data index. Intuitively, the push towards uniformity implies that if the posterior label weight $p_L(\mathbf{z} \mid \mathbf{x})$ is initially much larger than its destination weight $p_L(\mathbf{z}^* \mid \mathbf{x})$, this disparity shrinks as the temperature increases. Typically, this implies that an increasing probability of escape, $T^*_{\ell,\mathbf{z}}$, corresponds with a shrinking normalized posterior weight for the origin label, $p_\ell(\mathbf{z} \mid \mathbf{x})$. This potential coupling will prove central to our analysis.

With these building blocks established, we can outline the intuition behind our argument. The premise of simulated tempering is that when we are unlikely to transition away from $(L, \mathbf{z})$ in the *label space* (through the collapsed Gibbs sampler), we can instead march in the *annealing index* to some $(\ell', \mathbf{z})$, where $\beta_{\ell'}$ is hot enough so that escape (in the label space) is feasible (i.e. $T^*_{\ell',\mathbf{z}}$ is adequately large). However, if the increase in temperature causes the label weight $p_\ell(\mathbf{z} \mid \mathbf{x})$ to severely shrink, actually reaching the sufficiently hot temperature $\beta_{\ell'}$ may be difficult. In summary, if the increasing *escape probability* $T^*_{\ell,\mathbf{z}}$ is too tightly tied to the decreasing *normalized weight* $p_\ell(\mathbf{z} \mid \mathbf{x})$ (of the origin label) as the temperature rises, then the mixing bottleneck may persist.

To be precise, we adapt our earlier conductance argument to the simulated tempering joint space by selecting the *subset* $\mathbf{Q} := \{(\mathbf{z}, \ell) : \ell \in [L]\}$ (corresponding with the labeling $\mathbf{z}$ at *all* temperature levels), and analyzing its conductance. While the simulated tempering chain

alternates transitions on the label space and index space, the definition of $\mathbf{Q}$ ensures that the only potential for escape is through the label space (i.e. the transition kernel $T_\ell(\cdot \mid \cdot)$). Then, if $\pi(\ell, \mathbf{z})$ denotes the joint stationary distribution of the simulated tempering chain, the conductance of $\mathbf{Q}$ is bounded by

$$\begin{aligned}\Phi(\mathbf{Q}) &:= \frac{\sum_{\ell\in[L],\mathbf{z}'\neq\mathbf{z}} \pi(\ell,\mathbf{z})T_\ell(\mathbf{z}' \mid \mathbf{z})}{\sum_{\ell\in[L]}\pi(\ell,\mathbf{z})} \\ &\leq \frac{\sum_{\ell\in[L]} p_\ell(\mathbf{z}\mid\mathbf{x})T^*_{\ell,\mathbf{z}}}{\sum_{\ell\in[L]} p_\ell(\mathbf{z}\mid\mathbf{x})} \\ &\leq \sum_{\ell\in[L]} \frac{p_\ell(\mathbf{z}\mid\mathbf{x})}{p_L(\mathbf{z}\mid\mathbf{x})} T^*_{\ell,\mathbf{z}}. \end{aligned} \tag{3.10}$$

This clarifies the intuition behind our planned argument. At cold temperatures, $\mathbf{z}$ has high weight but a low escape probability, and at high temperatures, $\mathbf{z}$ has low weight but a high escape probability. This implies a potential bottleneck, as the maximum possible flow out of the subset for each label is limited by the weight of that label. Thus, our analysis will compare the change in normalized weight to the change in escape probability. The critical term is thus the ratio of the normalized weights, $\frac{p_\ell(\mathbf{z}|\mathbf{x})}{p_L(\mathbf{z}|\mathbf{x})}$, as the temperature changes. The challenge lies in the fact that we can only easily compute the *unnormalized* weights, and in Section 3.4.1, we introduce the technique we use to bound this ratio.

3.4.1 Growth Factors

The key term in Equation 3.10 is the ratio of normalized weights, $\frac{p_\ell(\mathbf{z}|\mathbf{x})}{p_L(\mathbf{z}|\mathbf{x})}$, which we wish to bound as a function of ℓ. While our conductance argument uses a discrete temperature schedule ($\ell \in [L]$), for the purposes of analysis we instead prefer the continuous parameter β. That is, our goal will be to bound the ratio of the normalized weights $\frac{p_\beta(\mathbf{z}|\mathbf{x})}{p_1(\mathbf{z}|\mathbf{x})}$, as a function of β.[10]

10. We note that the p_1 in this ratio refers to $\beta = 1$, which corresponds with $\beta_L = 1$, *not* $\ell = 1$. While we may use either p_β and p_ℓ, it should always be clear which we mean from the context of the expression.

We recall that we can write the normalized weights as

$$p_\beta(\mathbf{z} \mid \mathbf{x}) = \frac{\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x})}{\sum_{\tilde{\mathbf{z}}} \tilde{p}_\beta(\tilde{\mathbf{z}} \mid \mathbf{x})}, \tag{3.11}$$

where the unnormalized weights have a closed form (with $\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x}) = \int p_\beta(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta) d\theta$, in Equation 3.6). Thus, while we cannot compute ratios of normalized weights at different temperatures (the normalizing constants will not cancel), we can compute ratios of *unnormalized* weights at different temperatures. That is, we will define the *unnormalized growth factor* of a label $\mathbf{z}$ as

$$r_{\mathbf{z}}(\beta) := \frac{\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x})}{\tilde{p}_1(\mathbf{z} \mid \mathbf{x})},$$

which is the ratio between the *unnormalized* weight at some inverse temperature β and at the original target ($\beta = 1$).

An analysis of these growth factors can bound a ratio of normalized weights, and the intuition behind the argument is straightforward. Consider some "good" labeling $\mathbf{z}$ (which has high weight in the original target posterior), and some subset of "bad" labelings $\mathbf{Z}'$ (which have low weight). We will prove that as the temperature increases, the unnormalized weight of labels $\mathbf{z}' \in \mathbf{Z}'$ (assuming the subset is suitably chosen) will grow at a faster rate than the unnormalized weight of $\mathbf{z}$ (that is, $r_{\mathbf{z}}(\beta) < r_{\mathbf{z}'}(\beta)$). If $\mathbf{Z}'$ contains a sufficient proportion of the original total probability mass, then the unnormalized weight of $\mathbf{z}$ must be growing at a slower rate than its normalizing constant, and thus its normalized weight will shrink.

We formalize this argument in Lemma 3.4.1. While the premise is general, for clarity, we write it in the notation of our setting.

Lemma 3.4.1. *Consider some label* $\mathbf{z}$ *with growth factor* $r_{\mathbf{z}}(\beta)$*. For a given growth factor*

$r^*(\beta)$, let $\mathbf{Z}'$ denote a subset satisfying

$$r_{\mathbf{z}'}(\beta) \geq r^*(\beta)$$

for all $\mathbf{z}' \in \mathbf{Z}'$, and

$$\sum_{\mathbf{z}' \in \mathbf{Z}'} p_1(\mathbf{z}' \mid \mathbf{x}) \geq c^*,$$

so $\mathbf{Z}'$ contains at least the fraction c^* of the original (at $\beta = 1$) total probability mass. Then,

$$\frac{p_\beta(\mathbf{z} \mid \mathbf{x})}{p_1(\mathbf{z} \mid \mathbf{x})} \leq \frac{1}{c^*} \frac{r_{\mathbf{z}}(\beta)}{r^*(\beta)}.$$

With this tool in hand, we need only specify the conditions on the data that allow us to identify a suitable subset $\mathbf{Z}'$.

3.4.2 Conditions for Slow Mixing

The target of our analysis is the same setting as our earlier Theorem 2.2.1, which we recall established conditions under which the collapsed Gibbs sampler was slowly mixing. As a brief preview, in this section we will characterize *additional* conditions under which that original mixing bottleneck *cannot* be ameliorated through the use of simulated tempering.

We need not reproduce the full notation of that setting here (as most of the details are not required for this further analysis), and we focus on just the relevant parts. As discussed, the annealing on the fixed component density simply mirrors the annealing of the variable component, where we divide the variance by β (except for $\beta = 0$, which is defined to be uniform). That is, our annealed fixed density is now given by

$$p_\beta(x_i \mid z_i = 0) := \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, (\tilde{V}_{\mathbf{w}} \sigma^2 / \beta) I), \tag{3.12}$$

where we recall that $\mathbf{w}$ denotes the previously identified subset of data used for the fixed component. Under this definition, we can compute the unnormalized growth factor for any labeling $\mathbf{z}$ (it need not be the target labeling of interest used in the theorem itself). In our growth factor analysis, we always assume $\beta > 0$, as when $\beta = 0$ the result is trivial (the unnormalized weights are uniform, so the growth factor is just the inverse of the starting weight).

Lemma 3.4.2. *For the internally annealed greedy mixture posterior (Equation 3.6), with annealed fixed component density defined by Equation 3.12, the growth factor (assuming $\beta > 0$) for any labeling $\mathbf{z}$ is given by*

$$r_{\mathbf{z}}(\beta) = \beta^{\frac{Nd}{2}} \left(\frac{\alpha + \beta N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}} \right)^{\frac{d}{2}} \exp\left((1-\beta) \frac{[SS_{\mathbf{z}}]}{2\sigma^2} \right), \tag{3.13}$$

where

$$[SS_{\mathbf{z}}] := \sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\tilde{V}_{\mathbf{w}}} \sum_{i:z_i=0} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2. \tag{3.14}$$

The $[\mathrm{SS}_{\mathbf{z}}]$ notation refers to the sum of squares, as this term is primarily determined by the distances from the data to their corresponding sample means under the labeling $\mathbf{z}$. Intuitively, it measures whether the labeling is well-suited to the observed data.

Our argument will hinge on the comparison of two labels—the higher weight origin $\mathbf{z}$, and the lower weight destination $\mathbf{z}^*$. We will identify some subset of labels $\mathbf{Z}'$ (with sufficient total weight), whose growth factors are each at least as large as that of $\mathbf{z}^*$. The intuition is that $\mathbf{Z}'$ represents a set of labels that are a "worse" fit to the data than $\mathbf{z}^*$ (as measured by the sum of squares term, $[\mathrm{SS}_{\mathbf{z}}]$). As the sum of squares term tends to dominate the expression, there is typically a direct correspondence between a smaller weight and a larger growth factor. However, for our technical proof, we need to be precise and also consider the other term in the product, which is a function of the sample size $N_{\mathbf{z}}$. Thus, we define $\mathbf{Z}'$

through requirements on *both* the sum of squares term $[SS_{\mathbf{z}'}]$ and the sample size $N_{\mathbf{z}'}$, to ensure the correct inequality (however, the simple intuition behind this subset is simply that these labels are "at least as poor of a fit to the data as $\mathbf{z}^*$").

Lemma 3.4.3. *For two labels $\mathbf{z}', \mathbf{z}^*$ such that $[SS_{\mathbf{z}'}] \geq [SS_{\mathbf{z}^*}]$ and $N_{\mathbf{z}'} \leq N_{\mathbf{z}^*}$, we have*

$$r_{\mathbf{z}'}(\beta) \geq r_{\mathbf{z}^*}(\beta).$$

Then, our statement of the mixing bound (Theorem 3.4.4) takes the conditions from Theorem 2.2.1, and states that as long as such a subset of labels $\mathbf{Z}'$ exists, internal annealing cannot avoid a mixing bottleneck.

Theorem 3.4.4. *Consider the greedy Gaussian mixture posterior that follows the setting of Theorem 2.2.1, with label of interest $\mathbf{z}$. Consider the Markov chain that results from running simulated tempering (Algorithm 3) with the collapsed Gibbs transition kernel on an internal annealing (Equation 3.6) schedule for $0 = \beta_1 < \ldots < \beta_L = 1$. Let τ_{mix} denote the number of steps required so that the total variation distance to stationarity is at most $1/4$.*

Let $\mathbf{z}^* := \arg\max_{\mathbf{z}' \neq \mathbf{z}} \{T_L(\mathbf{z}' \mid \mathbf{z})\}$ denote the destination label that maximizes the probability of transitioning away from $\mathbf{z}$. We assume there exists some subset of labels $\mathbf{Z}'$ satisfying

$$[SS_{\mathbf{z}'}] \geq [SS_{\mathbf{z}^*}], \tag{3.15}$$

$$N_{\mathbf{z}'} \leq N_{\mathbf{z}^*}, \tag{3.16}$$

for $\mathbf{z}' \in \mathbf{Z}'$ (with $[SS_{\mathbf{z}'}]$ defined in Equation 3.14), and

$$\sum_{\mathbf{z}' \in \mathbf{Z}'} p_L(\mathbf{z}' \mid \mathbf{x}) \geq 1/10. \tag{3.17}$$

Then, the mixing time of the resulting Markov chain is still exponentially slow in our sepa-

ration parameters u and Δ, with a lower bound given by

$$\tau_{mix} \geq \frac{1}{80L} \min \left\{ \exp\left(\left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(\left[\frac{9 - 40r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\}. \tag{3.18}$$

We note that despite the L in the denominator of Equation 3.18, the actual computational *challenge* is not decreasing linearly in L. Rather, this only shows the bound on the mixing time of the *entire* joint chain. We recall that our goal is to generate samples at the cold temperature target, which only comprise a $1/L$ fraction of the total joint states of the simulated tempering chain. Thus, the actual computational challenge implied by this bound does not depend on L, and the theorem simply shows that the original mixing bottleneck is similarly problematic at *all* temperature levels.

Theorem 3.4.4 is best understood as a result that establishes a set of conditions under which internal annealing *fails* to address the mixing impediment identified by Theorem 2.2.1. The fundamental insight lies in the comparison between the normalized weight of $\mathbf{z}$ and the normalized weight of its escape destination $\mathbf{z}^*$. When the push towards uniformity from internal annealing leads to a tight inverse coupling of these weights (i.e. the growth of one implies the shrinking of the other), then we intuitively cannot fix the weight disparity through temperature transition. In Chapter 4, we consider alternative schedules that could potentially decouple these weight changes.

We also note that this bound may not be quite as broadly representative of the underlying mixing behavior as the original Theorem 2.2.1. Again, the conductance argument only considers the difficulty in escaping a single label (arising from an idealized cluster of data). In the original theorem, the dynamics of escaping a single label were broadly representative of the challenge in escaping a local subset (where the cluster need not be so sharply defined). In Theorem 3.4.4, we should not necessarily assume the same—when are able to transition amongst the subset, the narrower argument we make here may not apply, as we have more flexibility in how we reach the higher temperatures. Further, while the conditions for

identifying such a $\mathbf{Z}'$ are reasonably general, they are not universal—if the posterior weight is dominated by a small number of equally high weight labels, then we do not satisfy the conditions for Lemma 3.4.1.

In summary, despite the critical cautionary note provided by this theorem, the use of internal annealing is often still an effective technique in practical applications to avoid mixing bottlenecks under the collapsed Gibbs sampler. In particular, it offers some notable advantages compared to the typical external annealing methods that are not tailored to the specific properties of the posterior setting. We discussed the key points in Section 3.2.3, but we note a critical additional advantage—it makes the tuning of the algorithm more straightforward. That is, the selection of a viable temperature schedule, and the precise tuning of the Markov transition kernels at each temperature to facilitate mixing, are active areas of research (which vary depending on the application). This tuning is particularly difficult when the interpolating distributions follow the opaque form of a sum over exponentially many densities raised to the power of β. Under internal annealing, we preserve the form of the mixture (which is easier to visualize), and this enables our use of the collapsed Gibbs sampler as a transition kernel tailored to the structure of the problem.

In Section 3.4.3, we provide empirical simulations to illustrate how this technique is able to handle certain mixing bottlenecks in practice. However, this theorem proves we cannot rely on internal annealing to achieve rapid mixing in *all* cases, and that the challenge does not simply hinge on choosing the optimal temperature schedule—we confront a more fundamental barrier to mixing. This motivates our analysis of an alternative annealing scheme in Chapter 4, which has the potential to avoid this bottleneck.

3.4.3 Empirical Simulations

Despite the note of caution provided by Theorem 3.4.4, in practice, internal annealing is often a straightforward way to address practical mixing bottlenecks. In this section, we provide a simple example through empirical simulation. In short summary, we consider

the original experiment measuring the exponential relationship between u^2 and mixing time (Section 2.2.3), and show that the application of internal annealing is well-suited to improve this mixing behavior.

It is important to clarify that this is meant as an illustrative *demonstration*, and we will not overstate the implications of these narrow simulations. For example, we would hesitate to use simulations to conclusively argue that simulated tempering is infeasible, as it is difficult to know whether the tuning of the implementation (in particular, the count and spacing of the temperature schedule) is at fault. Rather, this example helps to illustrate the typical behavior.

We mirror the original three-cluster experiment with varying u^2 (Section 2.2.3), but now we run the simulated tempering algorithm until convergence is reached (rather than the collapsed Gibbs sampler). We build the chain using internal annealing with a linear inverse temperature schedule for $L = 5$ (the full experimental specification is provided in Appendix C.2.3). In Figure 3.2, we again plot the mean count of iterations until convergence is reached on the log scale for each level of u^2, and display two sets of results—the original results from the collapsed Gibbs sampler simulations, and the new results from the internal annealing simulations.

This experiment is *not* capable of drawing a precise comparison between the efficacy of the two methods (e.g. we do not include the process of estimating normalizing constants, and we are conflating different types of "iterations")—rather, we use it simply to characterize the general behavior of each. The rate of exponential growth observed under the collapsed Gibbs sampler was high enough that given our available computational resources, levels of u^2 above 6 soon became intractable, while smaller levels of u^2 were trivial to run. Under simulated tempering, the growth was comparatively quite slow, and there was relatively minimal difference in the practical difficulty of running the chain until convergence (i.e. among this whole range of u^2 inputs, the mean iteration count only varies from $10^{4.0}$ to $10^{4.6}$). Thus, for these experimental settings, the mixing bottleneck is relatively trivial

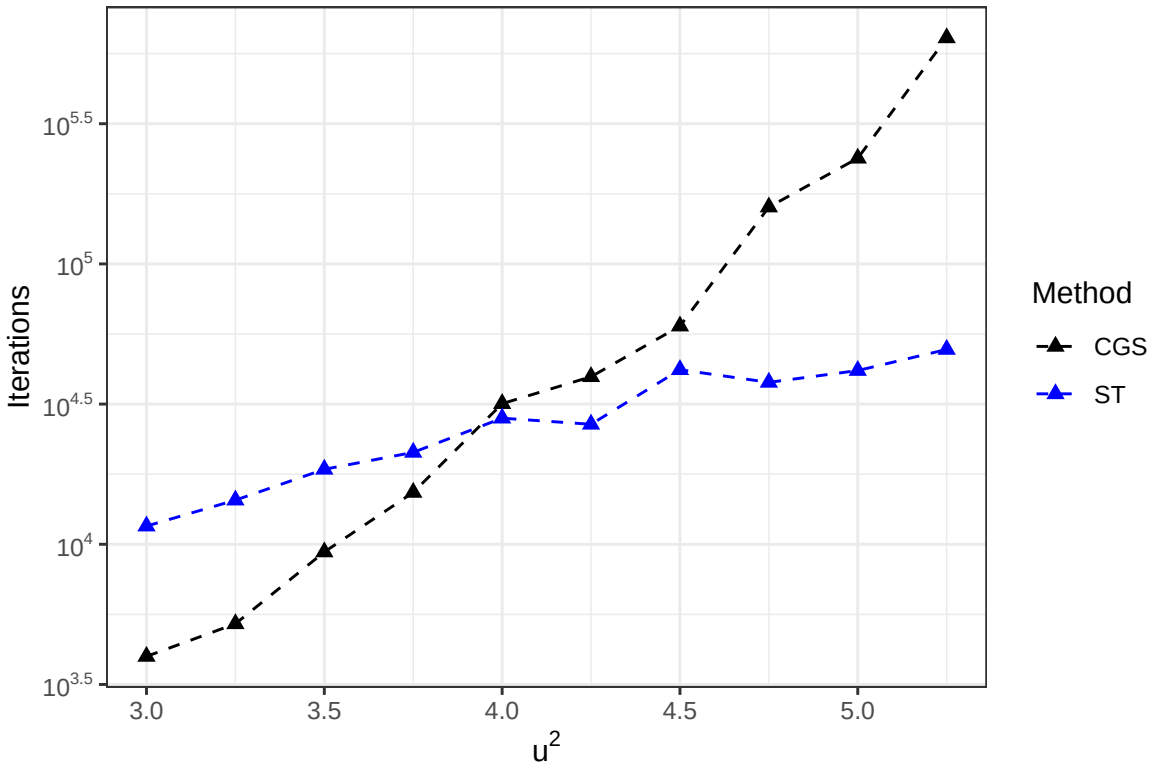


Figure 3.2: The mean number of iterations until convergence is reached (the vertical axis is defined on a log scale) for varying choices of u^2, under two algorithm types. "ST" = simulated tempering (via internal annealing), and "CGS" = the collapsed Gibbs sampler (reproducing the data from Figure 2.3a). This demonstration is not intended as a precise comparison between the run times of the two methods, rather it illustrates their individual behavior. See Appendix C.2 for details on methodology.

to address via simulated tempering, and essentially no careful tuning was required for its implementation. This result is unsurprising, as the actual requirements for mixing are quite light—we simply need to ensure that there is an occasional transfer between the two major isolated regions. This experiment serves as a practical illustration of that intuitive point.

We reiterate that this demonstration is not equipped to draw broader inferences about the mixing properties. For example, the plot appears to suggest steady exponential growth in the mixing time under simulated tempering (albeit with a gentle slope), but it is difficult to be as confident in the result—this only measures the growth under a *fixed* annealing schedule, and it would be necessary to tune the algorithm to adapt to more challenging settings. Thus, these simulations should only be viewed as an illustration of the premise, and in particular the simplicity of its implementation (as no careful tuning was required to facilitate mixing for this toy example).

3.5 Proofs for Chapter 3

3.5.1 Proofs for Section 3.3

Proof of Lemma 3.3.1. We consider a single posterior mixture component $p_\beta(\mathbf{x} \mid \theta, \mathbf{z})p(\theta)$ from Equation 3.5, and recall that the prior on the component center is normal, with mean zero and variance σ^2/α,

$$p(\theta) = \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2/\alpha}\|\theta\|^2\right).$$

For the conditional likelihood $p_\beta(\mathbf{x} \mid \theta, \mathbf{z})$, we mirror our earlier work (Equation 1.6), but now include the inverse temperature β. This scales the likelihood variance, and is included as a subscript in the fixed likelihood term $p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z})$ (which we can otherwise leave unspecified,

as before).

$$
\begin{aligned}
&p_\beta(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta) \\
&\quad = \left[p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2/\beta} \right)^{\frac{N_\mathbf{z} d}{2}} \exp\left(-\frac{\beta}{2\sigma^2} \left[\sum_{i:z_i=1} \|x_i - \bar{x}_\mathbf{z}\|^2 + N_\mathbf{z} \|\bar{x}_\mathbf{z} - \theta\|^2 \right] \right) \right] \\
&\qquad \times \left[\left(\frac{1}{2\pi\sigma^2/\alpha} \right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2/\alpha} \|\theta\|^2 \right) \right]
\end{aligned}
$$

We again complete the square (mirroring our work in Section 1.6.1), and factor out the term that depends on θ.

$$
\begin{aligned}
&= p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2/\beta} \right)^{\frac{N_\mathbf{z} d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha} \right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{\beta}{2\sigma^2} \left[\sum_{i:z_i=1} \|x_i - \bar{x}_\mathbf{z}\|^2 - \frac{1}{\frac{1}{\alpha} + \frac{1}{N_\mathbf{z}}} \|\bar{x}_\mathbf{z}\|^2 \right] \right) \\
&\qquad \underbrace{\times \exp\left(-\frac{1}{\frac{2\sigma^2}{\beta N_\mathbf{z} + \alpha}} \left\| \theta - \frac{\beta N_\mathbf{z}}{\beta N_\mathbf{z} + \alpha} \bar{x}_\mathbf{z} \right\|^2 \right)}_{\propto \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\beta}, \tilde{\sigma}^2_{\mathbf{z},\beta} I)}
\end{aligned}
$$

We identify a Gaussian dependence on θ, with mean $\tilde{\mu}_{\mathbf{z},\beta} := \frac{\beta N_{\mathbf{z}}}{\alpha+\beta N_{\mathbf{z}}}\bar{x}_{\mathbf{z}}$, and variance $\tilde{\sigma}^2_{\mathbf{z},\beta} := \sigma^2/(\alpha+\beta N_{\mathbf{z}})$. The terms that do not depend on θ form the posterior label weight.

$$
\begin{aligned}
&= p^{(0)}_{\beta}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2/\beta}\right)^{\frac{N_{\mathbf{z}}d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \left(2\pi\tilde{\sigma}^2_{\mathbf{z},\beta}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{\beta}{2\sigma^2}\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 - \frac{1}{\frac{1}{\alpha}+\frac{1}{N_{\mathbf{z}}}}\|\bar{x}_{\mathbf{z}}\|^2\right]\right) \\
&\qquad \times \left(\frac{1}{2\pi\tilde{\sigma}^2_{\mathbf{z},\beta}}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\tilde{\sigma}^2_{\mathbf{z},\beta}}\|\theta - \tilde{\mu}_{\mathbf{z},\beta}\|^2\right) \\
&= p^{(0)}_{\beta}(\mathbf{x} \mid \mathbf{z}) \left(\frac{1}{2\pi\sigma^2/\beta}\right)^{\frac{N_{\mathbf{z}}d}{2}} \left(\frac{\alpha}{\alpha+\beta N_{\mathbf{z}}}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{\beta}{2\sigma^2}\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}}+\frac{1}{\alpha}}\|\bar{x}_{\mathbf{z}}\|^2\right]\right) \\
&\qquad \times \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\beta}, \tilde{\sigma}^2_{\mathbf{z},\beta} I)
\end{aligned}
$$

Summing these posterior components over all labelings $\mathbf{z}$ produces the formula written in Equation 3.6.

□

Proof of Lemma 3.3.2. This computation largely mirrors the derivation (in Section 1.6.2) of the original collapsed Gibbs transition probabilities (Lemma 1.5.1). We need not reproduce this work in full, rather we simply note where the modified annealing form diverges. Starting from Equation 1.8), we instead observe

$$
\begin{aligned}
\frac{A^1_{\beta}(\mathbf{z}^{[i\to 1]})}{A^1_{\beta}(\mathbf{z}_{-i})} &= \int p_{\beta}(x_i \mid z_i = 1, \theta) \underbrace{\frac{p(\theta) q^1_{\theta,\beta}(\mathbf{z}_{-i})}{\int p(\theta') q^1_{\theta',\beta}(\mathbf{z}_{-i}) d\theta'}}_{\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{\sigma}^2_{\mathbf{z}_{-i},\beta} I)} d\theta \\
&= \int \mathcal{N}(x_i; \theta, (\sigma^2/\beta) I) \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{\sigma}^2_{\mathbf{z}_{-i},\beta} I) d\theta \\
&= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, (\tilde{\sigma}^2_{\mathbf{z}_{-i},\beta} + \sigma^2/\beta) I) \\
&= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\beta}, \tilde{V}_{\mathbf{z}_{-i},\beta} \sigma^2 I), \qquad (3.19)
\end{aligned}
$$

where once again $\tilde{V}_{\mathbf{z}_{-i},\beta} := \frac{1}{\beta} + \frac{1}{\beta N_{\mathbf{z}_{-i}} + \alpha} = (\tilde{\sigma}^2_{\mathbf{z}_{-i},\beta} + \sigma^2/\beta)/\sigma^2$ is the scaling constant for the posterior predictive variance. This is the only departure from the original derivation, and the formula in the lemma follows accordingly. □

3.5.2 Proofs for Section 3.4

Proof of Lemma 3.4.1.

$$\begin{aligned} p_\beta(\mathbf{z} \mid \mathbf{x}) &= \frac{\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x})}{\sum_{\tilde{\mathbf{z}}} \tilde{p}_\beta(\tilde{\mathbf{z}} \mid \mathbf{x})} \\ &= \frac{\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x})}{\sum_{\mathbf{z}' \in \mathbf{Z}'} \tilde{p}_\beta(\mathbf{z}' \mid \mathbf{x}) + \sum_{\tilde{\mathbf{z}} \notin \mathbf{Z}'} \tilde{p}_\beta(\tilde{\mathbf{z}} \mid \mathbf{x})} \end{aligned}$$

We drop the $\tilde{\mathbf{z}} \notin \mathbf{Z}'$ sum from the denominator, and rewrite the unnormalized weights at inverse temperature β using their growth factors and unnormalized weights under the original posterior (where $\beta = 1$),

$$= \frac{r_{\mathbf{z}}(\beta)\tilde{p}_1(\mathbf{z} \mid \mathbf{x})}{\sum_{\mathbf{z}' \in \mathbf{Z}'} r_{\mathbf{z}'}(\beta)\tilde{p}_1(\mathbf{z}' \mid \mathbf{x})}.$$

By construction, for $\mathbf{z}' \in \mathbf{Z}'$, we have $r_{\mathbf{z}'}(\beta) \geq r^*(\beta)$.

$$\begin{aligned} &= \frac{r_{\mathbf{z}}(\beta)\tilde{p}_1(\mathbf{z} \mid \mathbf{x})}{r^*(\beta) \sum_{\mathbf{z}' \in \mathbf{Z}'} \tilde{p}_1(\mathbf{z}' \mid \mathbf{x})} \\ &= \frac{r_{\mathbf{z}}(\beta) \frac{\tilde{p}_1(\mathbf{z}|\mathbf{x})}{\sum_{\tilde{\mathbf{z}}} \tilde{p}_1(\tilde{\mathbf{z}}|\mathbf{x})}}{r^*(\beta) \sum_{\mathbf{z}' \in \mathbf{Z}'} \frac{\tilde{p}_1(\mathbf{z}'|\mathbf{x})}{\sum_{\tilde{\mathbf{z}}} \tilde{p}_1(\tilde{\mathbf{z}}|\mathbf{x})}} \\ &= \frac{r_{\mathbf{z}}(\beta)p_1(\mathbf{z} \mid \mathbf{x})}{r^*(\beta) \sum_{\mathbf{z}' \in \mathbf{Z}'} p_1(\mathbf{z}' \mid \mathbf{x})} \end{aligned}$$

By construction, $\sum_{\mathbf{z}' \in \mathbf{Z}'} p_1(\mathbf{z}' \mid \mathbf{x}) \geq c^*$.

$$= \frac{1}{c^*} \frac{r_{\mathbf{z}}(\beta)}{r^*(\beta)} p_1(\mathbf{z} \mid \mathbf{x})$$

Thus, we divide both sides by $p_1(\mathbf{z} \mid \mathbf{x})$, and arrive at the stated bound,

$$\frac{p_\beta(\mathbf{z} \mid \mathbf{x})}{p_1(\mathbf{z} \mid \mathbf{x})} \leq \frac{1}{c^*} \frac{r_{\mathbf{z}}(\beta)}{r^*(\beta)}.$$

□

Proof of Lemma 3.4.2. First, we consider the form of the density product $p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z})$ for the fixed component under the labeling $\mathbf{z}$ (with fixed density defined by Equation 3.12). Let $N_{\mathbf{z}}^0 := N - N_{\mathbf{z}}$ denote the number of data points assigned to the fixed component under this labeling. Then, we can rewrite this product as

$$\begin{aligned} p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z}) &:= \prod_{i:z_i=0} p_\beta(x_i \mid z_i = 0) \\ &= \prod_{i:z_i=0} \left(\frac{1}{2\pi\sigma^2 \tilde{V}_{\mathbf{w}}/\beta} \right)^{\frac{d}{2}} \exp\left(-\frac{\beta}{2\sigma^2 \tilde{V}_{\mathbf{w}}} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2 \right) \\ &= \left(\frac{1}{2\pi\sigma^2 \tilde{V}_{\mathbf{w}}/\beta} \right)^{\frac{N_{\mathbf{z}}^0 d}{2}} \exp\left(-\frac{\beta}{2\sigma^2 \tilde{V}_{\mathbf{w}}} \sum_{i:z_i=0} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2 \right). \end{aligned} \tag{3.20}$$

Next, we take the full unnormalized posterior label weight (Equation 3.6)

$$\begin{aligned} \tilde{p}_\beta(\mathbf{z} \mid \mathbf{x}) := p_\beta^{(0)}(\mathbf{x} \mid \mathbf{z}) &\left(\frac{1}{2\pi\sigma^2/\beta} \right)^{\frac{N_{\mathbf{z}} d}{2}} \left(\frac{\alpha + \beta N_{\mathbf{z}}}{\alpha} \right)^{\frac{d}{2}} \\ &\times \exp\left(-\frac{\beta}{2\sigma^2} \left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 \right] \right), \end{aligned}$$

and substitute in Equation 3.20,

$$\begin{aligned} = \left(\frac{1}{2\pi\sigma^2/\beta} \right)^{\frac{Nd}{2}} &\tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}}^0 d}{2}} \left(\frac{\alpha + \beta N_{\mathbf{z}}}{\alpha} \right)^{\frac{d}{2}} \\ \times \exp\Bigg(-\frac{\beta}{2\sigma^2} &\underbrace{\left[\sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\tilde{V}_{\mathbf{w}}} \sum_{i:z_i=0} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2 \right]}_{[\mathrm{SS}_{\mathbf{z}}]} \Bigg). \end{aligned}$$

For clarity, we write $[\mathrm{SS}_{\mathbf{z}}]$ to denote the sum of squares term in the exponential for a given labeling, and the expression simplifies to

$$= \left(\frac{1}{2\pi\sigma^2/\beta}\right)^{\frac{Nd}{2}} \tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}}^0 d}{2}} \left(\frac{\alpha+\beta N_{\mathbf{z}}}{\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{\beta}{2\sigma^2}[\mathrm{SS}_{\mathbf{z}}]\right). \tag{3.21}$$

To compute the unnormalized growth factor, we simply examine the ratio between unnormalized weights (Equation 3.21) at a specified β and at $\beta = 1$. Several terms cancel, and this simplifies to our desired expression,

$$\begin{aligned} r_\beta(\mathbf{z}) &:= \frac{\tilde{p}_\beta(\mathbf{z} \mid \mathbf{x})}{\tilde{p}_1(\mathbf{z} \mid \mathbf{x})} \\ &= \frac{\left(\frac{1}{2\pi\sigma^2/\beta}\right)^{\frac{Nd}{2}} \tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}}^0 d}{2}} \left(\frac{\alpha+\beta N_{\mathbf{z}}}{\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{\beta}{2\sigma^2}[\mathrm{SS}_{\mathbf{z}}]\right)}{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{Nd}{2}} \tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}}^0 d}{2}} \left(\frac{\alpha+N_{\mathbf{z}}}{\alpha}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}[\mathrm{SS}_{\mathbf{z}}]\right)} \\ &= \beta^{\frac{Nd}{2}} \left(\frac{\alpha+\beta N_{\mathbf{z}}}{\alpha+N_{\mathbf{z}}}\right)^{\frac{d}{2}} \exp\left((1-\beta)\frac{[\mathrm{SS}_{\mathbf{z}}]}{2\sigma^2}\right). \end{aligned}$$

□

Proof of Lemma 3.4.3. By Lemma 3.4.2, the ratio between the growth factors is given by

$$\begin{aligned} \frac{r_\beta(\mathbf{z}')}{r_\beta(\mathbf{z}^*)} &= \frac{\beta^{\frac{Nd}{2}} \left(\frac{\alpha+\beta N_{\mathbf{z}'}}{\alpha+N_{\mathbf{z}'}}\right)^{\frac{d}{2}} \exp\left((1-\beta)\frac{[\mathrm{SS}_{\mathbf{z}'}]}{2\sigma^2}\right)}{\beta^{\frac{Nd}{2}} \left(\frac{\alpha+\beta N_{\mathbf{z}^*}}{\alpha+N_{\mathbf{z}^*}}\right)^{\frac{d}{2}} \exp\left((1-\beta)\frac{[\mathrm{SS}_{\mathbf{z}^*}]}{2\sigma^2}\right)} \\ &= \left(\frac{\frac{\alpha+\beta N_{\mathbf{z}'}}{\alpha+N_{\mathbf{z}'}}}{\frac{\alpha+\beta N_{\mathbf{z}^*}}{\alpha+N_{\mathbf{z}^*}}}\right)^{\frac{d}{2}} \exp\left(\frac{(1-\beta)}{2\sigma^2}\left([\mathrm{SS}_{\mathbf{z}'}] - [\mathrm{SS}_{\mathbf{z}^*}]\right)\right). \end{aligned}$$

As $N_{\mathbf{z}^*} \geq N_{\mathbf{z}'}$, for $\beta \in [0, 1]$, the first term is ≥ 1, and by assumption $[\mathrm{SS}_{\mathbf{z}^*}] \leq [\mathrm{SS}_{\mathbf{z}'}]$. Thus, we have the desired inequality between the growth factors,

$$\frac{r_\beta(\mathbf{z}')}{r_\beta(\mathbf{z}^*)} \geq 1.$$

□

Proof of Theorem 3.4.4. By Lemma 3.4.3 and our requirements (Equations 3.15 & 3.16) on the subset $\mathbf{Z}'$, we have $r_{\mathbf{z}'}(\beta) \geq r_{\mathbf{z}^*}(\beta)$ for all $\mathbf{z}' \in \mathbf{Z}'$. Thus, if we cite Lemma 3.4.1 with $r^*(\beta) := r_{\mathbf{z}^*}(\beta)$, $c^* := \frac{1}{10}$, and subset $\mathbf{Z}'$, we have

$$\frac{p_\beta(\mathbf{z} \mid \mathbf{x})}{p_1(\mathbf{z} \mid \mathbf{x})} \leq 10 \frac{r_{\mathbf{z}}(\beta)}{r_{\mathbf{z}^*}(\beta)}. \tag{3.22}$$

Further, the probability of a collapsed Gibbs transition is bounded by the density ratio. The reversibility of our Markov chain implies

$$p_\ell(\mathbf{z} \mid \mathbf{x}) T_\ell(\mathbf{z}^* \mid \mathbf{z}) = p_\ell(\mathbf{z}^* \mid \mathbf{x}) T_\ell(\mathbf{z} \mid \mathbf{z}^*).$$

As the normalizing constants cancel, and $T_\ell(\mathbf{z} \mid \mathbf{z}^*) \leq 1/N$ (the probability of selecting the corresponding index), we observe

$$\begin{aligned} \frac{\tilde{p}_\ell(\mathbf{z}^* \mid \mathbf{x})}{\tilde{p}_\ell(\mathbf{z} \mid \mathbf{x})} &\geq N T_\ell(\mathbf{z} \mid \mathbf{z}^*) \frac{\tilde{p}_\ell(\mathbf{z}^* \mid \mathbf{x})}{\tilde{p}_\ell(\mathbf{z} \mid \mathbf{x})} \\ &= N T_\ell(\mathbf{z}^* \mid \mathbf{z}). \end{aligned}$$

We defined the maximal probability of accepting a collapsed Gibbs transition as $T^*_{\ell,\mathbf{z}} := \max_{\mathbf{z}' \neq \mathbf{z}} \{N T_\ell(\mathbf{z}' \mid \mathbf{z})\}$, and thus by construction $T^*_{\ell,\mathbf{z}} = N T_\ell(\mathbf{z}^* \mid \mathbf{z})$, implying

$$\frac{\tilde{p}_\ell(\mathbf{z}^* \mid \mathbf{x})}{\tilde{p}_\ell(\mathbf{z} \mid \mathbf{x})} \geq T^*_{\ell,\mathbf{z}}. \tag{3.23}$$

Above, we defined $\mathbf{Q} := \{(\mathbf{z}, \ell) : \ell \in [L]\}$ as the subset of joint states for $\mathbf{z}$ at all temperature indices, and this provides the target for our conductance argument. By Equation 3.10,

$$\Phi(\mathbf{Q}) \leq \sum_{\ell \in [L]} \frac{p_\ell(\mathbf{z} \mid \mathbf{x})}{p_L(\mathbf{z} \mid \mathbf{x})} T^*_{\ell,\mathbf{z}},$$

and we substitute in Equation 3.22, translating the continuous β to our discrete schedule (with ℓ denoting β_ℓ and $\beta_L = 1$).

$$\begin{aligned}
&\leq 10 \sum_{\ell \in [L]} \frac{r_{\mathbf{z}}(\beta_\ell)}{r_{\mathbf{z}^*}(\beta_\ell)} T^*_{\ell,\mathbf{z}} \\
&= 10 \sum_{\ell \in [L]} \frac{\frac{\tilde{p}_\ell(\mathbf{z}|\mathbf{x})}{\tilde{p}_L(\mathbf{z}|\mathbf{x})}}{\frac{\tilde{p}_\ell(\mathbf{z}^*|\mathbf{x})}{\tilde{p}_L(\mathbf{z}^*|\mathbf{x})}} T^*_{\ell,\mathbf{z}} \\
&= 10 \sum_{\ell \in [L]} \frac{\frac{\tilde{p}_L(\mathbf{z}^*|\mathbf{x})}{\tilde{p}_L(\mathbf{z}|\mathbf{x})}}{\frac{\tilde{p}_\ell(\mathbf{z}^*|\mathbf{x})}{\tilde{p}_\ell(\mathbf{z}|\mathbf{x})}} T^*_{\ell,\mathbf{z}}
\end{aligned}$$

By Equation 3.23, we have $\frac{\tilde{p}_\ell(\mathbf{z}^*|\mathbf{x})}{\tilde{p}_\ell(\mathbf{z}|\mathbf{x})} \geq T^*_{\ell,\mathbf{z}}$, and thus

$$\begin{aligned}
&\leq 10 \sum_{\ell \in [L]} \frac{\tilde{p}_L(\mathbf{z}^* \mid \mathbf{x})}{\tilde{p}_L(\mathbf{z} \mid \mathbf{x})} \\
&= 10L \frac{T_L(\mathbf{z}^* \mid \mathbf{z})}{T_L(\mathbf{z} \mid \mathbf{z}^*)}.
\end{aligned}$$

This returns us squarely to a computation that was already completed in Section 2.2. Specifically, if we recall our derivation of Equation 2.28 (which bounds the maximal probability of transition), Equations 2.9 & 2.18 reflect this same ratio of transition probabilities. Thus, by Equation 2.28,

$$\leq 20L \max \left\{ \exp\left(- \left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(- \left[\frac{9 - 40 r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\}.$$

To complete the proof, we cite the Jerrum & Sinclair [30] mixing time bound (Equation 2.2), which implies

$$\begin{aligned}
\tau_{\text{mix}} &\geq \frac{1}{4\Phi(\mathbf{Q})} \\
&\geq \frac{1}{80L} \min \left\{ \exp\left(\left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(\left[\frac{9 - 40 r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\}.
\end{aligned}$$

□

Chapter 4

Subsample Annealing for the Mixture Posterior

In Chapter 4, we analyze the implementation and behavior of *subsample annealing*. Given the potential for a mixing bottleneck to persist under temperature annealing (as shown in Theorem 3.4.4), it is natural to consider alternatives. Annealing the posterior through the size of the observed subsample can be independently motivated by its computational benefits (Section 4.1). We introduce *fractional annealing*, a broader framework that allows us to adapt the premise of subsample annealing to the mixture posterior setting (Section 4.2). However, we offer a note of caution, as subsample annealing is highly sensitive to the ordering of the data. We characterize conditions under which the removal of a single datum has such a significant impact on the posterior that the mixing bottleneck must persist, and supplement this with broader empirical evidence of its potential fragility (Section 4.3).

4.1 Introduction

In a Bayesian setting, it is the process of *observing* data that updates our state of belief from the prior distribution to the posterior. Thus, one potential method to anneal the posterior is to use the *count of observed data* to control the distribution. That is, rather

than a continuous inverse temperature β, we use the size of the observed subsample n as the annealing parameter. Informally, we can view the impact of the temperature parameter under internal annealing as reducing the influence of the observed data on the posterior. Thus, a natural alternative is to instead directly limit *which* data influence the posterior. Both methods form a bridge between the prior (where we have no confidence in the observed data) and the posterior (where we have full confidence in the observed data).

For notation, we write the observed data as $\mathbf{x} = \mathbf{x}_{1:N} = (x_1, \ldots, x_N)$, generated with likelihood $p(\mathbf{x}_{1:N} \mid \theta)$, and we wish to draw samples from the target posterior $p(\theta \mid \mathbf{x}_{1:N}) \propto p(\theta)p(\mathbf{x}_{1:N} \mid \theta)$. Then, our annealed posterior is written as

$$p_n(\theta) := p(\theta \mid \mathbf{x}_{1:n}) \propto p(\theta)p(\mathbf{x}_{1:n} \mid \theta), \tag{4.1}$$

for $n \in \{0, \ldots, N\}$ (in the $n = 0$ case, we define the likelihood to be uniform, $p(\mathbf{x}_{1:0} \mid \theta) \propto 1$). If we consider $p_0, p_1, \ldots, p_N$ as a sequence of interpolating distributions, it is easy to see that they might plausibly satisfy our annealing criteria (described in Section 3.1, but now with the index beginning at $n = 0$). When $n = 0$ we have the prior (which should be easy to sample from), when $n = N$ we have our original target posterior, and barring notable outliers, the addition of any single datum should not lead to overly large spacing between adjacent distributions (this is the basic premise, but in later analysis we will see that this can be a dangerous assumption).[1] In practice, we need not define an interpolating distribution for each possible sample size n (we would likely follow a schedule where multiple data are added at each index), but this notation is convenient for illustrating the premise. In the remainder of this section, we discuss the potential motivations for this choice of annealing technique. This lays the foundation for the theoretical analysis of its mixing properties in Section 4.3.

1. We note that this definition implies *recursive* subsamples—i.e. $\mathbf{x}_{1:(n-1)} \subset \mathbf{x}_{1:n}$ for all n. We could instead follow a sequence of subsamples where this is violated, but that would clash with our annealing requirement that adjacent interpolating distributions are "close".

The initial motivation for this approach is simply computational, and does not rely on any subtle analysis. The computational complexity of each query scales with the sample size (which may be large), so the use of subsamples offers an obvious potential speed-up. In this area, the most directly relevant prior work is that of van de Meent et al. [32], who study the use of this framework (which they refer to as "subsample tempering") for parallel tempering and tempered transitions.[2] Their estimate of the potential speed-up (due to the faster queries) relative to temperature annealing is a factor somewhere 2 and 10 (depending on the problem setting), and they provide initial experimental evidence for its efficacy. Their work provides a useful motivating proof-of-concept, although our interest in this chapter will diverge from theirs, as we specialize our implementation and analysis to the Gaussian mixture posterior setting. While our focus is on the use of subsamples within *annealing*, it is worth noting that subsampling is of interest as a broadly important tool for speeding up MCMC computation in a variety of settings (for a recent survey, see Quiroz et al. [43]). Finally, subsample annealing itself was used by Obermeyer et al. [44] as a basis for simulated annealing. As our interest lies in time homogeneous Markov chains, their work does not directly apply, but in summary we can see that this underdeveloped area of study has drawn interest from a range of perspectives.

While our focus is strictly computational, we briefly note that subsample annealing could offer potential inferential benefits as well. Just as temperature annealing has a natural interpretation in physical simulations, the interpolating distributions under subsample annealing have their own natural interpretation—they are exactly the posterior when we only observe a subset of data. It is easy to imagine settings where these intermediate distributions are useful in their own right (perhaps the subsamples are structured as a time series, or perhaps some broader form of online learning). These applications are beyond the scope of this study, but the simple underlying point is that subsample annealing has added motivation as

2. We recall these are two alternative MCMC implementations of the annealing framework, and that their theoretical mixing behavior can generally be assumed to approximately mirror that of simulated tempering.

a natural fit to the setting, rather than being chosen purely for mathematical convenience.

Before starting our analysis, we highlight two key concerns. First, the choice of *which* data are contained in our size n subsample will prove critical. Above, we have simply written that it follows the ordering of the data indices,[3] but this is an important question for any implementation (and it will be the focus of our later analysis). Second, if we define the collapsed Gibbs sampler directly on the subsample annealing posterior (Equation 4.1) for each n, then the state space of the labels changes with the annealing parameter.

That is, if we consider some simulated tempering chain that operates in the state space of the parameters θ (e.g. using the standard Gibbs sampler of Algorithm 1 as a transition kernel), we can readily use interpolating distributions created via subsample annealing with no other adjustments required. However, if we consider some simulated tempering chain operating in the state space of the *labels* (i.e. using the *collapsed* Gibbs sampler as the transition kernel), we cannot directly use a subsample annealing schedule without additional modification, as the simulated tempering annealing index transitions assume that the state space is the same at all indices. In Section 4.2, we will introduce the *fractional annealing* framework as a method for implementing subsample annealing (which allows us to again define our simulated tempering chain directly on the state space of the labels). However, first (in Section 4.1.1) we use our earlier analysis to build our intuition for the properties of subsample annealing.

4.1.1 Graph-based Analysis

Subsample annealing is a topic of independent interest (as described in Section 4.1), but we can supplement this motivation through an intuitive analysis of the mixing arguments described in Chapter 3. In short, the likely paths of flow that emerge under subsample annealing diverge from those under temperature annealing. This is appealing because it

3. We interchangeably refer to this issue as either the *ordering* of the data, or the *composition* of the subsamples, which are equivalent.

has the potential to avoid the mixing bottlenecks associated with temperature annealing (Theorem 3.4.4), but other issues can arise in their place.

To clarify the structure of this chapter, Sections 4.2 - 4.4 will introduce and analyze our implementation of subsample annealing, and the discussion in this section (Section 4.1.1) is not fundamentally required for that work. Rather, the purpose is to provide motivation and context. This is not based on rigorous proof, and instead it draws on the framework of graph-based analysis described in Section 3.2. We use this framework to build our intuition for the setting—in particular, it offers a perspective on the differences in the underlying structure between the two annealing approaches.

The fundamental challenge of the mixture setting lies in the transfer of information between isolated mixture components, as locally-based MCMC techniques cannot easily traverse the low-density valleys that separate the individually unimodal regions. The high level premise of simulated tempering is that it enables new paths between otherwise isolated regions of the target state space through an auxiliary random variable (the annealing index). In Section 3.2, we articulate this through *graph-based analysis*. As a brief reminder (with full details contained in the earlier section), we use a weighted graph to encode the viable flow between the mixture components. The graph nodes represent mixture components at a given annealing index, the node weights correspond with the distribution of the labels under the annealed posterior, and the edges represent the flow tracked by our analysis.

The graph that encodes the flow for the standard simulated tempering premise (including the work of Ge et al. [25] and Woodard et al. [15] studied in Section 3.2) is shown in Figure 4.1 (this is a reproduction of the earlier Figure 3.1). The vertical edges indicate that for a small enough gap in temperature, there should be flow between the adjacent temperature indices. The horizontal edges at the highest temperature ($\ell = 1$) indicate that there is ample flow throughout the state space under the rapidly mixing base distribution. However, we omit the horizontal edges at other temperatures, because we cannot assume that our transition kernel can traverse the low-density valley if the components are well-separated. Again, this

is a model of the reliable flow—some nominal trickle will cross even the deepest valley, but we cannot assume it is enough to enable rapid mixing. This graph is the natural articulation of the simulated tempering premise—we assume that traversal between previously separated points is enabled by exploration at the rapidly mixing high temperature distribution.

The core work of Ge et al. [25] and Woodard et al. [15] lies in their technical arguments proving that the mixing properties of a hypothetical chain defined on this graph correspond with the mixing properties of the true simulated tempering chain. Crucially, for our internally annealed mixture posterior, we have the luxury of operating *directly* on the state space of the labels (through the collapsed Gibbs sampler), and require no further technical argument to make the connection.

It is illustrative to use this perspective to frame our earlier theoretical analysis. That is, Theorem 2.2.1 characterizes conditions where the collapsed Gibbs sampler will struggle to escape from a mixture component, creating a mixing bottleneck. This mirrors the premise of Figure 4.1—we do not assume we can rely on horizontal edges at cold temperatures, as the components may be well-separated. Then, in Theorem 3.4.4, we analyze the conductance of a specified subset, and demonstrate a mixing bottleneck. This comprehensive examination (i.e. it includes all potential horizontal transitions to other components at all temperatures) is necessary for a rigorous proof of slow mixing. However, it is instructive to consider an informal analysis of rapid mixing through the graph-based framework of Figure 4.1. If we restrict our attention to just the flow that passes through those included edges, *could* this be sufficient to facilitate rapid mixing?

While Ge et al. [25] and Woodard et al. [15] are able to place assumptions on their generic mixtures that ensure that the flow following this graph *is* sufficient for rapid mixing, the exponential component count of the mixture posterior appears to be problematic for this approach. Under the base distribution ($\ell = 1$), all nodes have uniform weight, whereas at the cold temperature target, we have observed that "good" labels tend to be exponentially heavier than "bad" labels. Thus, under such a weighted graph (Figure 4.1), the path between

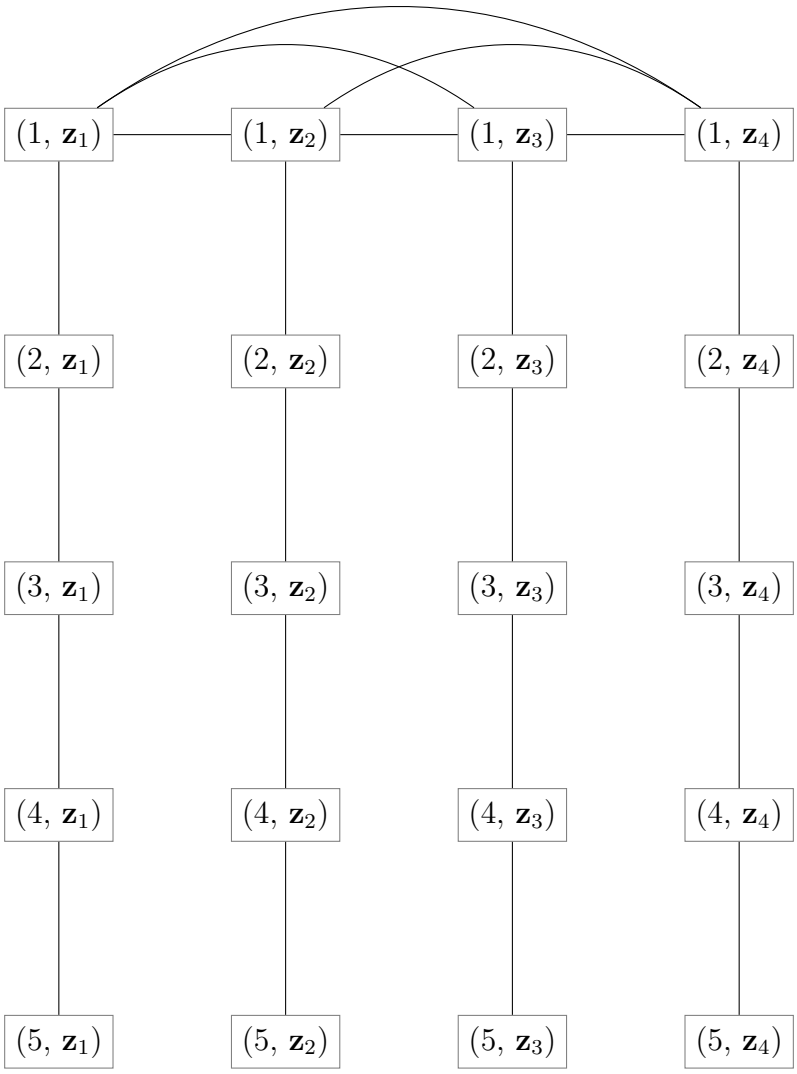


Figure 4.1: The simulated tempering premise (for generic mixtures), encoded as a graph. Each $(\ell, \mathbf{z})$ node represents a duple of temperature index and mixture component (with $L = 5$ and $\mathcal{Z} := \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$). The set of edges models the flow in the simulated tempering chain that we can reliably use in our analysis. This is a reproduction of Figure 3.1, included for convenience.

high weight labels (at the cold target) must pass through the exponentially low weight labels at the base distribution. Or, to view the issue from a different perspective, we note that transitions at the high temperature base distribution are uniform. Thus, if the target posterior is dominated by a small number of labels, it will take exponentially long to "find" such labels through uniform transitions.

This is an intuitive argument for why the set of paths shown in Figure 4.1 is unlikely to be sufficient in demonstrating rapid mixing, not a rigorous proof that the mixing of the true chain must be slow (such a proof would require us to consider all possible horizontal transitions at all temperatures, as shown in our proof of Theorem 3.4.4). But the broader insight from this perspective is that when the component count is exponential, we should not hope to rely *solely* on the transitions between mixture components at the highest temperature distribution for exploration. Increasing the temperature pushes the mixture weights towards uniformity, facilitating (horizontal) movement between the labels, but uniformity among exponentially many labels makes it difficult to "find" the important ones. Thus, we wish to consider annealing techniques enabling paths of flow that follow a *different* structure than the graph shown in Figure 4.1

In the generic mixture setting, the components are essentially independent, and thus a structure of flow that looks like Figure 4.1 is seemingly necessary (we can only assume viable transitions once annealing fully removes the barriers of separation). However, the components of the mixture posterior need not be viewed as independent—their structure is governed by the underlying latent variable framework. Intriguingly, subsample annealing provides a potential restructuring of the viable paths for the simulated tempering chain, one that is shaped by that latent variable framework.

A subsample of size n implies 2^n distinct posterior labels, and we can imagine that the addition of a new datum "splits" each of these labels in two—defining a hypothetical branching binary tree based on this parent-child structure. As a parent and its child will only differ by the removal of a single datum from the observed data, we might typically

expect them to be "close". This is shown in Figure 4.2, where each node represents a single label vector $\mathbf{z}$ for $N = 3$, and the edges display the parent-child relationships. This binary branching tree offers a new set of potential vertical connections, governed by the omission of data to reveal shared ancestry. This is not intended to be comprehensive (there are other ways in which labels might also be "close"), but it reveals a new structure for plausible paths of flow which can be studied. Under Figure 4.1, we assumed the only way to reach an isolated label was through horizontal transitions at the highest temperature, whereas under Figure 4.2, the labels are iteratively built, datum by datum, guided by the latent variable structure. The critical implicit assumption is that the omission of data offers sufficient control over the "closeness" of these labels—this will be the focus of our conductance analysis in Section 4.3.

Thus, the premise of the conductance argument in Theorem 3.4.4 does not apply under this restructured setting. However, while the informal intuition behind this approach is promising, it implies the potential for *different* problematic bottlenecks to arise due to the label weights. In particular, depending on the ordering of the data, it is quite possible for a low weight parent to beget a high weight child, which would impede flow under a binary branching tree. For a toy example, we imagine two symmetric data points, both equidistant (and very far) from the origin, where the fixed density is centered. The $(0, 1)$ and $(1, 0)$ labels will have equal and high weight, as they equivalently provide the best fit to the observed data. In contrast, the (0) label (which is the parent of the $(0, 1)$ label) will be very low weight (the single observed data point is a poor fit for the fixed density, and thus it was likely generated by the variable Gaussian). While the $N = 2$ case is trivial, this same behavior is unavoidable for similarly separated data. In Section 4.3, we analyze this concern, but first, we must introduce the fractional annealing framework, which will allow us to implement simulated tempering under subsample annealing while using the collapsed Gibbs sampler.

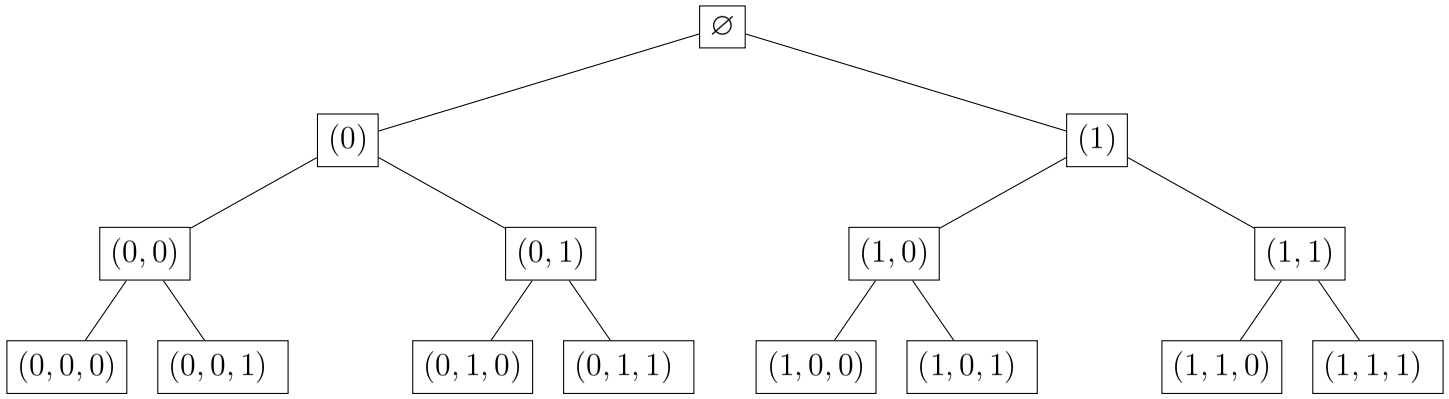


Figure 4.2: Imagining the likely "connections" between the $\mathbf{z}$ labels as a branching tree, in the $N = 3$ case. Each vertical level denotes a subsample size $n \in \{0, 1, 2, 3\}$, and each node denotes a binary label vector $\mathbf{z}$ (with length depending on its level n). Each parent-child pair only differs by the removal of a single datum, and thus we might plausibly expect that they are generally "close" (illustrated by a connecting edge).

4.2 Fractional Annealing

In Section 4.1, we noted that we cannot directly define simulated tempering for subsample annealing on the state space of the labels, because the state space itself changes with the sample size. The basic premise of the fix is straightforward—we simply define the posterior over the full state space of 2^N labels for *all* subsample sizes n, and "ignore" all data outside of the specified subsample. However, this perspective suggests a more general framework—if a subsample is defined by the inclusion of data, we can easily define their *fractional* inclusion. This dovetails neatly with the premise of temperature annealing, where a hotter temperature flattens the likelihood component densities, which weakens the impact of the data. We can view both temperature and subsample annealing as examples of the same broader framework that controls the *inclusion* of the observed data in the model.

This premise requires individualized control over each datum. Thus, rather than use a continuous temperature β or subsample size n, our annealing parameter will be a *vector* $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_N)$, with an annealing value $\beta_i \in [0, 1]$ for each datum representing its fractional inclusion in the model. Specifically, we change the implied generative model for our data to now follow the variable density

$$p_{\beta_i}(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, (\sigma^2/\beta_i)I).$$

Throughout this chapter, we explicitly define the $\beta_i = 0$ case to be the improper uniform density. The resulting posterior matches our intuition for this fractional inclusion—as β_i decreases, it attenuates the influence of x_i on the posterior distribution of θ, until at $\beta_i = 0$ it is fully ignored. Thus, subsample annealing and temperature annealing are both examples of this fractional annealing framework, they just differ on the allowable domain of the parameter vector. Fractional annealing allows $\boldsymbol{\beta}$ anywhere within the N dimensional hypercube, while subsample annealing restricts it to the hypercube corners (the binary vectors $\boldsymbol{\beta} \in \{0, 1\}^N$), and temperature annealing restricts it to a single line (satisfying $0 \leq \beta_i = \beta_j \leq 1$ for all i, j).

Fractional annealing is the natural method to anneal the mixture posterior given the constraint of preserving the latent variable structure. As the posterior is shaped by the effect of observing each individual datum, it offers precise control over the construction. This flexibility has intriguing theoretical implications—the mixing bottleneck under temperature annealing (identified by the conductance argument of Theorem 3.4.4) arises due to the assumption that we treat each datum the same (thus coupling the weight changes), and such bottlenecks could potentially be avoided. It also has practical advantages—one frustration with tuning subsample annealing is that when the sample size is small, the discrete parameter n is not sufficiently granular, which complicates our spacing of the interpolating distributions. Thus, even when implementing subsample annealing, it may be convenient to "smooth out" the schedule through the fractional inclusion of data (i.e. we "ramp up" to their full inclusion). This should be viewed as a practical convenience rather than a major theoretical change, but it proves useful for our simulations in Section 4.3.

However, the cost of this flexibility lies in the difficulty of picking the right fractional annealing schedule, given the vastly increased dimension of the potential options. That is, fractional annealing has the *potential* to ameliorate a given mixing bottleneck, but it is difficult to translate this potential into *general* instructions. This complicates our ability to make broad theoretical claims about the mixing behavior. Thus, for the remainder of this chapter, our analysis will focus on the use of *subsample annealing*. As discussed in Section 4.1, this technique has a variety of strong prior motivations, and thus there is significant value in improving our understanding of its mixing behavior.

In summary, we have introduced fractional annealing for two reasons. First, we will use it to *implement* subsample annealing (with binary $\boldsymbol{\beta}$ vectors), as we need a method to preserve the state space of the labels (and the "ramp up" technique will prove convenient for our empirical simulations). More broadly, we believe that fractional annealing provides a promising foundation for future study in its own right. In the literature, the construction of interpolating distributions for annealing typically follows the same narrow techniques (usu-

ally direct exponentiation), with minimal specialization to the setting (beyond the spacing of the temperature schedule). Fractional annealing offers the natural framework for tailoring these interpolating distributions to the specific structure of the Bayesian mixture posterior. Its flexibility complicates our ability to make sweeping claims about its theoretical properties, but its potential lies in its capacity for specialization to the specific demands of an application. Thus, in this chapter, we narrow our focus to its use in subsample annealing (a particularly intriguing specialization), but before we begin that analysis, in Section 4.2.1 we must derive the explicit form of the conjugate posterior.

4.2.1 Conjugate Posterior

The derivation of the conjugate posterior mirrors our earlier work. To avoid division by zero, it is convenient to define $S_{\boldsymbol{\beta}} := \{i : \beta_i > 0\}$ as the subset of "included" data indices. Further, we assume that each fixed component $p_{\beta_i}(x_i \mid z_i = 0)$ is also parametrized by β_i. As before, we initially leave its definition to be flexible, but in our later analysis we mirror the variable component and divide the variance by β_i.

Thus, we define our fractional annealing likelihood as

$$\begin{aligned} p_{\boldsymbol{\beta}}(\mathbf{x}|\theta) &\propto \prod_{i \in S_{\boldsymbol{\beta}}} \frac{1}{2} \left[p_{\beta_i}(x_i|z_i = 0) + p_{\beta_i}(x_i \mid \theta, z_i = 1) \right] \\ &\propto \sum_{\mathbf{z}} p_{\boldsymbol{\beta}}(\mathbf{x}|\theta, \mathbf{z}). \end{aligned}$$

We use this annealed likelihood to compute our posterior, which is proportional to

$$p_{\boldsymbol{\beta}}(\theta|\mathbf{x}) \propto p_{\boldsymbol{\beta}}(\mathbf{x}|\theta)p(\theta) \propto \sum_{\mathbf{z}} p_{\boldsymbol{\beta}}(\mathbf{x}|\theta, \mathbf{z})p(\theta). \tag{4.2}$$

We derive the full formula for the conjugate fractional posterior in Lemma 4.2.1. For nota-

tion, given a labeling $\mathbf{z}$, we define

$$N_{\mathbf{z},\boldsymbol{\beta}} := \sum_{i:z_i=1} \beta_i,$$
$$\bar{x}_{\mathbf{z},\boldsymbol{\beta}} := \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}}} \sum_{i:z_i=1} \beta_i x_i,$$

as the fractional annealing equivalents of the sample size and sample mean (now suitably weighted), and define

$$S_{\mathbf{z},\boldsymbol{\beta}} := \{i : z_i = 1, \beta_i > 0\},$$

as the set of included data indices assigned to the variable component (which we again use to avoid division by zero).

Lemma 4.2.1. *For the Bayesian mixture model under fractional annealing described in Section 4.2, the formula for the conjugate posterior is given by*

$$p_{\boldsymbol{\beta}}(\theta|\mathbf{x}) \propto \sum_{\mathbf{z}} \tilde{p}_{\boldsymbol{\beta}}(\mathbf{z} \mid \mathbf{x}) p_{\boldsymbol{\beta}}(\theta \mid \mathbf{z}, \mathbf{x}), \tag{4.3}$$

where,

$$\tilde{p}_{\boldsymbol{\beta}}(\mathbf{z} \mid \mathbf{x}) = \left[\prod_{i:z_i=0} p_{\beta_i}(x_i \mid z_i = 0)\right] \left(\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \frac{1}{2\pi\sigma^2/\beta_i}\right)^{\frac{d}{2}} \left(\frac{\alpha}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha}\right)^{\frac{d}{2}} \times \exp\left(-\frac{1}{2\sigma^2}\left[\frac{\alpha N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha}\|\bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2\right]\right),$$

$$p_{\boldsymbol{\beta}}(\theta \mid \mathbf{z}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} I),$$

and,

$$\tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}} := \frac{N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \bar{x}_{\mathbf{z},\boldsymbol{\beta}},$$
$$\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} := \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \sigma^2.$$

This latent structure matches the original posterior, and thus we can compute the collapsed Gibbs transition probabilities in the same fashion. Intuitively, the densities that weight the potential destinations mirror those under internal annealing (Lemma 3.3.2), and we must simply update the parameters.

Lemma 4.2.2. *For the fractional annealing Bayesian mixture posterior, with annealing parameter vector $\boldsymbol{\beta}$, and data index $i \in \{1, \ldots, N\}$ such that $\beta_i > 0$, the collapsed Gibbs conditional transition probabilities are given by*

$$p_\beta(z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \begin{cases} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}}\sigma^2 I)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}}\sigma^2 I) + p_{\beta_i}(x_i \mid z_i = 0)}, & \text{for } z_i = 1, \\ \frac{p_{\beta_i}(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}}\sigma^2 I) + p_{\beta_i}(x_i \mid z_i = 0)}, & \text{for } z_i = 0, \end{cases}$$

for $\tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}} := \frac{N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \bar{x}_{\mathbf{z},\boldsymbol{\beta}}$ and $\tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}} := \frac{1}{\beta_i} + \frac{1}{N_{\mathbf{z}_{-i},\boldsymbol{\beta}} + \alpha}$.

4.3 Subsample Annealing Conductance

The properties of subsample annealing are of particular interest due to the potential computational speed-up (as discussed in Section 4.1). However, for both better and worse, the flow of its simulated tempering chain can have notably divergent properties from that of temperature annealing. In Section 4.1.1, we outlined its *potential* to avoid the mixing bottlenecks that prove problematic under temperature annealing (and highlighted some alternative concerns).

The intuition behind that potential advantage is straightforward. Consider some high

weight label $\mathbf{z}$ which is difficult to "escape" at cold temperatures. This creates the original bottleneck, which we could attempt to address through simulated tempering. The crux of Theorem 3.4.4 is the inverse coupling between the normalized weight of the origin label $\mathbf{z}$, and the normalized weight of the escape destination label $\mathbf{z}^*$. At a hot temperature, it is easy to transition away from the current label, but this is achieved by both *increasing* the weight of the destination label and *decreasing* the weight of the origin label (hence, the "coupling"). The theorem implies that under the specified conditions, any temperature that is sufficiently hot to enable escape must correspond with such a sharp decrease in the origin label's weight, that a mixing bottleneck will emerge.

Under subsample annealing, this coupling logic does not apply. For example, if the origin and destination label diverge on only a single datum, the removal of that datum immediately maximizes the probability of transition (the labels become identical), but the impact on the normalized weight of the origin may be minimal (as it was only a single datum). While this can be easily demonstrated via computational example (comparing the ratios of the weights under temperature annealing and subsample annealing), delving any deeper into the specifics would unnecessarily complicate the key simple point—there is no assumption that the weights must follow the same coupling as before.

This informally illustrates the potential for subsample annealing to avoid this prior bottleneck, but it is not so simple to prove that it actually solves the broader underlying issue. Thus, in summary, we have diverse motivations for the use of subsample annealing, and our goal is to begin to characterize the mixing behavior of its simulated tempering chain. In particular, we will examine the *sensitivity* of the flow to the ordering of the data. In Section 4.3.1, we consider the original mixing bottleneck from Theorem 2.2.1, and establish conditions under which the removal of a single datum causes such a large shift in the annealed posterior that the bottleneck is guaranteed to persist. In Section 4.3.2, we supplement this theoretical analysis with evidence from empirical experiments exploring this sensitivity. Thus, despite the numerous potential advantages of subsample annealing, this analysis offers

a note of caution on its blind application to new settings without due diligence—the very conditions that make mixing difficult in the first place can make the technique particularly sensitive to the ordering of the data.

4.3.1 Conditions for Slow Mixing

The construction of an inverse temperature schedule $\beta_1 < \ldots < \beta_L$ (as in Chapter 3) only requires two choices—the spacing and the count of the inverse temperatures. In contrast, under subsample annealing the data can be removed in any order, and there may be dramatic variation in the shape of the posterior depending on the choice. In this section, we build our understanding of this behavior by assessing the potential impact of the removal of a *single* datum. Specifically, we establish conditions under which the shift in the posterior is so dramatic that this removal creates a mixing bottleneck in the simulated tempering chain.

We again consider the setting of Theorem 2.2.1. As before, we identify an isolated cluster of data, which implies the existence of a labeling $\mathbf{z}$ that is hard to "escape" through a collapsed Gibbs transition. The premise of simulated tempering is that for a sufficiently small subsample, it will be easy to transition to a different label (this may require a subsample size of $n = 0$), and thus we can escape if we reach this annealing index. However, the simple *existence* of such a path is not enough to ensure the mixing of the chain—the path must have sufficient *capacity* for the volume of flow that needs to pass through it. Or, in simple terms, if the removal of data causes the normalized weight of $\mathbf{z}$ to drop too rapidly, then escape will still be difficult. In our analysis, we focus on the removal of a *single datum*, and the conditions that cause the resultant weight change to create a bottleneck.

We make this setting explicit in our notation. We consider the annealing indices L and $L-1$ of a fractional annealing schedule, where $\boldsymbol{\beta}_L := (1, \ldots, 1)$ is the original target posterior, and $\boldsymbol{\beta}_{L-1} := (0, 1, \ldots, 1)$ removes just the x_1 data point from the observed set. We again leverage a conductance argument, but whereas in Theorem 3.4.4 we considered the subset defined by the labeling $\mathbf{z}$ at all annealing indices, here the simpler subset of

$\mathbf{Q} := \{(L, \mathbf{z}), (L-1, \mathbf{z})\}$ will suffice. We need not specify the rest of the annealing schedule $(\boldsymbol{\beta}_{L-2}, \ldots, \boldsymbol{\beta}_1)$, as this conductance argument will show that the weight change from the removal of the x_1 datum is problematic *no matter* how the rest of the schedule is set.

As before, let $T_\ell(\cdot \mid \cdot)$ denote the collapsed Gibbs transition kernel at annealing index ℓ, and define $T^*_{\ell,\mathbf{z}} := \max_{\mathbf{z}' \neq \mathbf{z}} \{N T_\ell(\mathbf{z}' \mid \mathbf{z})\}$ as the corresponding maximal probability of escape from $\mathbf{z}$. We again note that this is *not* growing with the sample size—the transition kernel implicitly includes the $1/N$ probability of selecting any given index, which cancels with the factor of N (i.e. we imagine that we always select the maximizing index). A simple upper bound follows from the definition of conductance (Equation 2.1).

$$\begin{aligned}
\Phi(\mathbf{Q}) &= \frac{\sum_{\mathbf{z}' \neq \mathbf{z}} \pi(L, \mathbf{z}) T_L(\mathbf{z}' \mid \mathbf{z})}{\pi(L, \mathbf{z}) + \pi(L-1, \mathbf{z})} + \frac{\sum_{\mathbf{z}' \neq \mathbf{z}} \pi(L-1, \mathbf{z}) T_{L-1}(\mathbf{z}' \mid \mathbf{z})}{\pi(L, \mathbf{z}) + \pi(L-1, \mathbf{z})} \\
&\leq T^*_{L,\mathbf{z}} + \frac{\pi(L-1, \mathbf{z})}{\pi(L, \mathbf{z})} \\
&= T^*_{L,\mathbf{z}} + \frac{p_{L-1}(\mathbf{z} \mid \mathbf{x})}{p_L(\mathbf{z} \mid \mathbf{x})} \qquad (4.4)
\end{aligned}$$

In short, the conductance of the two-node subset $\mathbf{Q}$ is bounded above by the probability of escape from $\mathbf{z}$ at the original posterior ($T^*_{L,\mathbf{z}}$), plus the ratio between the normalized weights (which measures the capacity for flow through this label). The original premise (which will follow from Theorem 2.2.1) is that $T^*_{L,\mathbf{z}}$ is small (hence the need for simulated tempering), and thus we need only study the ratio of normalized weights. While we have chosen the intuitive framing where L is the target and $L-1$ removes the first datum, this is broadly revealing of the properties of the posterior under subsample annealing. We could apply a similar analysis to any ℓ and ℓ' which differ by the omission of a datum (the resulting mixing bound is simply most intuitive when we focus on the target L), and the behavior when removing a singular datum illustrates similar dynamics when removing a larger subsample.

While we frame this using fractional annealing, we are specialized to a specific case of subsample annealing, involving the posterior given $\mathbf{x}$ and the posterior given $\mathbf{x}_{-1}$. Thus, it is convenient to translate our notation into a more familiar form that omits the use of $\boldsymbol{\beta}$ (i.e.

we manually write out the data indices, rather than referencing $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}_{L-1}$ throughout). We recall the setting of Theorem 2.2.1. We need not reproduce the full setup (as some details are not relevant to any of the new work), but we again note the key notation. For the subset of data assigned to the variable component under $\mathbf{z}$, let $N_{\mathbf{z}}$ denote its sample size, $\bar{x}_{\mathbf{z}}$ denote its sample mean, and now let $N_{\mathbf{z}}^0 := N - N_{\mathbf{z}}$ denote the sample size assigned to the *fixed* component. For some previously identified label $\mathbf{w}$, the fixed component density is defined as $p(x_i \mid z_i = 0) := \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, \tilde{V}_{\mathbf{w}}\sigma^2 I)$, where $\tilde{\mu}_{\mathbf{w}} := \frac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}}\bar{x}_{\mathbf{w}}$ and $\tilde{V}_{\mathbf{w}} := 1 + \frac{1}{N_{\mathbf{w}} + \alpha}$. Under this construction, we can compute $\tilde{p}_L(\mathbf{z} \mid \mathbf{x})$ and $\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x})$, the unnormalized posterior weights for $\mathbf{z}$ at the annealing indices L and $L-1$, respectively. These unnormalized weights are the key building blocks for our analysis, and we relegate their full formulae to the proofs at the end of the chapter (Equations 4.14 & 4.15), so that this intuitive argument does not get bogged down in messy notation.

The target of our analysis is the ratio of *normalized* weights, $p_{L-1}(\mathbf{z} \mid \mathbf{x})/p_L(\mathbf{z} \mid \mathbf{x})$. In our proof of Theorem 3.4.4 (i.e. under temperature annealing), we used Lemma 3.4.1 and an analysis of the *unnormalized growth factors* to upper bound the ratio of normalized weights. In this case, we again leverage the growth factors to create our bound, although the structure of the analysis will be different. In the original statement of Lemma 3.4.1, a growth factor was defined on a continuous variable β. Now, we are *only* interested in the growth factor for a single annealing index $L-1$, and we switch our notation to reflect this, defining

$$r_{\mathbf{z}}(L-1) := \frac{\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x})}{\tilde{p}_L(\mathbf{z} \mid \mathbf{x})}.$$

We will still cite Lemma 3.4.1 to translate the analysis of growth factors into a bound on the ratio of normalized weights, the only change is in the notation (as otherwise the statement of the lemma is identical, we need not reproduce it in full). In short summary, the role of $r_{\mathbf{z}}(\beta)$ is now filled by $r_{\mathbf{z}}(L-1)$, again reflecting the natural correspondence between the continuous parameter β and a discretized annealing schedule.

In Lemma 4.3.1, we provide the full formula for the growth factor of an arbitrary labeling $\mathbf{z}$ (it need not be the target labeling we specify in the mixing bound). The comparison between growth factors for different labels will form the basis of the proof. For clarity, we define $[\text{SSD}^1_{\mathbf{z}}]$ as a term measuring the sum of squares difference in the exponential for the label $\mathbf{z}$, given the removal of the 1st datum index. The growth factor analysis will require us to identify a subset of labels with desirable properties—specifically, they must have a larger growth factor than that of our label of interest $\mathbf{z}$.

Lemma 4.3.1. *For observed data* $\mathbf{x}$*, with fixed density based on the subset* $\mathbf{w}$*, and subsample annealing schedule where the* $L-1$ *index simply removes the* x_1 *datum from the sample, the growth factor is given by*

$$r_{\mathbf{z}}(L-1) = \begin{cases} (2\pi\sigma^2)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2}[SSD^1_{\mathbf{z}}]\right) & \text{for } \ z_1 = 1, \\ (2\pi\sigma^2)^{\frac{d}{2}} \tilde{V}_{\mathbf{w}}^{-\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2\tilde{V}_{\mathbf{w}}}\|x_1 - \tilde{\mu}_{\mathbf{w}}\|^2\right) & \text{for } \ z_1 = 0, \end{cases}$$

where,

$$[SSD^1_{\mathbf{z}}] := \|x_1 - \bar{x}_{\mathbf{z}}\|^2 - \left[\frac{\alpha(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2 - \frac{\alpha N_{\mathbf{z}}}{N_{\mathbf{z}}+\alpha}\|\bar{x}_{\mathbf{z}}\|^2\right] - \sum_{\substack{i:z_i=1,\\ i>1}} \left[\|x_i - \bar{x}_{\mathbf{z}_{-1}}\|^2 - \|x_i - \bar{x}_{\mathbf{z}}\|^2\right].$$

We are using the data parameterization of Theorem 2.2.1, and for convenience we briefly reproduce the key notation here. We make one slight modification—if $\mathbf{z}$ denotes the labeling of interest for our analysis (which is difficult to "escape"), we require that $z_1 = 1$ (that is, our annealing step removes a datum that was previously assigned to the variable component under $\mathbf{z}$). We recall the parameters used to characterize the data (originally illustrated in Figure 2.1, with full explanation provided in Section 2.2).

Notation Reminder:

Reproduction of the data setting in Theorem 2.2.1.

$$\begin{aligned}
\delta &:= \max_{i:z_i=1} \|\bar{x}_{\mathbf{z}} - x_i\| \\
p(x_i \mid z_i = 0) &:= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{w}}, \tilde{V}_{\mathbf{w}}\sigma^2 I) \\
u &:= \|\bar{x}_{\mathbf{w}} - \bar{x}_{\mathbf{z}}\| \\
\Delta &:= \min_{i:z_i=0} \|\bar{x}_{\mathbf{z}} - x_i\| \\
r_\delta &:= \frac{\delta}{u} \\
R &:= \max_{i:z_i=0} \frac{\|\bar{x}_{\mathbf{w}} - x_i\|}{\|\bar{x}_{\mathbf{z}} - x_i\|}
\end{aligned}$$

Further, Theorem 2.2.1 requires that the sample sizes $N_{\mathbf{z}}$ and $N_{\mathbf{w}}$ (i.e. the count of data assigned to the variable component under this labeling) be sufficiently large. In short, this limits the impact from removing a single datum on these parameters (e.g. the sample sizes must scale with dimension and the magnitude of the data). For convenience, we make a slight modification to instead require $d + 1$ as a minimum, rather than d in the original theorem (otherwise the requirements are unchanged).

Sample Size Requirement:

For $N^* := \min\{N_{\mathbf{z}}, N_{\mathbf{w}}\}$, we require

$$N^* \geq \max\{d + 1, 9\}, \tag{4.5}$$

and for any index i, we require

$$N^* \geq \begin{cases} \frac{1}{\delta}\|x_i\| + 1 - \alpha & \text{if } z_i = 1, \\ \frac{10\alpha}{R}\frac{\|x_i\|}{\|\bar{x}_{\mathbf{z}} - x_i\|} - \alpha & \text{if } z_i = 0. \end{cases} \tag{4.6}$$

With the setting established, we can clarify the plan for the overall proof. In brief

summary, Theorem 2.2.1 establishes conditions that cause a mixing bottleneck for the collapsed Gibbs sampler. In this setting, we will establish further conditions under which the bottleneck will *persist* despite the use of subsample annealing (implemented via simulated tempering). By the conductance argument of Equation 4.4, the key is to show that the ratio of normalized weights $p_{L-1}(\mathbf{z} \mid \mathbf{x})/p_L(\mathbf{z} \mid \mathbf{x})$ is exponentially small. To do this, we will identify a subset of labels $\mathbf{Z}'$ whose growth factors $r_{\mathbf{z}'}(L-1)$ are significantly larger than that of $r_{\mathbf{z}}(L-1)$, for all $\mathbf{z}' \in \mathbf{Z}'$. This would imply that the unnormalized weight of $\mathbf{z}$ is growing slower than its normalizing constant, and thus its normalized weight is shrinking. If we can show that the weight change is sufficiently severe, then the proof is complete. Thus, the final missing piece in this argument is to define this subset $\mathbf{Z}'$.

Informally, we imagine $\mathbf{Z}'$ as a subset of labels for which the datum x_1 is a poor fit. Thus, the removal of that datum (under annealing) will have a particularly *large* increase on their unnormalized weight (relative to the other labels), implying the desired inequality on the growth factors. To make this concrete, for any label $\mathbf{z}' \in \mathbf{Z}'$, let $\bar{x}_{\mathbf{z}'}$ denote its sample mean. We will require that $\bar{x}_{\mathbf{z}'}$ is sufficiently far from the removed datum x_1. We recall that the premise of the original bottleneck relied on the separation between the sample means of the labels $\mathbf{z}$ (the label that is difficult to escape) and $\mathbf{w}$ (the basis for the fixed component), given by some suitably large $u := \|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{w}}\|$. Thus, for all $\mathbf{z}' \in \mathbf{Z}'$, we will require that $\|\bar{x}_{\mathbf{z}'} - \bar{x}_{\mathbf{w}}\| > u + \delta$. That is, we require that $\bar{x}_{\mathbf{z}}$ must be *further* from this $\bar{x}_{\mathbf{z}'}$ than it is from $\bar{x}_{\mathbf{w}}$, by an additional distance of at least δ. As the datum that is removed via annealing (x_1) is at most distance δ from the sample mean $\bar{x}_{\mathbf{z}}$, this implies that $\|\bar{x}_{\mathbf{z}'} - x_1\| > u$. which ensures that the growth factor is sufficiently large.

The precise technical requirements used for the proof are provided by Equations 4.7-4.9. For all $\mathbf{z}' \in \mathbf{Z}'$, in addition to this requirement on the sample mean $\bar{x}_{\mathbf{z}'}$, we place a familiar (albeit looser) requirement on the minimum sample size $N_{\mathbf{z}'}$, so that the removal of any datum does not have too large an impact on the parameters. Finally, we specify that the total probability mass of the subset $\mathbf{Z}'$ be at least some constant fraction of the total

(we choose 1/10), as required by Lemma 3.4.1 (guaranteeing the growth of the normalizing constant).

Requirements on the Subset $\mathbf{Z}'$:

Let $\mathbf{Z}'$ denote a subset of labelings such that all $\mathbf{z}' \in \mathbf{Z}$ satisfy

$$\|\bar{x}_{\mathbf{z}'} - \bar{x}_{\mathbf{z}}\| \geq \|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{w}}\| + \delta, \tag{4.7}$$

$$N_{\mathbf{z}'} \geq \max\left\{5, d, \tfrac{1}{\delta}\|\bar{x}_{\mathbf{z}'}\| + 1\right\}, \tag{4.8}$$

and such that the total normalized weight of $\mathbf{Z}'$ is originally at least 1/10,

$$\sum_{\mathbf{z}' \in \mathbf{Z}'} p_L(\mathbf{z}' \mid \mathbf{x}) \geq 1/10. \tag{4.9}$$

With this foundation established, we can state the mixing bound in full, and then walk through the underlying argument used in its proof.

Theorem 4.3.2. *Consider the Gaussian mixture posterior that follows the construction of Theorem 2.2.1. As in the original theorem, for observed data* $\mathbf{x}$*, let* $\mathbf{z}$ *and* $\mathbf{w}$ *denote labels such that* $R < \frac{1}{2}$*,* $r_\delta < \frac{9}{40}$*, and whose sample sizes satisfy Equations 4.5 & 4.6. Let* $\mathbf{Z}'$ *denote a subset of labels satisfying Equations 4.7 - 4.9.*

We assume $z_1 = 1$*, and define a fractional annealing schedule such that* $\boldsymbol{\beta}_L := (1, \ldots, 1)$*, and* $\boldsymbol{\beta}_{L-1} := (0, 1, \ldots, 1)$*. Consider the Markov chain that results from running simulated tempering (Algorithm 3) on this annealing schedule. Let* τ_{mix} *denote the number of steps required so that the total variation distance to stationarity is at most* 1/4*.*

Then, the mixing time of the resulting Markov chain is exponentially slow in our separation parameters u *and* Δ*, with a lower bound given by*

$$\tau_{mix} \geq \frac{5}{48} \min\left\{\exp\left(\left[\frac{7 - 14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right), \exp\left(\left[\frac{9 - 40r_\delta}{20}\right]\frac{u^2}{\sigma^2}\right)\right\}.$$

The proof of Theorem 4.3.2 leverages the conductance argument of Equation 4.4. By the

proof of Theorem 2.2.1, we know that $T^*_{L,\mathbf{z}}$ is small, and thus we need only show that the normalized weight ratio is similarly small. We prove this through growth factor analysis. The first step is to upper bound the growth factor for the labeling $\mathbf{z}$, to ensure it cannot be too large.

Lemma 4.3.3. *Given the setting of Theorem 4.3.2, the growth factor for the target label* $\mathbf{z}$ *is bounded by*

$$r_{\mathbf{z}}(L-1) \leq \left(2\pi\sigma^2\right)^{\frac{d}{2}} 2 \exp\left(\left[\frac{5r_\delta^2}{2}\right]\frac{u^2}{\sigma^2}\right).$$

Next, we establish a lower bound for the growth factor of any label $\mathbf{z}' \in \mathbf{Z}'$.

Lemma 4.3.4. *Given the setting of Theorem 4.3.2, the growth factor for any label* $\mathbf{z}' \in \mathbf{Z}'$ *is bounded by*

$$r_{\mathbf{z}'}(L-1) \geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \frac{1}{2} \exp\left(\left[15/32 - r_\delta\right]\frac{u^2}{\sigma^2}\right).$$

By Lemmas 4.3.3 & 4.3.4 the ratio of growth factors $r_{\mathbf{z}}(L-1)/r_{\mathbf{z}'}(L-1)$ is small for any $\mathbf{z}' \in \mathbf{Z}'$, and thus by Lemma 3.4.1, we can bound the ratio of normalized weights $p_{L-1}(\mathbf{z} \mid \mathbf{x})/p_L(\mathbf{z} \mid \mathbf{x})$. This completes the conductance argument.

4.3.2 Empirical Experiments

We can supplement the theoretical analysis in Section 4.3.1 with evidence from empirical experimentation, demonstrating the sensitivity of subsample annealing to the ordering of the data. In particular, we compare the mixing behavior when the subsamples are drawn *randomly*, versus the mixing behavior when they follow a *pre-set* order (chosen to avoid a likely bottleneck). The full specification of this experiment is written in Appendix C.2.4, but the key details are described here.

We consider data comprised by three well-separated clusters of equal sample size (drawn

from a multivariate Gaussian), whose centers form an equilateral triangle (equidistant from the origin). While earlier experiments measured the relationship between cluster separation and mixing time, in this section the precise data arrangement is less important—we simply require a shared setting where mixing is slow, which we can use to compare the efficacy of *four* different MCMC techniques. First, we run collapsed Gibbs sampling, to provide a baseline technique that we do not expect to converge (these clusters are isolated, and escape from a local region is unlikely). Then, we consider simulated tempering under three different implementations of fractional annealing. The first is temperature annealing, where the $\boldsymbol{\beta}_\ell$ vectors are uniform-valued (this is equivalent to the internal annealing formulation of Section 3.3). This provides our second baseline comparison—an annealing method that is well-suited to the setting, and should be able to converge. Sampling in this setting is straightforward as long as the chain is able to occasionally transfer between the three well-separated clusters, and this is a case where temperature annealing will prove effective.

The final two techniques are implementations of subsample annealing, and their behavior is our primary focus. Both follow a schedule with the same subsample *sizes*, but they diverge in the subsample *compositions*. The first technique randomizes the order of the data, whereas the second technique follows a pre-set order, requiring every subsample to contain an *equal* count of data from each of the three clusters. Their comparison allows us to explore the broader concern implied by Theorem 4.3.2. In short, when data is removed from one cluster (but not another), the cluster separation ensures a dramatic shift in the posterior weights of the labels, which makes annealing index transitions difficult. Our pre-set schedule avoids this cluster imbalance, but our random ordering may encounter such a bottleneck.

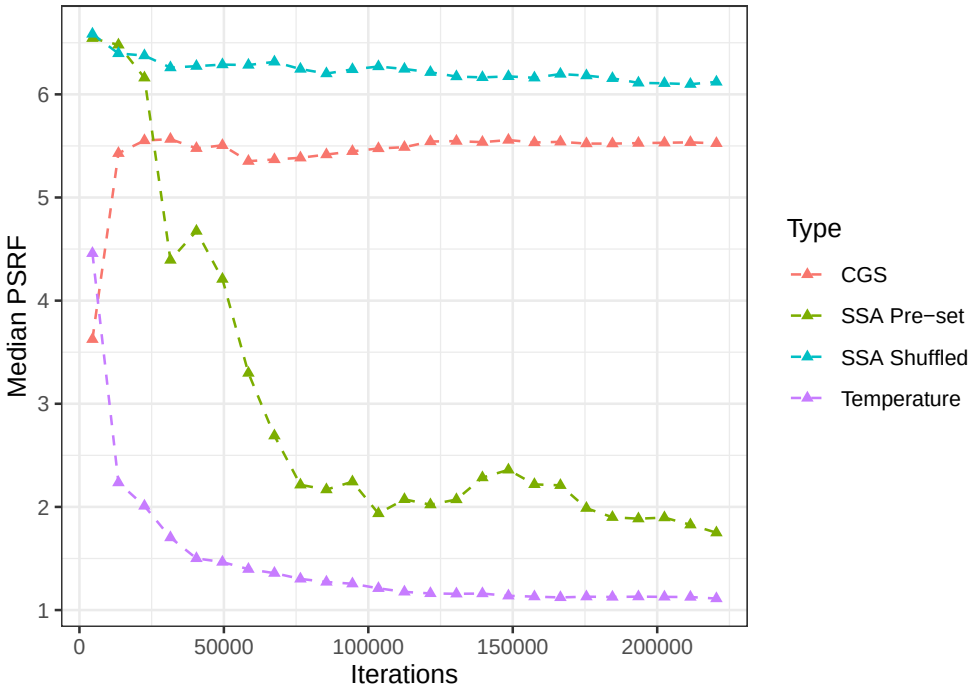
We generate 50 such datasets, and apply the four techniques to each. We track the evolution of the *potential scale reduction factor* (PSRF, introduced in Appendix C.1) as the iteration count grows. The PSRF is our chosen convergence criterion—in earlier simulations, we simply tracked when convergence was reached (requiring a PSRF below 1.10), here we instead track the PSRF itself. The results are shown in Figure 4.3. As expected, the collapsed

Gibbs PSRF is roughly constant, as the well-separated clusters make any escape from the isolated starting region unlikely. Temperature annealing has quickly decreasing PSRF, as this heating schedule is sufficient to enable the simple transfer between isolated regions (and the bottleneck concerns of Theorem 3.4.4 do not apply).

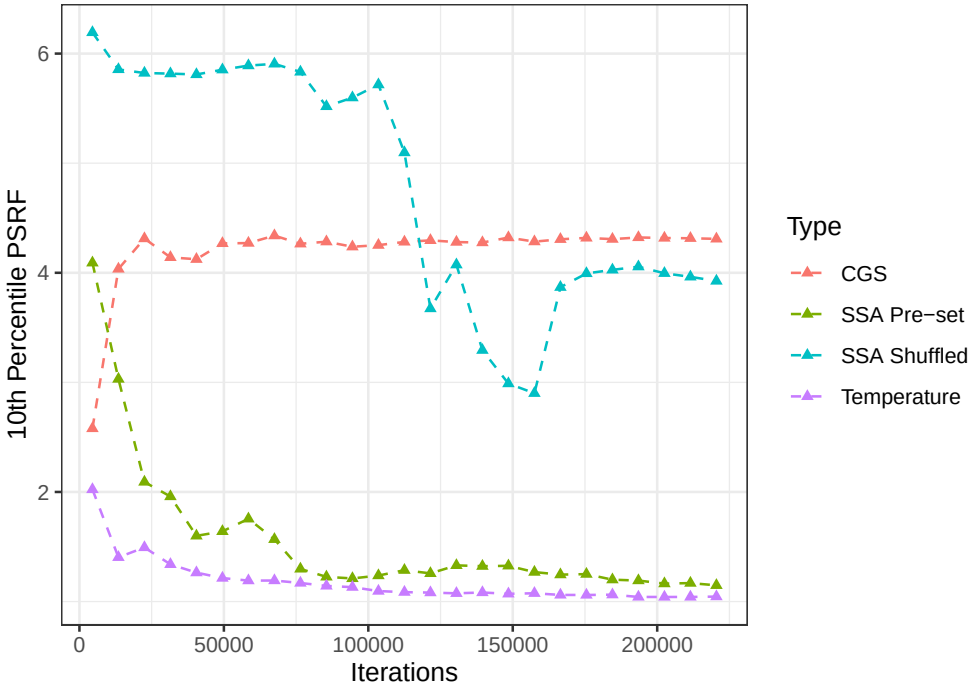
The sensitivity of subsample annealing to the ordering of the data can be observed in the poor performance when the order is randomized ("SSA Shuffled"). This illustrates the insight of Theorem 4.3.2—as the clusters are well-separated, the removal of even a small amount of data causes a dramatic shift in the posterior weights, which creates a mixing bottleneck. While escape from an isolated local region would be possible at a sufficiently small sample size, the transitions required to *reach* that annealing index can be just as difficult. We note that while the median performance of the randomly shuffled chain is worse than the median performance of the collapsed Gibbs sampler, its 10th percentile is perhaps slightly better. This is unsurprising—it implies that the median ordering is poor enough that the chain cannot readily transition to smaller sample sizes (and thus behaves like a slower version of the collapsed Gibbs sampler), but in a minority of cases, the random ordering is sufficient to enable some additional flow.

When we instead follow a pre-set schedule ("SSA Pre-set") which maintains the cluster balance, the annealing index transitions are viable, and the PSRF is steadily decreasing towards convergence. While the median performance is still significantly worse than that of temperature annealing, the 10th percentile shows comparatively little difference between the two. This again illustrates the *sensitivity* of subsample annealing. There is little variation in performance under temperature annealing, while the variability in the randomly generated *data* leads to a wider range of behaviors under subsample annealing (even when the cluster representation is guaranteed to be equal). Thus, when the generated data are favorable, there is little difference between the two techniques, but there is greater potential for problematic datasets under subsample annealing.

We emphasize that these simulations are intended to be illustrative, and do not offer



(a) Median PSRF



(b) 10th Percentile PSRF

Figure 4.3: PSRF Percentiles, by algorithm type. "CGS" = collapsed Gibbs sampler, "SSA Pre-set" = subsample annealing following a pre-set schedule, "SSA Shuffled" = subsample annealing under randomly ordered data, "Temperature" = internal annealing by inverse temperature. The full experiment specification is found in Appendix C.2.4.

comprehensive characterizations of efficacy, particularly in the comparison between temperature and subsample annealing. A core impetus for subsample annealing is the speed-up for each query in large data settings. We only measure the iteration count (not the speed of the queries), nor is the synthetic dataset large enough to make the choice desirable (a formal exploration of computational efficiency in practical settings is beyond the scope of this theoretical analysis). Rather, as subsample annealing is a technique of independent interest in the literature, this analysis provides a note of caution against its blind use without careful consideration. We have demonstrated that the separation conditions that lead to the original mixing bottleneck (and thus our use of annealing in the first place), can make the posterior particularly sensitive to the removal of certain data. Thus, the use of subsample annealing solely for its superior query speed should be given its due scrutiny.

4.4 Variable Schedule

We will conclude this chapter by informally outlining a potential direction for further study. We introduce a technique that provides an example of how our earlier theoretical analysis could guide the development of methods for mixture posterior sampling.

The fundamental allure of fractional annealing lies in its capacity for *specialization*. The canonical annealing implementations (e.g. direct exponentiation) are ill-suited to the unusual properties of the mixture posterior, while fractional annealing is tailored to the latent variable structure. This offers precise control over the shape of the resulting interpolating distributions, which provides the potential to avoid specific bottlenecks. However, concomitant with this flexibility is the vast increase in the range of possible annealing schedules. This poses a challenge—the mixing behavior is highly sensitive to the selection of the annealing schedule, and yet we often lack prior knowledge to help us make this choice. For example, even when we narrow our attention to just subsample annealing, the analysis of Section 4.3 shows that small changes to the ordering of the data (even the removal of a single datum)

can cause a bottleneck.

Given this challenge, we note two particular paths forward. First, fractional annealing can be viewed as a framework that can be tailored to the needs of a certain domain. Thus, the appeal lies in its flexibility, and we are guided to the right choice of annealing schedule by the demands of a specific application. This is intriguing, and worthy of note, but in this theoretical study we will not say much further on the topic (the details are unique to the practitioner). On the other hand, if we lack such specific guidance, the alternative is to develop a technique that is *robust* to a poorly chosen annealing schedule. Thus, to mitigate the impact of a poor choice, a natural solution is to *regularly change* the schedule in use.

We refer to this strategy as a *variable annealing schedule*. Its broad appeal is that we will not become permanently stuck with some poor selection, but perhaps the more intriguing perspective is that this helps to ensure the eventual exploration of the state space. If the current annealing schedule has a mixing bottleneck that precludes flow between two regions, we simply wait for some future schedule where the bottleneck disappears. Intuitively, even if *each* annealing schedule has a bottleneck causing slow mixing, it still might be quite beneficial to follow a variable schedule (as the location of the bottleneck can shift, enabling flow between isolated regions over time). As a useful contrast, we recall that the empirical simulations of Section 4.3.2 included subsample annealing on a randomized ordering of the data. However, each chain followed a *single* randomized ordering—it is no surprise that a bottleneck would arise, and a region of state space would become isolated. If this ordering was regularly shuffled, eventually paths might form between these different regions, allowing for full exploration.

Unfortunately, simulated tempering is highly resistant to the use of a variable annealing schedule. While our analysis has focused on its theoretical mixing properties, the implementation of simulated tempering requires the estimation of the relative normalizing constants for that sequence of interpolating distributions (as introduced in Appendix C.3). Under a variable annealing schedule, this estimation process would need to be repeated with *every*

update to the schedule, which is computationally impractical.

However, simulated tempering is not the only MCMC implementation of the annealing framework, and a natural candidate for a variable annealing schedule is the *tempered transitions* algorithm, originally proposed by Radford Neal [36]. A comprehensive comparison of these annealing techniques is beyond the scope of this brief overview, but in short, simulated tempering and tempered transitions can both be viewed as single-chain implementations of the original parallel tempering premise. While the simulated tempering joint chain is defined on both the state space and annealing index, the tempered transitions chain operates on solely the original state space, and uses the sequence of interpolating distributions to build its Metropolis-Hastings proposal. Intuitively, each update crawls down and up the annealing indices (following a sequence of descending and ascending transition kernels that preserve stationarity for that index), generating a sequence of intermediate states, and the final proposal state is accepted with probability determined by the entire sequence (the full details are explained below). Crucially, this update does *not* require the use of normalizing constants, which enables us to follow a variable schedule. This advantage of tempered transitions was also noted by van de Meent et al. [32] in their subsample annealing empirical simulations, although otherwise their interests diverge from our own.

A full analysis of the properties of tempered transitions is beyond the scope of this initial exploration. It is instructive to first explicitly state the algorithm, before we highlight some critical points.

Tempered Transitions Algorithm:

1. Initialize the simulated tempering chain.

    - Let $\tilde{p}_1, \ldots, \tilde{p}_L$ denote a sequence of unnormalized interpolating distributions, on state space $\mathcal{Y}$.
    - For $\ell \in \{2, \ldots, L\}$, let $\hat{T}_\ell(\cdot \mid \cdot)$ and $\check{T}_\ell(\cdot \mid \cdot)$ denote our ascending and descend-

ing state space transition kernels (respectively), where $\hat{T}_\ell$ preserves invariance for p_ℓ, and $\check{T}_\ell$ preserves invariance for $p_{\ell+1}$.

- Initialize starting state $y^{(0)}$.
- Initialize $t \leftarrow 1$ to mark our current iteration.

2. Perform a tempered transitions update.

   - Generate a candidate state $\check{y}_L$ as follows:
     - Set $\hat{y}_L \leftarrow y^{(t-1)}$.
     - Generate $\hat{y}_{L-1} \sim \hat{T}_{L-1}(\cdot \mid y^{(t-1)})$.
     - Generate $\hat{y}_{L-2} \sim \hat{T}_{L-2}(\cdot \mid \hat{y}_{L-1})$.
     - $\ldots$
     - Generate $\bar{y}_1 \sim \hat{T}_1(\cdot \mid \hat{y}_2)$.
     - Generate $\check{y}_2 \sim \check{T}_1(\cdot \mid \bar{y}_1)$.
     - $\ldots$
     - Generate $\check{y}_L \sim \check{T}_{L-1}(\cdot \mid \check{y}_{L-1})$.
   - Compute the acceptance probability:

$$Q \leftarrow \min\left\{1, \frac{\tilde{p}_{L-1}(\hat{y}_L)}{\tilde{p}_L(\hat{y}_L)} \cdots \frac{\tilde{p}_1(\hat{y}_2)}{\tilde{p}_2(\hat{y}_2)} \frac{\tilde{p}_2(\check{y}_2)}{\tilde{p}_1(\check{y}_2)} \cdots \frac{\tilde{p}_L(\check{y}_L)}{\tilde{p}_{L-1}(\check{y}_L)}\right\}.$$

   - With probability Q, accept the proposed transition, and set $y^{(t)} \leftarrow \check{y}_L$. Otherwise, reject the proposed transition, and set $y^{(t)} \leftarrow y^{(t-1)}$

3. If the convergence criterion is satisfied, halt the algorithm. Otherwise, set $t \leftarrow t+1$, and return to step #2.

This can be easily adapted to fractional annealing—we use the collapsed Gibbs transition rule for both the descending and ascending transition kernels. Crucially, we are free to follow any annealing schedule for these updates, without the need to estimate normalizing

constants. The natural choice would be to simply set a new annealing schedule after a fixed number of tempered transitions updates, although it would be worth considering the potential for an adaptive approach (which modifies the schedule based on the observed behavior).

However, while this approach has intriguing theoretical potential, we must also note the practical concerns that arise in its implementation. A principal challenge of tempered transitions is its *tuning*. Under temperature annealing, the only choice in setting the schedule is the count L and the spacings between the inverse temperatures. Still, there is significant literature devoted to the precise selection of this schedule (e.g. the work of Behrens et al. [45], although most applications will address this topic). Intuitively, as the acceptance probability is a product over a sequence of intermediate states, imprecise tuning (i.e. intermediate proposals that are too aggressive or conservative) tend to lead to acceptance probabilities near 0 or 1. Unfortunately, fractional annealing is particularly sensitive—there is greater flexibility when setting the schedule, and operating in the discrete space offers less fine grained control over the Markov kernels used for the transitions. This issue of tuning is not insurmountable, but it does complicate the immediate application of the premise.

Broadly, the use of tempered transitions with a variable fractional annealing schedule provides a case study in how the earlier theoretical analysis can guide the development of mixture posterior sampling techniques. By characterizing the impediments to mixing (under both Gibbs sampling and the annealing framework), we highlight the importance of specializing methods to this domain. Further work is needed to understand the practical relevance of these mixing impediments, and the viability of potential alternatives.

4.5 Proofs for Chapter 4

4.5.1 Proofs for Section 4.2

Proof of Lemma 4.2.1. Before we begin the derivation, for convenient reference we reproduce the notation introduced for the fractional annealing model. The use of $S_{\mathbf{z},\boldsymbol{\beta}}$ allows us to avoid dividing by zero (i.e. the annealed densities are specifically defined to be the improper uniform when $\beta_i = 0$, and thus are ignored from the likelihood product).

$$\begin{aligned} S_{\boldsymbol{\beta}} &:= \{i : \beta_i > 0\} \\ S_{\mathbf{z},\boldsymbol{\beta}} &:= \{i : z_i = 1, \beta_i > 0\} \\ N_{\mathbf{z},\boldsymbol{\beta}} &:= \sum_{i:z_i=1} \beta_i \\ \bar{x}_{\mathbf{z},\boldsymbol{\beta}} &:= \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}}} \sum_{i:z_i=1} \beta_i x_i \end{aligned}$$

We also note the following weighted sum of squares identity (modified to fit our notation), which will prove useful.

$$\sum_{i:z_i=1} \beta_i \|x_i - \theta\|^2 = \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + N_{\mathbf{z},\boldsymbol{\beta}} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}} - \theta\|^2. \tag{4.10}$$

We begin with the conditional likelihood (which appears in each component of Equation

4.2).

$$
\begin{aligned}
&p_{\boldsymbol{\beta}}(\mathbf{x} \mid \theta, \mathbf{z}) \\
&\quad = \underbrace{\left[\prod_{i:z_i=0} p_{\beta_i}(x_i \mid z_i = 0)\right]}_{p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x}|\mathbf{z})} \left[\prod_{i:z_i=1} p_{\beta_i}(x_i \mid z_i = 1, \theta)\right] \\
&\quad = p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left[\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \mathcal{N}(x_i; \theta, (\sigma^2/\beta_i) I)\right] \\
&\quad = p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left[\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \left(\frac{1}{2\pi\sigma^2/\beta_i}\right)^{\frac{d}{2}}\right] \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1} \beta_i \|x_i - \theta\|^2\right]\right)
\end{aligned}
$$

We cite the weighted sum of squares identity (Equation 4.10), and observe

$$
\begin{aligned}
&= p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left[\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \left(\frac{1}{2\pi\sigma^2/\beta_i}\right)^{\frac{d}{2}}\right] \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + N_{\mathbf{z},\boldsymbol{\beta}} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}} - \theta\|^2\right]\right).
\end{aligned}
$$

We recall that the prior on the component center is normal, with mean zero and variance σ^2/α. We combine the conditional likelihood and prior to compute a single term in the posterior sum (Equation 4.2),

$$
\begin{aligned}
&p_{\boldsymbol{\beta}}(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta) = \\
&\quad p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left[\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \left(\frac{1}{2\pi\sigma^2/\beta_i}\right)^{\frac{d}{2}}\right] \left(\frac{1}{2\pi\sigma^2/\alpha}\right)^{\frac{d}{2}} \\
&\qquad \times \exp\left(-\frac{1}{2\sigma^2} \underbrace{\left[\sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + N_{\mathbf{z},\boldsymbol{\beta}} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}} - \theta\|^2 + \alpha\|\theta\|^2\right]}_{A_1}\right).
\end{aligned} \tag{4.11}
$$

We examine the term in the exponential (A_1), and consider its dependence on θ (by conju-

gacy, we know the posterior component will take the form of a Gaussian).

$$
\begin{aligned}
A_1 &= \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + N_{\mathbf{z},\boldsymbol{\beta}} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}} - \theta\|^2 + \alpha\|\theta\|^2 \\
&= (\alpha + N_{\mathbf{z},\boldsymbol{\beta}})\|\theta\|^2 - 2N_{\mathbf{z},\boldsymbol{\beta}} \bar{x}_{\mathbf{z},\boldsymbol{\beta}}^T \theta + N_{\mathbf{z},\boldsymbol{\beta}} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2
\end{aligned}
$$

We complete the square, and factor so that only one term depends on θ,

$$
= (N_{\mathbf{z},\boldsymbol{\beta}} + \alpha) \left\| \theta - \frac{N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \bar{x}_{\mathbf{z},\boldsymbol{\beta}} \right\|^2 + \frac{\alpha N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2.
$$

We observe the quadratic form of a Gaussian, with posterior component mean $\tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}} := \frac{N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}}+\alpha} \bar{x}_{\mathbf{z},\boldsymbol{\beta}}$ and variance $\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} := \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}}+\alpha}\sigma^2$.

$$
= \frac{\sigma^2}{\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}}} \|\theta - \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + \underbrace{\frac{\alpha N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \|\bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2 + \sum_{i:z_i=1} \beta_i \|x_i - \bar{x}_{\mathbf{z},\boldsymbol{\beta}}\|^2}_{A_2}
$$

For simplicity, we write A_2 for the terms that do not depend on θ. If we consider the original exponential term in Equation 4.11, we observe the desired Gaussian density,

$$
\begin{aligned}
\exp\left(-\frac{1}{2\sigma^2} A_1\right) &= \exp\left(-\frac{1}{2\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}}} \|\theta - \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}\|^2\right) \exp\left(-\frac{1}{2\sigma^2} A_2\right) \\
&= \left(2\pi\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}}\right)^{\frac{d}{2}} \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} I) \exp\left(-\frac{1}{2\sigma^2} A_2\right).
\end{aligned}
$$

We substitute this result into Equation 4.11, and simplify.

$$\begin{aligned}
p_{\boldsymbol{\beta}}(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta) &= p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left(\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \frac{1}{2\pi\sigma^2/\beta_i} \right)^{\frac{d}{2}} \left(\frac{1}{2\pi\sigma^2/\alpha} \right)^{\frac{d}{2}} \\
&\quad \times \left(2\pi\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}}\right)^{\frac{d}{2}} \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} I) \exp\left(-\frac{1}{2\sigma^2} A_2 \right) \\
&= p^{(0)}_{\boldsymbol{\beta}}(\mathbf{x} \mid \mathbf{z}) \left(\prod_{i \in S_{\mathbf{z},\boldsymbol{\beta}}} \frac{1}{2\pi\sigma^2/\beta_i} \right)^{\frac{d}{2}} \left(\frac{\alpha}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2} A_2 \right) \\
&\quad \times \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} I)
\end{aligned}$$

We recall $p_{\boldsymbol{\beta}}(\theta \mid \mathbf{x}) \propto \sum_{\mathbf{z}} p_{\boldsymbol{\beta}}(\mathbf{x} \mid \theta, \mathbf{z}) p(\theta)$, thus to compute the posterior we sum these component densities over all possible labelings $\mathbf{z}$, and arrive at the formula stated in Equation 4.3.

□

Proof of Lemma 4.2.2. This computation mirrors the derivation of the collapsed Gibbs transition probabilities under internal annealing (Section 3.5.1), where we simply substitute in our fractional annealing densities instead. We need not repeat the whole process here, and we simply begin at the ratio of marginal distributions, which now takes the form

$$\begin{aligned}
\frac{A^1_{\boldsymbol{\beta}}(\mathbf{z}^{[i \to 1]})}{A^1_{\boldsymbol{\beta}}(\mathbf{z}_{-i})} &= \frac{\int p(\theta) p_{\beta_i}(x_i \mid z_i = 1, \theta) q^1_{\theta,\boldsymbol{\beta}}(\mathbf{z}_{-i}) d\theta}{\int p(\theta) q^1_{\theta,\boldsymbol{\beta}}(\mathbf{z}_{-i}) d\theta}, \\
&= \int p_{\beta_i}(x_i \mid z_i = 1, \theta) \underbrace{\frac{p(\theta) q^1_{\theta,\boldsymbol{\beta}}(\mathbf{z}_{-i})}{\int p(\theta') q^1_{\theta',\boldsymbol{\beta}}(\mathbf{z}_{-i}) d\theta'}}_{\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z}_{-i},\boldsymbol{\beta}} I)} d\theta.
\end{aligned}$$

We have bracketed the term that is the posterior distribution of θ under the labeling $\mathbf{z}_{-i}$, with known parameters.

$$= \int p_{\beta_i}(x_i \mid z_i = 1, \theta) \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z}_{-i},\boldsymbol{\beta}} I) d\theta$$

This is just the formula for the convolution of the normal. Thus, as $p_{\beta_i}(x_i \mid z_i = 1, \theta) = \mathcal{N}(x_i; \theta, (\sigma^2/\beta_i)I)$, we have

$$\begin{aligned} &= \int \mathcal{N}(x_i; \theta, (\sigma^2/\beta_i)I)\mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z}_{-i},\boldsymbol{\beta}} I)d\theta \\ &= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, (\tilde{\sigma}^2_{\mathbf{z}_{-i}} + \sigma^2/\beta_i)I). \end{aligned} \tag{4.12}$$

We define $\tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}} := \frac{1}{\beta_i} + \frac{1}{N_{\mathbf{z}_{-i},\boldsymbol{\beta}}+\alpha}$ as the scaling for the posterior predictive variance, and thus

$$= \mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i},\boldsymbol{\beta}}, \tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}}\sigma^2 I), \tag{4.13}$$

completing the proof. □

4.5.2 Proofs for Section 4.3

Proof of Lemma 4.3.1. We begin with the unnormalized posterior weights, $\tilde{p}_L(\mathbf{z} \mid \mathbf{x})$ and $\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x})$, whose full formulae follow from the structure of the conjugate posterior (either under fractional annealing with $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}_{L-1}$, or under the original formulation with differing observed datasets $\mathbf{x}$ and $\mathbf{x}_{-1}$).

$$\tilde{p}_L(\mathbf{z} \mid \mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{Nd}{2}} \tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}}^0 d}{2}} \left(\frac{\alpha}{N_{\mathbf{z}} + \alpha}\right)^{\frac{d}{2}} \times \exp\left(-\frac{1}{2\sigma^2}\left[\frac{\alpha N_{\mathbf{z}}}{N_{\mathbf{z}} + \alpha}\|\bar{x}_{\mathbf{z}}\|^2 + \sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\tilde{V}_{\mathbf{w}}} \sum_{i:z_i=0} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2\right]\right) \tag{4.14}$$

$$\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{(N-1)d}{2}} \tilde{V}_{\mathbf{w}}^{\frac{N_{\mathbf{z}_{-1}}^0 d}{2}} \left(\frac{\alpha}{N_{\mathbf{z}_{-1}} + \alpha}\right)^{\frac{d}{2}} \times \exp\left(-\frac{1}{2\sigma^2}\left[\frac{\alpha N_{\mathbf{z}_{-1}}}{N_{\mathbf{z}_{-1}} + \alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2 + \sum_{\substack{i:z_i=1,\\ i>1}} \|x_i - \bar{x}_{\mathbf{z}_{-1}}\|^2 + \frac{1}{\tilde{V}_{\mathbf{w}}} \sum_{\substack{i:z_i=0,\\ i>1}} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2\right]\right) \tag{4.15}$$

The growth factor is the ratio between Equation 4.15 and Equation 4.14, and we consider the two separate cases (determined by the binary value of z_1).

Case #1: Assume $z_1 = 0$. The ratio simplifies, which leaves a single $\tilde{V}_{\mathbf{w}}$ term in the product (as $N_{\mathbf{z}}^0 = N_{\mathbf{z}_{-1}}^0 + 1$, but $N_{\mathbf{z}_{-1}} = N_{\mathbf{z}}$), and a single term in the exponent (as the sample means $\bar{x}_{\mathbf{z}_{-1}} = \bar{x}_{\mathbf{z}}$ are equal).

$$\begin{aligned} r_{\mathbf{z}}(L-1) &:= \frac{\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x})}{\tilde{p}_L(\mathbf{z} \mid \mathbf{x})} \\ &= \left(2\pi\sigma^2\right)^{\frac{d}{2}} \tilde{V}_{\mathbf{w}}^{-\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2 \tilde{V}_{\mathbf{w}}}\|x_1 - \tilde{\mu}_{\mathbf{w}}\|^2\right) \end{aligned}$$

Case #2: Assume $z_1 = 1$. As the sample means are no longer equal, what remains in the

exponential is a difference between sums of squares.

$$
\begin{aligned}
r_{\mathbf{z}}(L-1) &:= \frac{\tilde{p}_{L-1}(\mathbf{z} \mid \mathbf{x})}{\tilde{p}_{L}(\mathbf{z} \mid \mathbf{x})} \\
&= \left(2\pi\sigma^2\right)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \\
&\quad \times \exp\Bigg(-\tfrac{1}{2\sigma^2}\Bigg[\tfrac{\alpha(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2 - \tfrac{\alpha N_{\mathbf{z}}}{N_{\mathbf{z}}+\alpha}\|\bar{x}_{\mathbf{z}}\|^2 \\
&\qquad\qquad + \sum_{\substack{i:z_i=1,\\ i>1}} \left[\|x_i-\bar{x}_{\mathbf{z}_{-1}}\|^2 - \|x_i-\bar{x}_{\mathbf{z}}\|^2\right] - \|x_1-\bar{x}_{\mathbf{z}}\|^2\Bigg]\Bigg) \\
&= \left(2\pi\sigma^2\right)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2}[\mathrm{SSD}^1_{\mathbf{z}}]\right)
\end{aligned}
$$

We use $[\mathrm{SSD}^1_{\mathbf{z}}]$ as convenient shorthand to refer to this term (capturing the difference in the sum of squares, for the labeling $\mathbf{z}$, with the superscript denoting the index that is excluded),

$$
\begin{aligned}
[\mathrm{SSD}^1_{\mathbf{z}}] := \|x_1-\bar{x}_{\mathbf{z}}\|^2 &- \left[\tfrac{\alpha(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2 - \tfrac{\alpha N_{\mathbf{z}}}{N_{\mathbf{z}}+\alpha}\|\bar{x}_{\mathbf{z}}\|^2\right] \\
&- \sum_{\substack{i:z_i=1,\\ i>1}} \left[\|x_i-\bar{x}_{\mathbf{z}_{-1}}\|^2 - \|x_i-\bar{x}_{\mathbf{z}}\|^2\right],
\end{aligned}
$$

and this completes the proof. □

Proof of Lemma 4.3.3. By Lemma 4.3.1, and the assumption that $z_1 = 1$, we have

$$
r_{\mathbf{z}}(L-1) = \left(2\pi\sigma^2\right)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2}[\mathrm{SSD}^1_{\mathbf{z}}]\right). \tag{4.16}
$$

The sample size requirement of Equation 4.5 ensures $N_{\mathbf{z}} \geq d+1$, and as the ratio is decreasing in $N_{\mathbf{z}}+\alpha$, we have

$$
\left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \leq \left(\frac{d+1}{d}\right)^{\frac{d}{2}},
$$

which we bound using Lemma 2.3.1,

$$\left(\frac{N_{\mathbf{z}}+\alpha}{N_{\mathbf{z}}-1+\alpha}\right)^{\frac{d}{2}} \leq 2. \tag{4.17}$$

Next, we decompose the resulting $[\mathrm{SSD}^1_{\mathbf{z}}]$ term in the exponent.

$$\begin{aligned}[\mathrm{SSD}^1_{\mathbf{z}}] = \|x_1 - \bar{x}_{\mathbf{z}}\|^2 + \underbrace{\left[\tfrac{\alpha N_{\mathbf{z}}}{N_{\mathbf{z}}+\alpha}\|\bar{x}_{\mathbf{z}}\|^2 - \tfrac{\alpha(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2\right]}_{A_1} \\ + \underbrace{\sum_{\substack{i:z_i=1,\\ i>1}} \left[\|x_i - \bar{x}_{\mathbf{z}}\|^2 - \|x_i - \bar{x}_{\mathbf{z}_{-1}}\|^2\right]}_{A_2}.\end{aligned} \tag{4.18}$$

We can control the size of A_1, as our assumptions ensure that $\|\bar{x}_{\mathbf{z}_{-1}}\|$ and $\|\bar{x}_{\mathbf{z}}\|$ are not too far apart.

$$\begin{aligned}A_1 &= \alpha\left[\tfrac{N_{\mathbf{z}}}{N_{\mathbf{z}}+\alpha}\|\bar{x}_{\mathbf{z}}\|^2 - \tfrac{(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\|\bar{x}_{\mathbf{z}_{-1}}\|^2\right] \\ &= \alpha\left[\tfrac{\alpha}{(N_{\mathbf{z}}-1+\alpha)(N_{\mathbf{z}}+\alpha)}\|\bar{x}_{\mathbf{z}}\|^2 + \tfrac{(N_{\mathbf{z}}-1)}{N_{\mathbf{z}}-1+\alpha}\left[\|\bar{x}_{\mathbf{z}}\|^2 - \|\bar{x}_{\mathbf{z}_{-1}}\|^2\right]\right]\end{aligned}$$

The sample size requirement of Equation 4.6 implies that $\|\bar{x}_{\mathbf{z}}\|/(N_{\mathbf{z}}-1) \leq \delta$, thus

$$\leq \alpha^2\delta^2 + \alpha\left|\|\bar{x}_{\mathbf{z}}\|^2 - \|\bar{x}_{\mathbf{z}_{-1}}\|^2\right|. \tag{4.19}$$

Intuitively, as the cluster for label $\mathbf{z}$ is tightly packed, the squared norms of the sample means must be similar. To formalize this claim, we expand the terms

$$\begin{aligned}\left|\|\bar{x}_{\mathbf{z}}\|^2 - \|\bar{x}_{\mathbf{z}_{-1}}\|^2\right| &= \left|\left[\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{z}_{-1}}\right]^T\left[\bar{x}_{\mathbf{z}} + \bar{x}_{\mathbf{z}_{-1}}\right]\right| \\ &\leq \left\|\bar{x}_{\mathbf{z}} - \bar{x}_{\mathbf{z}_{-1}}\right\| \ \ \left\|\bar{x}_{\mathbf{z}} + \bar{x}_{\mathbf{z}_{-1}}\right\|.\end{aligned}$$

We recall the identity that translates between the two sample means, $\bar{x}_{\mathbf{z}_{-1}} = \frac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}} - x_1}{N_{\mathbf{z}}-1}$ (Equation 2.24).

$$
\begin{aligned}
&= \left\| \bar{x}_{\mathbf{z}} - \frac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}} - x_1}{N_{\mathbf{z}} - 1} \right\| \quad \left\| \bar{x}_{\mathbf{z}} + \frac{N_{\mathbf{z}}\bar{x}_{\mathbf{z}} - x_1}{N_{\mathbf{z}} - 1} \right\| \\
&= \left\| \frac{x_1 - \bar{x}_{\mathbf{z}}}{N_{\mathbf{z}} - 1} \right\| \quad \left\| 2\bar{x}_{\mathbf{z}} + \frac{\bar{x}_{\mathbf{z}} - x_1}{N_{\mathbf{z}} - 1} \right\| \\
&\leq \left[\tfrac{1}{N_{\mathbf{z}}-1} \|x_1 - \bar{x}_{\mathbf{z}}\| \right] \left[2\|\bar{x}_{\mathbf{z}}\| + \tfrac{1}{N_{\mathbf{z}}-1} \|\bar{x}_{\mathbf{z}} - x_1\| \right]
\end{aligned}
$$

By construction, $\|\bar{x}_{\mathbf{z}} - x_1\| \leq \delta$.

$$
\leq \left[\tfrac{\delta}{N_{\mathbf{z}}-1} \right] \left[2\|\bar{x}_{\mathbf{z}}\| + \tfrac{\delta}{N_{\mathbf{z}}-1} \right].
$$

The sample size requirement of Equation 4.6 implies that $\|\bar{x}_{\mathbf{z}}\|/(N_{\mathbf{z}} - 1) \leq \delta$.

$$
\leq \delta^2 \left[2 + \tfrac{1}{(N_{\mathbf{z}}-1)^2} \right]
$$

We substitute this back into Equation 4.19,

$$
A_1 \leq \alpha^2\delta^2 + \alpha \left[\delta^2 \left[2 + \tfrac{1}{(N_{\mathbf{z}}-1)^2} \right] \right]
$$

and as $\alpha \leq 1$, we observe

$$
\leq \delta^2 \left[3 + \tfrac{1}{(N_{\mathbf{z}}-1)^2} \right]. \tag{4.20}
$$

Before tackling the A_2 term, we note the following sum of squares identity for the sample mean,

$$
\sum_{\substack{i:z_i=1, \\ i>1}} \|x_i - \bar{x}_{\mathbf{z}}\|^2 = \sum_{\substack{i:z_i=1, \\ i>1}} \|x_i - \bar{x}_{\mathbf{z}_{-1}}\|^2 + (N_{\mathbf{z}} - 1)\|\bar{x}_{\mathbf{z}_{-1}} - \bar{x}_{\mathbf{z}}\|^2. \tag{4.21}
$$

Substituting Equation 4.21 into our expression for A_2, we observe

$$\begin{aligned} A_2 &:= \sum_{\substack{i:z_i=1,\\ i>1}} \left[\|x_i - \bar{x}_{\mathbf{z}}\|^2 - \|x_i - \bar{x}_{\mathbf{z}_{-1}}\|^2\right] \\ &= (N_{\mathbf{z}} - 1)\|\bar{x}_{\mathbf{z}_{-1}} - \bar{x}_{\mathbf{z}}\|^2. \end{aligned}$$

We mirror the computation above, and reach the bound

$$\begin{aligned} &= (N_{\mathbf{z}} - 1)\left\|\frac{x_1 - \bar{x}_{\mathbf{z}}}{N_{\mathbf{z}} - 1}\right\|^2 \\ &\leq \frac{\delta^2}{N_{\mathbf{z}} - 1}. \end{aligned} \tag{4.22}$$

We substitute Equations 4.20 & 4.22 into Equation 4.18, and recall that by construction, $\|x_1 - \bar{x}_{\mathbf{z}}\| \leq \delta$.

$$\begin{aligned} [\mathrm{SSD}_{\mathbf{z}}^1] &= \|x_1 - \bar{x}_{\mathbf{z}}\|^2 + A_1 + A_2 \\ &\leq \delta^2 + \delta^2\left[3 + \tfrac{1}{(N_{\mathbf{z}}-1)^2}\right] + \frac{\delta^2}{N_{\mathbf{z}} - 1} \end{aligned} \tag{4.23}$$

By our stated sample size requirement (Equation 4.5), this is bounded by

$$\leq 5\delta^2. \tag{4.24}$$

To complete the proof, we substitute Equations 4.17 & 4.24 into our original expression for the growth factor (Equation 4.16), and reach the desired bound,

$$\begin{aligned} r_{\mathbf{z}}(L-1) &= \left(2\pi\sigma^2\right)^{\frac{d}{2}}\left(\frac{N_{\mathbf{z}} + \alpha}{N_{\mathbf{z}} - 1 + \alpha}\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2}[\mathrm{SSD}_{\mathbf{z}}^1]\right) \\ &\leq \left(2\pi\sigma^2\right)^{\frac{d}{2}} 2\exp\left(\frac{5}{2\sigma^2}\delta^2\right) \\ &= \left(2\pi\sigma^2\right)^{\frac{d}{2}} 2\exp\left(\left[\frac{5r_\delta^2}{2}\right]\frac{u^2}{\sigma^2}\right). \end{aligned}$$

□

Proof of Lemma 4.3.4. We need to consider two cases, for the two potential values of the x_1 label under the given $\mathbf{z}'$.

Case #1: Let $z'_1 = 1$. By Lemma 4.3.1, we have

$$r_{\mathbf{z}'}(L) = \left(2\pi\sigma^2\right)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}'} + \alpha}{N_{\mathbf{z}'} - 1 + \alpha}\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2}[\mathrm{SSD}^1_{\mathbf{z}'}]\right). \tag{4.25}$$

We follow a similar structure to the proof of Lemma 4.3.3, but we flip the signs of A_1 and A_2, so that we once again are constructing an upper bound.

$$\begin{aligned}[\mathrm{SSD}^1_{\mathbf{z}'}] = \|x_1 - \bar{x}_{\mathbf{z}'}\|^2 - \underbrace{\left[\tfrac{\alpha(N_{\mathbf{z}'}-1)}{N_{\mathbf{z}'}-1+\alpha}\|\bar{x}_{\mathbf{z}'_{-1}}\|^2 - \tfrac{\alpha N_{\mathbf{z}'}}{N_{\mathbf{z}'}+\alpha}\|\bar{x}_{\mathbf{z}'}\|^2\right]}_{A_1} \\ - \underbrace{\sum_{\substack{i:z'_i=1,\\ i>1}} \left[\|x_i - \bar{x}_{\mathbf{z}'_{-1}}\|^2 - \|x_i - \bar{x}_{\mathbf{z}'}\|^2\right]}_{A_2}.\end{aligned} \tag{4.26}$$

Our computation mirrors the proof of Lemma 4.3.3.

$$\begin{aligned}A_1 &= \alpha\left[\tfrac{(N_{\mathbf{z}'}-1)}{N_{\mathbf{z}'}-1+\alpha}\|\bar{x}_{\mathbf{z}'_{-1}}\|^2 - \tfrac{N_{\mathbf{z}'}}{N_{\mathbf{z}'}+\alpha}\|\bar{x}_{\mathbf{z}'}\|^2\right] \\ &\leq \tfrac{\alpha N_{\mathbf{z}'}}{N_{\mathbf{z}'}+\alpha}\left|\|\bar{x}_{\mathbf{z}'_{-1}}\|^2 - \|\bar{x}_{\mathbf{z}'}\|^2\right|\end{aligned}$$

We begin with the familiar decomposition for this difference of squares.

$$\left|\|\bar{x}_{\mathbf{z}'_{-1}}\|^2 - \|\bar{x}_{\mathbf{z}'}\|^2\right| \leq \left\|\bar{x}_{\mathbf{z}'_{-1}} - \bar{x}_{\mathbf{z}'}\right\| \quad \left\|\bar{x}_{\mathbf{z}'_{-1}} + \bar{x}_{\mathbf{z}'}\right\|$$

We recall that $\bar{x}_{\mathbf{z}'_{-1}} = \frac{N_{\mathbf{z}'}\bar{x}_{\mathbf{z}'} - x_1}{N_{\mathbf{z}'}-1}$.

$$
\begin{aligned}
&= \left\| \frac{N_{\mathbf{z}'}\bar{x}_{\mathbf{z}'} - x_1}{N_{\mathbf{z}'}-1} - \bar{x}_{\mathbf{z}'} \right\| \left\| \frac{N_{\mathbf{z}'}\bar{x}_{\mathbf{z}'} - x_1}{N_{\mathbf{z}'}-1} + \bar{x}_{\mathbf{z}'} \right\| \\
&= \left\| \frac{\bar{x}_{\mathbf{z}'} - x_1}{N_{\mathbf{z}'}-1} \right\| \left\| 2\bar{x}_{\mathbf{z}'} + \frac{\bar{x}_{\mathbf{z}'} - x_1}{N_{\mathbf{z}'}-1} \right\| \\
&\leq \left[\tfrac{1}{N_{\mathbf{z}'}-1} \|\bar{x}_{\mathbf{z}'} - x_1\| \right] \left[2\|\bar{x}_{\mathbf{z}'}\| + \tfrac{1}{N_{\mathbf{z}'}-1} \|\bar{x}_{\mathbf{z}'} - x_1\| \right]
\end{aligned}
$$

The sample size requirement (Equation 4.8) implies that $\frac{1}{N_{\mathbf{z}'}-1}\|\bar{x}_{\mathbf{z}'}\| \leq \delta$.

$$
\leq \|\bar{x}_{\mathbf{z}'} - x_1\| \left[2\delta + \tfrac{1}{(N_{\mathbf{z}'}-1)^2} \|\bar{x}_{\mathbf{z}'} - x_1\| \right]
$$

We recall that $\delta = r_\delta u$.

$$
= 2r_\delta u \|\bar{x}_{\mathbf{z}'} - x_1\| + \tfrac{1}{(N_{\mathbf{z}'}-1)^2} \|\bar{x}_{\mathbf{z}'} - x_1\|^2
$$

By construction, we recall that $u := \|\bar{x}_{\mathbf{w}} - \bar{x}_{\mathbf{z}}\| \leq \|\bar{x}_{\mathbf{z}'} - \bar{x}_{\mathbf{z}}\| - \delta$ (Equation 4.7). As $\|\bar{x}_{\mathbf{z}} - x_1\| \leq \delta$, this implies $\|\bar{x}_{\mathbf{z}'} - \bar{x}_{\mathbf{z}}\| - \delta \leq \|\bar{x}_{\mathbf{z}'} - \bar{x}_1\|$, and we have

$$
\begin{aligned}
&\leq 2r_\delta \|\bar{x}_{\mathbf{z}'} - x_1\|^2 + \tfrac{1}{(N_{\mathbf{z}'}-1)^2} \|\bar{x}_{\mathbf{z}'} - x_1\|^2 \\
&= \left[2r_\delta + \tfrac{1}{(N_{\mathbf{z}'}-1)^2} \right] \|\bar{x}_{\mathbf{z}'} - x_1\|^2.
\end{aligned}
$$

The sample size requirement of Equation 4.8 implies $1/(N_{\mathbf{z}'}-1)^2 \leq 1/16$,

$$
\leq [2r_\delta + 1/16] \|\bar{x}_{\mathbf{z}'} - x_1\|^2.
$$

For the A_2 term, we note that the sample mean minimizes the sum of squared distances, so

this term is nonpositive:

$$A_2 = \sum_{\substack{i:z'_i=1,\\ i>1}} \left[\|x_i - \bar{x}_{\mathbf{z}'_{-1}}\|^2 - \|x_i - \bar{x}_{\mathbf{z}'}\|^2 \right],$$
$$\leq 0.$$

We substitute A_1 and A_2 into Equation 4.26,

$$\begin{aligned}
[\mathrm{SSD}^1_{\mathbf{z}'}] &= \|x_1 - \bar{x}_{\mathbf{z}'}\|^2 - A_1 - A_2 \\
&\geq \|x_1 - \bar{x}_{\mathbf{z}'}\|^2 - [2r_\delta + 1/16] \|x_1 - \bar{x}_{\mathbf{z}'}\|^2 \\
&= [1 - 2r_\delta - 1/16] \|x_1 - \bar{x}_{\mathbf{z}'}\|^2,
\end{aligned}$$

and substitute $[\mathrm{SSD}^1_{\mathbf{z}'}]$ into Equation 4.25,

$$\begin{aligned}
r_{\mathbf{z}'}(L-1) &= \left(2\pi\sigma^2\right)^{\frac{d}{2}} \left(\frac{N_{\mathbf{z}'} + \alpha}{N_{\mathbf{z}'} - 1 + \alpha} \right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2} [\mathrm{SSD}^1_{\mathbf{z}'}] \right) \\
&\geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2} [1 - 2r_\delta - 1/16] \|x_1 - \bar{x}_{\mathbf{z}'}\|^2 \right).
\end{aligned}$$

By the construction of $\mathbf{Z}'$ (Equation 4.7), we have $\|x_1 - \bar{x}_{\mathbf{z}'}\| \geq u$,

$$\geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \exp\left([15/32 - r_\delta] \frac{u^2}{\sigma^2} \right). \tag{4.27}$$

Case #2: Let $z'_1 = 0$. By Lemma 4.3.1, we have:

$$r_{\mathbf{z}'}(L-1) = \left(2\pi\sigma^2\right)^{\frac{d}{2}} \tilde{V}_{\mathbf{w}}^{-\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2 \tilde{V}_{\mathbf{w}}} \|x_1 - \tilde{\mu}_{\mathbf{w}}\|^2 \right) \tag{4.28}$$

The sample size requirement of Equation 4.5 implies $N_{\mathbf{w}} + \alpha \geq d$, and as the ratio is

increasing in $N_{\mathbf{w}} + \alpha$, we again have

$$\begin{aligned}\tilde{V}_{\mathbf{w}}^{-\frac{d}{2}} &= \left(\frac{N_{\mathbf{w}} + \alpha}{N_{\mathbf{w}} + \alpha + 1}\right)^{\frac{d}{2}} \\ &\geq \left(\frac{d}{d+1}\right)^{\frac{d}{2}}.\end{aligned}$$

and by Lemma 2.3.1,

$$\geq \frac{1}{2}.. \tag{4.29}$$

Next, we lower bound the distance $\|x_1 - \tilde{\mu}_{\mathbf{w}}\|^2$.

$$\begin{aligned}\|\tilde{\mu}_{\mathbf{w}} - x_1\| &= \left\| \tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} \bar{x}_{\mathbf{w}} - x_1 \right\| \\ &= \left\| \tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} \left[\bar{x}_{\mathbf{w}} - x_1\right] - \tfrac{\alpha}{\alpha + N_{\mathbf{w}}} x_1 \right\| \\ &\geq \tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} \|\bar{x}_{\mathbf{w}} - x_1\| - \tfrac{\alpha}{\alpha + N_{\mathbf{w}}} \|x_1\|\end{aligned}$$

The sample size requirement of Equation 4.6 implies $\frac{1}{\alpha + N_{\mathbf{w}}} \|x_1\| \leq \delta$, and by construction, $\|\bar{x}_{\mathbf{w}} - x_1\| \geq u - \delta$,

$$\geq \tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} \left[u - \delta\right] - \alpha\delta.$$

As $r_\delta := \delta / u$, we have

$$= u \left[\tfrac{N_{\mathbf{w}}}{\alpha + N_{\mathbf{w}}} (1 - r_\delta) - \alpha r_\delta \right]. \tag{4.30}$$

Returning to Equation 4.28, and substituting in Equations 4.29 & 4.30, we have

$$r_{\mathbf{z}'}(L-1) = \left(2\pi\sigma^2\right)^{\frac{d}{2}} \tilde{V}_{\mathbf{w}}^{-\frac{d}{2}} \exp\left(\frac{1}{2\sigma^2\tilde{V}_{\mathbf{w}}}\|x_1 - \tilde{\mu}_{\mathbf{w}}\|^2\right)$$
$$\geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \frac{1}{2} \exp\left(\frac{\left[\frac{N_{\mathbf{w}}}{\alpha+N_{\mathbf{w}}}(1-r_\delta) - \alpha r_\delta\right]^2}{2\frac{N_{\mathbf{w}}+\alpha+1}{N_{\mathbf{w}}+\alpha}} \frac{u^2}{\sigma^2}\right)$$

We note that $r_\delta < 9/40$, $\alpha \in (0,1]$, $\frac{N_{\mathbf{w}}}{\alpha+N_{\mathbf{w}}} \geq 9/10$, and $\frac{N_{\mathbf{w}}+\alpha+1}{N_{\mathbf{w}}+\alpha} \leq 10/9$ (as implied by Equation 4.5). We expand out the square, plug in these terms, and simplify the resulting quadratic bound (in terms of r_δ) so that it has a clean denominator.

$$\geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \frac{1}{2} \exp\left(\left[1/3 - 3r_\delta/2 + 3r_\delta^2\right] \frac{u^2}{\sigma^2}\right) \tag{4.31}$$

Thus, between the two cases (Equations 4.27 & 4.31), for any $\mathbf{z}' \in \mathbf{Z}'$, the growth factor is at least

$$r_{\mathbf{z}'}(L-1) \geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \min\left\{\exp\left(\left[15/32 - r_\delta\right] \frac{u^2}{\sigma^2}\right), \right.$$
$$\left. \frac{1}{2}\exp\left(\left[1/3 - 3r_\delta/2 + 3r_\delta^2\right] \frac{u^2}{\sigma^2}\right)\right\}.$$

Comparing the two bracketed scaling factors, we observe that $1/3 - 3r_\delta/2 + 3r_\delta^2 > 15/32 - r_\delta$ for $r_\delta \in [0, 9/40)$, which leads to the desired bound,

$$\geq \left(2\pi\sigma^2\right)^{\frac{d}{2}} \frac{1}{2} \exp\left(\left[15/32 - r_\delta\right] \frac{u^2}{\sigma^2}\right).$$

□

We use these lemmas to create the final missing piece in the theorem proof—a bound on the *ratio* of growth factors.

Lemma 4.5.1. *For the setup given by Theorem 4.3.2, the ratio of growth factors between*

the label $\mathbf{z}$*, and any label* $\mathbf{z}' \in \mathbf{Z}'$*, is bounded by*

$$\frac{r_{\mathbf{z}}(L-1)}{r_{\mathbf{z}'}(L-1)} \leq 4 \exp\left(- \left[15/32 - r_\delta - 5r_\delta^2/2 \right] \frac{u^2}{\sigma^2} \right).$$

Proof of Lemma 4.5.1. We cite Lemma 4.3.3 for an upper bound on $r_{\mathbf{z}}(L-1)$ and Lemma 4.3.4 for a lower bound on $r_{\mathbf{z}'}(L-1)$, and observe

$$\begin{aligned} \frac{r_{\mathbf{z}}(L-1)}{r_{\mathbf{z}'}(L-1)} &\leq \frac{(2\pi\sigma^2)^{\frac{d}{2}} \, 2 \exp\left(\left[\frac{5r_\delta^2}{2} \right] \frac{u^2}{\sigma^2} \right)}{(2\pi\sigma^2)^{\frac{d}{2}} \, \frac{1}{2} \exp\left(\left[15/32 - r_\delta \right] \frac{u^2}{\sigma^2} \right)} \\ &= 4 \exp\left(- \left[15/32 - r_\delta - 5r_\delta^2/2 \right] \frac{u^2}{\sigma^2} \right). \end{aligned}$$

We note that for $r_\delta \in [0, 9/40]$, the bracketed term is positive, as desired.

□

Proof of Theorem 4.3.2. For the given labeling $\mathbf{z}$, we define our conductance subset as $\mathbf{Q} := \{(L, \mathbf{z}), (L-1, \mathbf{z})\}$, the chosen label $\mathbf{z}$ at the two annealing indices, L and $L-1$ (which differ just by the removal of the datum x_1). We recall the premise of the conductance argument (Equation 4.4),

$$\begin{aligned} \Phi(\mathbf{Q}) &= \frac{\sum_{\mathbf{z}' \neq \mathbf{z}} \pi(L, \mathbf{z}) T_L(\mathbf{z}' \mid \mathbf{z})}{\pi(L, \mathbf{z}) + \pi(L-1, \mathbf{z})} + \frac{\sum_{\mathbf{z}' \neq \mathbf{z}} \pi(L-1, \mathbf{z}) T_{L-1}(\mathbf{z}' \mid \mathbf{z})}{\pi(L, \mathbf{z}) + \pi(L-1, \mathbf{z})} \\ &\leq T^*_{L,\mathbf{z}} + \frac{p_{L-1}(\mathbf{z} \mid \mathbf{x})}{p_L(\mathbf{z} \mid \mathbf{x})}. \end{aligned} \tag{4.32}$$

We wish to bound the ratio of normalized densities using the ratio of growth factors. We cite Lemma 3.4.1, with $\mathbf{Z}'$ as the subset of labels whose total probability mass is at least $c^* = 1/10$. Let $r^*(L-1) := \min_{\mathbf{z}' \in \mathbf{Z}'} \{ r_{\mathbf{z}'}(L-1) \}$ denote a bound on the growth factor for

any $\mathbf{z}' \in \mathbf{Z}'$.

$$\begin{aligned}\frac{p_{L-1}(\mathbf{z} \mid \mathbf{x})}{p_L(\mathbf{z} \mid \mathbf{x})} &\leq \frac{1}{10} \frac{r_{\mathbf{z}}(L-1)}{r^*(L-1)} \\ &= \frac{1}{10} \max_{\mathbf{z}' \in \mathbf{Z}'} \left\{ \frac{r_{\mathbf{z}}(L-1)}{r_{\mathbf{z}'}(L-1)} \right\}\end{aligned}$$

We cite Lemma 4.5.1 to upper bound the ratio of growth factors for any $\mathbf{z}' \in \mathbf{Z}'$,

$$\leq \frac{2}{5} \exp\left(- \left[15/32 - r_\delta - 5r_\delta^2/2 \right] \frac{u^2}{\sigma^2} \right).$$

We recall that $T^*_{L,\mathbf{z}}$ denotes maximal probability of transition away from the labeling $\mathbf{z}$ under the original posterior. This theorem shares the conditions of Theorem 2.2.1, and thus we cite that original proof (Equation 2.28) to bound this probability.

$$T^*_{L,\mathbf{z}} \leq 2 \max \left\{ \exp\left(- \left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(- \left[\frac{9 - 40r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\}$$

Thus, the upper bound on the conductance in Equation 4.32 is a sum of two terms that are exponentially small (in u and Δ). In order to take its inverse for the mixing time bound, it is convenient to combine these terms.

$$\begin{aligned}\Phi(\mathbf{Q}) \leq\ & 2 \max \left\{ \exp\left(- \left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(- \left[\frac{9 - 40r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\} \\ & + \frac{2}{5} \exp\left(- \left[15/32 - r_\delta - 5r_\delta^2/2 \right] \frac{u^2}{\sigma^2} \right)\end{aligned}$$

We note that $9/20 - 2r_\delta < 15/32 - r_\delta - 5r_\delta^2/2$ for $r_\delta \in [0, 9/40]$, and thus we can simply use the smaller bracketed scaling factor for the purposes of the maximum.

$$\begin{aligned}
&\leq 2 \max\left\{ \exp\left(-\left[\frac{7-14R}{20}\right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(-\left[\frac{9-40r_\delta}{20}\right] \frac{u^2}{\sigma^2} \right) \right\} \\
&\quad + \frac{2}{5} \exp\left(-\left[\frac{9-40r_\delta}{20}\right] \frac{u^2}{\sigma^2} \right) \\
&\leq \frac{12}{5} \max\left\{ \exp\left(-\left[\frac{7-14R}{20}\right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(-\left[\frac{9-40r_\delta}{20}\right] \frac{u^2}{\sigma^2} \right) \right\} \qquad (4.33)
\end{aligned}$$

Finally, we again translate an upper bound on the conductance (Equation 4.33) into a lower bound on the mixing time by the bound of Jerrum & Sinclair (Equation 2.2),

$$\begin{aligned}
\tau_{\text{mix}} &\geq \frac{1}{4\Phi(\mathbf{Q})} \\
&\geq \frac{5}{48} \min\left\{ \exp\left(\left[\frac{7-14R}{20}\right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(\left[\frac{9-40r_\delta}{20}\right] \frac{u^2}{\sigma^2} \right) \right\},
\end{aligned}$$

which completes the proof. □

Appendix A

Supplemental Notation Reference

This appendix provides a supplemental reference for the notation used in the dissertation (every term is first defined in the main text). It is not comprehensive, and it is primarily intended for terms that *we* define (not just general notational choices), and are used multiple times throughout the work. This appendix is not necessary for any aspect of the document—it is simply intended as a convenience, in case it is ever difficult to find the introduction of a term.

A.1 General Model Notation

- Observed Data: $\mathbf{x} = (x_1, \ldots, x_N)$.
- Latent Label: $\mathbf{z} = (z_1, \ldots, z_N) \in \{0, 1\}^N$.
- Variable Density: $p(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, \sigma^2 I)$.
- Fixed Density: $p(x_i \mid z_i = 0)$, with its definition left flexible.
- Likelihood: $p(\mathbf{x} \mid \theta) = \prod_{i=1}^{N} \frac{1}{2}[p(x_i \mid z_i = 0) + p(x_i \mid \theta, z_i = 1)]$.
- Conditional Likelihood: $p(\mathbf{x} \mid \theta, \mathbf{z}) := \prod_{i=1}^{N} p(x_i \mid z_i, \theta)$.

- Prior: $p(\theta) := \mathcal{N}(\theta; 0, (\sigma^2/\alpha)I)$, for $\alpha \in (0, 1]$.
- Posterior: $p(\theta|\mathbf{x}) \propto p(\theta)p(\mathbf{x} \mid \theta) \propto \sum_{\mathbf{z}} p(\mathbf{x} \mid \theta, \mathbf{z})p(\theta)$.

A.2 Notation for Chapter 1

- The full conjugate posterior formula uses parameters defined for a labeling $\mathbf{z}$—the sample size, $N_{\mathbf{z}} = \sum_{i=1}^{N} z_i$, and the sample mean, $\bar{x}_{\mathbf{z}} = \frac{1}{N_{\mathbf{z}}} \sum_{i:z_i=1} x_i$, of the data assigned to the variable component under $\mathbf{z}$.
- The conjugate posterior component density (for a given $\mathbf{z}$) is Gaussian,

$$p(\theta \mid \mathbf{z}, \mathbf{x}) = \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z}}, \tilde{\sigma}^2_{\mathbf{z}} I),$$

  with parameters

$$\tilde{\mu}_{\mathbf{z}} := \frac{N_{\mathbf{z}}}{\alpha + N_{\mathbf{z}}} \bar{x}_{\mathbf{z}},$$
$$\tilde{\sigma}^2_{\mathbf{z}} := \frac{1}{\alpha + N_{\mathbf{z}}} \sigma^2.$$

- Let $\mathbf{z} \in \mathcal{Z}$ denote the state space of the posterior labels. Under our greedy construction, it is the set of all length N binary vectors, $\mathcal{Z} := \{0, 1\}^N$.
- Let $\mathbf{z}^{[i \to 1]}$ and $\mathbf{z}^{[i \to 0]}$ refer to the labeling $\mathbf{z}$ with the ith index overwritten to be equal to 1 or 0 (respectively).
- We define $\tilde{V}_{\mathbf{z}} := 1 + \frac{1}{N_{\mathbf{z}} + \alpha}$ as the scaling factor for the posterior predictive variance.
- Let $\theta \in \Omega$ denote the state space of the parameters. Under our greedy construction, it is simply $\Omega := \mathbb{R}^d$.
- Let $\mathbf{w}$ refer to the label identifying the subset of data that defines the fixed component

in our mixing analysis (each of the subsequent chapters share this notation for their mixing bound).

A.3 Notation for Chapter 2

- In Section 2.2, let $T(\cdot \mid \cdot)$ denote the collapsed Gibbs transition kernel. This combines the probability of selecting an index i with the collapsed Gibbs conditional transition probability described in Lemma 1.5.1. Written explicitly, if $\mathbf{z}'$ denotes a destination label differing from the current label $\mathbf{z}$ on solely the ith index (i.e. $z_i' = 1 - z_i$, and $z_j = z_j'$ for $j \neq i$), then

$$T(\mathbf{z}' \mid \mathbf{z}) = \frac{1}{N} p(z_i' \mid \mathbf{z}_{-i}, \mathbf{x}).$$

- $T^*_{\mathbf{z}} := \max_{\mathbf{z}' \neq \mathbf{z}} \{NT(\mathbf{z}' \mid \mathbf{z})\} = \max_i \{p(1-z_i \mid \mathbf{z}_{-i}, \mathbf{x})\}$ is an upper bound on the maximal probability of "escape" from the label $\mathbf{z}$, under the collapsed Gibbs sampler.

- The terminology used in the setting of Theorem 2.2.1 (as illustrated by Figure 2.1):
  - $\delta := \max_{i:z_i=1} \{\|\bar{x}_{\mathbf{z}} - x_i\|\}$ is the maximal radius of the target cluster $\mathbf{z}$.
  - $u := \|\bar{x}_{\mathbf{w}} - \bar{x}_{\mathbf{z}}\|$ is the separation between the sample mean of the target cluster, and the subset used to build the fixed density.
  - $\Delta := \min_{i:z_i=0} \{\|\bar{x}_{\mathbf{z}} - x_i\|\}$ is the minimum distance from the target cluster center to any new datum not currently included.
  - $r_\delta := \frac{\delta}{u}$ and $R := \max_{i:z_i=0} \left\{ \frac{\|\bar{x}_{\mathbf{w}} - x_i\|}{\|\bar{x}_{\mathbf{z}} - x_i\|} \right\}$ are the ratios that are used to ensure minimum sufficient cluster separation.

A.4 Notation for Chapter 3

- Under temperature annealing, an inverse temperature schedule is given by $0 \leq \beta_1 < \beta_2 < \ldots < \beta_L = 1$, where β_1 provides the high temperature base distribution, and $\beta_L = 1$ provides the (cold) original density of interest. Throughout this work, we note that increasing ℓ implies increasing *inverse* temperature, and thus decreasing temperature.

- We denote the joint stationary distribution of a simulated tempering chain with $\pi(\ell, \cdot)$, for $\ell \in [L]$ as the annealing index. If we denote our state space as $y \in \mathcal{Y}$, then its conditional distribution matches our interpolating distributions, $\pi(y \mid \ell) = p_\ell(y)$.

- Aspects of the internal annealing model.

  - Variable Density: $p_\beta(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, (\sigma^2/\beta)I)$.
  - Likelihood: $p_\beta(\mathbf{x}|\theta) := \prod_{i=1}^N \frac{1}{2}\left[p_\beta(x_i|z_i = 0) + p_\beta(x_i \mid \theta, z_i = 1)\right]$.
  - Posterior: $p_\beta(\theta|\mathbf{x}) \propto \sum_{\mathbf{z}} \tilde{p}_\beta(\mathbf{z} \mid \mathbf{x}) p_\beta(\theta \mid \mathbf{z}, \mathbf{x})$.
  - Posterior Component Density: $p_\beta(\theta \mid \mathbf{z}, \mathbf{x}) := \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\beta}, \tilde{\sigma}^2_{\mathbf{z},\beta} I)$.
  - Posterior Component Mean: $\tilde{\mu}_{\mathbf{z},\beta} := \frac{\beta N_{\mathbf{z}}}{\alpha + \beta N_{\mathbf{z}}} \bar{x}_{\mathbf{z}}$.
  - Posterior Component Variance: $\tilde{\sigma}^2_{\mathbf{z},\beta} := \frac{1}{\alpha + \beta N_{\mathbf{z}}} \sigma^2$.
  - Variance Scaling: $\tilde{V}_{\mathbf{z}_{-i},\beta} := \frac{1}{\beta} + \frac{1}{\beta N_{\mathbf{z}_{-i}} + \alpha}$ (used in the posterior predictive density).

- $\mathbf{z}^* := \arg\max_{\mathbf{z}' \neq \mathbf{z}} \{T(\mathbf{z}' \mid \mathbf{z})\}$ denotes the destination label that maximizes the escape probability.

- $T^*_{\ell,\mathbf{z}} := \max_{\mathbf{z}' \neq \mathbf{z}} \{N T_\ell(\mathbf{z}' \mid \mathbf{z})\}$ denotes the maximal probability of escape from the state $\mathbf{z}$ under the collapsed Gibbs transition kernel, $T_\ell(\cdot \mid \cdot)$.

- The unnormalized growth factor $r_{\mathbf{z}}(\beta) := \frac{\tilde{p}_\beta(\mathbf{z}|\mathbf{x})}{\tilde{p}_1(\mathbf{z}|\mathbf{x})}$ measures the change in the unnormalized posterior label weight as a function of the inverse temperature (with the original target at $\beta = 1$ providing the baseline in the denominator).

- We use $[SS_{\mathbf{z}}]$ as convenient shorthand for the sum of squares term that arises in the posterior label weight

$$[\mathrm{SS}_{\mathbf{z}}] := \sum_{i:z_i=1} \|x_i - \bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\frac{1}{N_{\mathbf{z}}} + \frac{1}{\alpha}} \|\bar{x}_{\mathbf{z}}\|^2 + \frac{1}{\tilde{V}_{\mathbf{w}}} \sum_{i:z_i=0} \|x_i - \tilde{\mu}_{\mathbf{w}}\|^2,$$

  measuring whether the label is well-suited to the data.

A.5 Notation for Chapter 4

- We cite the following convenient notation when defining the fractional posterior.

$$\begin{aligned} S_{\boldsymbol{\beta}} &:= \{i : \beta_i > 0\} \\ S_{\mathbf{z},\boldsymbol{\beta}} &:= \{i : z_i = 1, \beta_i > 0\} \\ N_{\mathbf{z},\boldsymbol{\beta}} &:= \sum_{i:z_i=1} \beta_i \\ \bar{x}_{\mathbf{z},\boldsymbol{\beta}} &:= \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}}} \sum_{i:z_i=1} \beta_i x_i \end{aligned}$$

- Aspects of the fractional annealing model.
  - Variable Density: $p_{\beta_i}(x_i \mid \theta, z_i = 1) := \mathcal{N}(x_i; \theta, (\sigma^2/\beta_i)I)$.
  - Likelihood: $p_{\boldsymbol{\beta}}(\mathbf{x}|\theta) \propto \prod_{i \in S_{\boldsymbol{\beta}}} \frac{1}{2} [p_{\beta_i}(x_i|z_i = 0) + p_{\beta_i}(x_i \mid \theta, z_i = 1)]$.
  - Posterior: $p_{\boldsymbol{\beta}}(\theta|\mathbf{x}) \propto \sum_{\mathbf{z}} \tilde{p}_{\boldsymbol{\beta}}(\mathbf{z} \mid \mathbf{x}) p_{\boldsymbol{\beta}}(\theta \mid \mathbf{z}, \mathbf{x})$.
  - Posterior Component Density: $p_{\boldsymbol{\beta}}(\theta \mid \mathbf{z}, \mathbf{x}) := \mathcal{N}(\theta; \tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}}, \tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} I)$.
  - Posterior Component Mean: $\tilde{\mu}_{\mathbf{z},\boldsymbol{\beta}} := \frac{N_{\mathbf{z},\boldsymbol{\beta}}}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \bar{x}_{\mathbf{z},\boldsymbol{\beta}}$.
  - Posterior Component Variance: $\tilde{\sigma}^2_{\mathbf{z},\boldsymbol{\beta}} := \frac{1}{N_{\mathbf{z},\boldsymbol{\beta}} + \alpha} \sigma^2$.
  - Variance Scaling: $\tilde{V}_{\mathbf{z}_{-i},\boldsymbol{\beta}} := \frac{1}{\beta_i} + \frac{1}{N_{\mathbf{z}_{-i},\boldsymbol{\beta}} + \alpha}$ (for the posterior predictive density).

Appendix B

Related Models

B.1 Variable Weights

The greedy mixture model introduced in Section 1.4 assumes constant, equal weights for the fixed and variable Gaussian likelihood mixture components. As we have discussed, our model reflects a step in a general greedy procedure, and thus the choice of weights will vary with the application of interest.

A relevant alternative model that requires further consideration is the use of *variable* weights. Under our original model, the variable component center θ is our inference target, but there are a variety of potential component parameters that are natural to study. In this appendix, we extend our analysis to the case where the likelihood mixture weights $\boldsymbol{\omega} := (\omega_0, \omega_1)$ are additional variable parameters, with a known prior distribution $p(\boldsymbol{\omega})$. Depending on the application, this may better reflect our *a priori* knowledge of the setting, or the greedy computational procedure might simply benefit from this added flexibility (compared to the rigid assumption that we know *a priori* the weight of each new component). We note that $\boldsymbol{\omega}$ is effectively one-dimensional (as $\omega_0 + \omega_1 = 1$), but it is often convenient to refer to it as a length 2 vector, as that better mirrors the general case with K mixture components.

Our updated likelihood now conditions on two parameters: θ and $\boldsymbol{\omega}$.

$$p(x_i \mid \theta, \boldsymbol{\omega}) := \omega_0 p(x_i \mid z_i = 0) + \omega_1 p(x_i \mid \theta, z_i = 1)]$$

As before, we can write the mixture likelihood as either a product of sums, or a sum over exponentially many potential labelings.

$$\begin{aligned} p(\mathbf{x} \mid \theta, \boldsymbol{\omega}) &= \prod_{i=1}^{N} [\omega_0 p(x_i \mid z_i = 0) + \omega_1 p(x_i \mid \theta, z_i = 1)] \\ &= \sum_{\mathbf{z}} \underbrace{\left[\prod_{i=1}^{N} p(z_i \mid \boldsymbol{\omega})\right]}_{p(\mathbf{z}|\boldsymbol{\omega})} \underbrace{\left[\prod_{i=1}^{N} p(x_i \mid z_i, \theta)\right]}_{p(\mathbf{x}|\mathbf{z},\theta)} \\ &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \boldsymbol{\omega}) p(\mathbf{x} \mid \mathbf{z}, \theta) \end{aligned}$$

The posterior distribution is a function of the likelihood $p(\mathbf{x} \mid \theta, \boldsymbol{\omega})$, the Gaussian prior $p(\theta)$, and the prior we assign to the weights, $p(\boldsymbol{\omega})$. The canonical choice for mixture models is the Dirichlet distribution, which is conjugate to the multinomial distribution on the data indices.[1] In the greedy setting, the labels are binary vectors, so this could be equivalently viewed as a beta-binomial model (the one-dimensional version of the Dirichlet and multinomial). In this section, we will stick with the notation of the Dirichlet-multinomial model, as that we better mirrors the literature standard (which generally assumes K variable components).

Thus, we define

$$\begin{aligned} \boldsymbol{\omega} &:= (\omega_0, \omega_1), \\ p(\boldsymbol{\omega}) &:= \text{Dirichlet}(\boldsymbol{\omega}; \alpha_0, \alpha_1) = \frac{1}{B(\alpha_0, \alpha_1)} \omega_0^{\alpha_0 - 1} \omega_1^{\alpha_1 - 1}, \end{aligned}$$

1. The literature typically refers to this distribution as "multinomial", and we mirror that terminology. However, it is worth noting that in other settings, $p(\mathbf{z} \mid \boldsymbol{\omega})$ would be referred to as a *sequence of categorical variables*. While the distinction is slight, and the Dirichlet is conjugate for both multinomial and categorical likelihoods, the posterior formulae diverge (and we note that we are citing the categorical case).

for $\alpha_0, \alpha_1 > 0$, $\omega_0 = 1 - \omega_1$, and $\omega_0, \omega_1 \in [0, 1]$, with

$$B(\alpha_0, \alpha_1) := \frac{\Gamma(\alpha_0)\Gamma(\alpha_1)}{\Gamma(\alpha_0 + \alpha_1)},$$

where $\Gamma(\cdot)$ is the Gamma function.

The target for our inference is still the parameter θ, but the key impact from this change is on the collapsed Gibbs transition probabilities. Under the standard Gibbs sampler (Algorithm 1), the introduction of variable weights would require an additional intermediate sampling step, where we generate weights $\boldsymbol{\omega}$ conditioned on the current label $\mathbf{z}$. Under the collapsed Gibbs sampler (Algorithm 2), the weights are instead an additional variable to integrate out in the computation of the transition probabilities. The full formula is shown in Lemma B.1.1, with the proof provided at the end of this appendix.

Lemma B.1.1. *For the Bayesian mixture posterior described above, with known prior distribution $\boldsymbol{\omega} \sim$ Dirichlet(α_0, α_1) on the weights, and data index $i \in \{1, \dots, N\}$, the collapsed Gibbs conditional transition probabilities are given by*

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}) = \begin{cases} \frac{(\alpha_1 + N_{\mathbf{z}_{-i}})\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)}{(\alpha_1 + N_{\mathbf{z}_{-i}})\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I) + (\alpha_0 + N^0_{\mathbf{z}_{-i}})p(x_i|z_i=0)}, & \text{for } z_i = 1, \\ \frac{(\alpha_0 + N^0_{\mathbf{z}_{-i}})p(x_i|z_i=0)}{(\alpha_1 + N_{\mathbf{z}_{-i}})\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I) + (\alpha_0 + N^0_{\mathbf{z}_{-i}})p(x_i|z_i=0)}, & \text{for } z_i = 0, \end{cases} \tag{B.1}$$

with $N^0_{\mathbf{z}} := N - N_{\mathbf{z}}$, $\tilde{\mu}_{\mathbf{z}_{-i}} := \frac{N_{\mathbf{z}_{-i}}}{N_{\mathbf{z}_{-i}} + \alpha} \bar{x}_{\mathbf{z}_{-i}}$, *and* $\tilde{V}_{\mathbf{z}_{-i}} := 1 + \frac{1}{N_{\mathbf{z}_{-i}} + \alpha}$.

Crucially, the *only* departure from the previous transition probabilities (Lemma 1.5.1) are the scaling factors $(\alpha_1 + N_{\mathbf{z}_{-i}})$ and $(\alpha_0 + N^0_{\mathbf{z}_{-i}})$ applied to the two original densities. These scaling factors can be understood as implicit estimates of the mixture weights. Thus, if we instead used constant but non-uniform mixture weights, those exact weights would replace the role of these scaling factors (and our analysis would otherwise be similar).

B.1.1 Conditions for Slow Mixing

To understand the impact of variable weights on our theoretical mixing analysis, we consider the effects on the conditions for slow mixing established in Theorem 2.2.1. We start with the step in the original proof that must be updated to reflect these new transition probabilities.

In the $z_i = 0$ case, we modify Equation 2.9 and observe

$$\begin{aligned}\max_{i:z_i=0} \mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) &\leq \max_{i:z_i=0} \frac{(N_{\mathbf{z}_{-i}} + \alpha_1)\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{(N^0_{\mathbf{z}_{-i}} + \alpha_0)p(x_i \mid z_i = 0)} \\ &= \frac{(N_{\mathbf{z}} + \alpha_1)}{(N^0_{\mathbf{z}} - 1 + \alpha_0)} \max_{i:z_i=0} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{p(x_i \mid z_i = 0)}. \end{aligned} \tag{B.2}$$

The ratio of densities can be bounded by our earlier work, and we need only bound the ratio of scaling factors. In the $z_i = 1$ case, we have a similar form (updating Equation 2.18).

$$\begin{aligned}\max_{i:z_i=1} \mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) &\leq \max_{i:z_i=1} \frac{(N^0_{\mathbf{z}_{-i}} + \alpha_0)p(x_i \mid z_i = 0)}{(N_{\mathbf{z}_{-i}} + \alpha_1)\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \\ &= \frac{(N^0_{\mathbf{z}} + \alpha_0)}{(N_{\mathbf{z}} - 1 + \alpha_1)} \max_{i:z_i=1} \frac{p(x_i \mid z_i = 0)}{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}_{-i}}, \tilde{V}_{\mathbf{z}_{-i}}\sigma^2 I)} \end{aligned} \tag{B.3}$$

Thus, in our updated version of the theorem, we can simply take the crude maximum of these ratios,

$$r_{\boldsymbol{\omega}} := \max\left\{ \frac{N_{\mathbf{z}} + \alpha_1}{N^0_{\mathbf{z}} - 1 + \alpha_0}, \frac{N^0_{\mathbf{z}} + \alpha_0}{N_{\mathbf{z}} - 1 + \alpha_1} \right\}. \tag{B.4}$$

as an additional factor scaling our mixing time bound (Theorem B.1.2 matches the original result, except for this narrow change).

Theorem B.1.2. *Consider the greedy Gaussian mixture posterior with variable weights described above, with known prior*

$$\boldsymbol{\omega} \sim \textit{Dirichlet}(\alpha_0, \alpha_1).$$

Consider the corresponding Markov chain generated by the collapsed Gibbs sampler (Algorithm 2) on this posterior. Let τ_{mix} denote the number of steps required so that the total variation distance to stationarity is at most $1/4$. For observed data $\mathbf{x}$, let $\mathbf{z}$ and $\mathbf{w}$ denote labels such that $R < \frac{1}{2}$, $r_\delta < \frac{9}{40}$, and whose sample sizes satisfy Equations 2.6 & 2.7. Define $r_{\boldsymbol{\omega}}$ as in Equation B.4.

Then, the mixing time of the resulting Markov chain is exponentially slow in our separation parameters u and Δ, with a lower bound

$$\tau_{mix} \geq \frac{1}{8r_{\boldsymbol{\omega}}} \min\left\{ \exp\left(\left[\frac{7 - 14R}{20} \right] \frac{\Delta^2}{\sigma^2} \right), \exp\left(\left[\frac{9 - 40r_\delta}{20} \right] \frac{u^2}{\sigma^2} \right) \right\}. \tag{B.5}$$

While this strategy is crude, capturing the impact of the variable weights using $r_{\boldsymbol{\omega}}$ is sufficient because our primary interest lies in distinguishing polynomial and exponential time mixing. As the theorem statement already places assumptions on the minimum sample size (Equations 2.6 & 2.7), we need not worry about the case of a problematically small denominator, and $r_{\boldsymbol{\omega}}$ will be of polynomial order in any reasonable setting. As noted earlier, these scaling factors can be viewed as estimates of the relative weights, and if we instead defined our mixture using non-uniform constant weights, we would arrive at the same result (but with the exact weight ratio, rather than the estimates). The conditions that guarantee that the mixing bottleneck will persist under temperature annealing (Theorem 3.4.4) or subsample annealing (Theorem 4.3.2) are both extensions of this original bottleneck, and thus they could be adjusted in a similar fashion.

The choice of weights is highly influential in other aspects of model construction, but it has little impact on the fundamental mixing behavior. The collapsed Gibbs transition probabilities are simply scaled by the ratio of the weights (either exact, or their current estimate). Typical models of interest will not exhibit extreme weight ratios, and a likelihood mixture component with tiny weight has little impact on the posterior (this stands in contrast to the task of estimation, where the minimum weight is often a key assumption). This is the

fundamental motivation for the use of constant uniform weights in our model—alternative choices do not change the underlying impediments to mixing, and thus we use the definition that clarifies our analysis.

B.2 Gibbs Sampler Variants

The mixing analysis in this dissertation leverages the collapsed Gibbs sampler as a Markov chain transition rule (as described in Algorithm 2), and it is instructive to provide some further context on its construction. The collapsed Gibbs sampler we have defined updates a single index at a time, selecting index i uniformly at random, and then generating a new $z_i \sim p(\cdot \mid \mathbf{z}_{-i}, \mathbf{x})$ based on the collapsed Gibbs conditional transition probabilities. This is the *random scan* implementation of the Gibbs sampler. A common alternative is *systematic scan*, which updates the coordinates in a deterministic order. We mirror the typical literature and follow a random scan as it enables easier theoretical analysis, and it is generally assumed that the fundamental underlying behavior of the two approaches is similar. While the precise factor by which their convergence rates can diverge is not entirely resolved in the literature, this is beyond the scope of our work, and random scan is sufficient for our purposes (given our goal of distinguishing between exponential and polynomial convergence).

Further, Algorithm 2 only updates one index at a time, rather than drawing a wholly new vector (i.e. $z_i' \sim p(\cdot \mid \mathbf{z}_{-i}, \mathbf{x})$, not $\mathbf{z}' \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$). The simplicity of updating a single data index is critical for the clean analysis shown in the document, and it mirrors the common approach in the literature. Further, this is necessary in order to "collapse" the Gibbs sampler. That is, we could generate a full vector $\mathbf{z}' \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$ through an intermediate step—we draw $\theta \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$ (which is normally distributed), and then use that θ to draw $\mathbf{z}' \sim p(\cdot \mid \theta, \mathbf{x})$ (which factors into a product, and thus can be sampled). However, we cannot easily "collapse" the Gibbs sampler if we wish to draw an entirely new $\mathbf{z}$ vector. In short, $\mathbf{z}' \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$ is *not* independent across the data indices, and thus

even if we can compute its density for any given $\mathbf{z}'$, we cannot easily generate samples. The posterior labels are only independent when conditioned on a specific θ, and thus when we attempt to integrate out this θ, we introduce a complicated dependence structure. For the purposes of this dissertation, the single-index update random scan technique is natural—it captures the key underlying mixing behavior, and mirrors the broader research literature.

B.3 Proofs for Appendix B

Proof of Lemma B.1.1. For this updated derivation, it is illustrative to begin with the full model distribution.

$$p(\mathbf{z}, \theta, \boldsymbol{\omega}, \mathbf{x}) = p(\boldsymbol{\omega})p(\mathbf{z} \mid \boldsymbol{\omega})p(\theta)p(\mathbf{x} \mid \mathbf{z}, \theta).$$

The (proportional) posterior on the data labels results from integrating out both θ and $\boldsymbol{\omega}$ (we recall that $\boldsymbol{\omega}$ is a one-dimensional object in this integration).

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{x}) &\propto \int \int p(\mathbf{z}, \theta, \boldsymbol{\omega}, \mathbf{x}) \; d\boldsymbol{\omega} \; d\theta \\ &= \left[\int p(\boldsymbol{\omega})p(\mathbf{z} \mid \boldsymbol{\omega}) d\boldsymbol{\omega} \right] \left[\int p(\theta)p(\mathbf{x} \mid \mathbf{z}, \theta) d\theta \right] \end{aligned}$$

The integral over θ mirrors our previous work (Section 1.6.2),

$$= \underbrace{\left[\int p(\boldsymbol{\omega})p(\mathbf{z} \mid \boldsymbol{\omega}) d\boldsymbol{\omega} \right]}_{p(\mathbf{z})} \left[A^0(\mathbf{z}) A^1(\mathbf{z}) \right]. \tag{B.6}$$

The additional factor is this integral over $\boldsymbol{\omega}$ (previously, the marginal distribution $p(\mathbf{z})$ was uniform, and thus this term disappeared). The marginal distribution of $\mathbf{z}$ can be computed using the conjugate posterior for the Dirichlet distribution. As mentioned above, $p(\mathbf{z} \mid \boldsymbol{\omega}) := \omega_0^{N_{\mathbf{z}}^0} \omega_1^{N_{\mathbf{z}}}$ should technically be referred to as a sequence of categorical variables, not multinomial (which would require a different normalizing constant on the conjugate

posterior).

$$
\begin{aligned}
p(\mathbf{z}) &= \int p(\boldsymbol{\omega}) p(\mathbf{z} \mid \boldsymbol{\omega}) \; d\boldsymbol{\omega} \\
&= \int \frac{1}{B(\alpha_0, \alpha_1)} \omega_0^{\alpha_0 - 1} \omega_1^{\alpha_1 - 1} \omega_0^{N_{\mathbf{z}}^0} \omega_1^{N_{\mathbf{z}}} \; d\omega_1 \\
&= \frac{1}{B(\alpha_0, \alpha_1)} \int \omega_0^{N_{\mathbf{z}}^0 + \alpha_0 - 1} \omega_1^{N_{\mathbf{z}} + \alpha_1 - 1} \; d\omega_1 \\
&= \frac{B(N_{\mathbf{z}}^0 + \alpha_0, N_{\mathbf{z}} + \alpha_1)}{B(\alpha_0, \alpha_1)} \\
&= \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \frac{\Gamma(\alpha_0 + N_{\mathbf{z}}^0)\Gamma(\alpha_1 + N_{\mathbf{z}})}{\Gamma(\alpha_0 + \alpha_1 + N)} \\
&= \left[\frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0 + \alpha_1 + N)} \right] \left[\frac{\Gamma(\alpha_0 + N_{\mathbf{z}}^0)}{\Gamma(\alpha_0)} \right] \left[\frac{\Gamma(\alpha_1 + N_{\mathbf{z}})}{\Gamma(\alpha_1)} \right]
\end{aligned}
$$

We can substitute this result back into Equation B.6, and we observe

$$
p(\mathbf{z} \mid \mathbf{x}) \propto \Gamma(\alpha_0 + N_{\mathbf{z}}^0)\Gamma(\alpha_1 + N_{\mathbf{z}}) A^0(\mathbf{z}) A^1(\mathbf{z}).
$$

Using this marginal distribution, we can modify our derivation of the transition probabilities under the collapsed Gibbs sampler.

$$
\begin{aligned}
&\mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) \\
&\qquad \propto \mathbb{P}(z_i = 1, \mathbf{z}_{-i} \mid \mathbf{x}) \\
&\qquad = \Gamma(\alpha_0 + N_{\mathbf{z}_{-i}}^0)\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}} + 1) A^0(\mathbf{z}_{-i}) A^1(\mathbf{z}^{[i \to 1]}) \\
&\qquad = \left[\Gamma(\alpha_0 + N_{\mathbf{z}_{-i}}^0)\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}}) A^0(\mathbf{z}_{-i}) A^1(\mathbf{z}_{-i}) \right] \frac{\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}} + 1)}{\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}})} \frac{A^1(\mathbf{z}^{[i \to 1]})}{A^1(\mathbf{z}_{-i})}
\end{aligned}
$$

We use this factorization because it offers a natural comparison to the $z_i = 0$ case.

$$\begin{aligned}
&\mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) \\
&\qquad \propto \Gamma(\alpha_0 + N^0_{\mathbf{z}_{-i}} + 1)\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}})A^0(\mathbf{z}^{[i\to 0]})A^1(\mathbf{z}_{-i}) \\
&\qquad = \left[\Gamma(\alpha_0 + N^0_{\mathbf{z}_{-i}})\Gamma(\alpha_1 + N_{\mathbf{z}_{-i}})A^0(\mathbf{z}_{-i})A^1(\mathbf{z}_{-i})\right] \frac{\Gamma(\alpha_0 + N^0_{\mathbf{z}_{-i}} + 1)}{\Gamma(\alpha_0 + N^0_{\mathbf{z}_{-i}})} \frac{A^0(\mathbf{z}^{[i\to 0]})}{A^0(\mathbf{z}_{-i})}
\end{aligned}$$

In the normalization, the matching bracketed term disappears, the ratios of the Gamma functions simplify, and the ratios of A^1 and A^0 match our earlier derivation (in Section 1.6.1). This provides the conditional probabilities shown in Lemma B.1.1. □

Proof of Theorem B.1.2. We cite the proof of Theorem 2.2.1, with slight modification. By Equation B.2, we have

$$\max_{i:z_i=0} \mathbb{P}(z_i = 1 \mid \mathbf{z}_{-i}, \mathbf{x}) \leq \frac{(N_{\mathbf{z}} + \alpha_1)}{(N^0_{\mathbf{z}} - 1 + \alpha_0)} \max_{i:z_i=0} \frac{\mathcal{N}(x_i; \tilde{\mu}_{\mathbf{z}}, \tilde{V}_{\mathbf{z}}\sigma^2 I)}{p(x_i \mid z_i = 0)}.$$

By the definition of $r_{\boldsymbol{\omega}}$ (Equation B.4), and the work in the original proof (Equation 2.17), we have

$$\leq 2r_{\boldsymbol{\omega}} \exp\left(-\left[\frac{7 - 14R}{20}\right]\frac{\Delta^2}{\sigma^2}\right)$$

Similarly, by Equation B.3 and the original derivation (Equation 2.27), we have

$$\max_{i:z_i=1} \mathbb{P}(z_i = 0 \mid \mathbf{z}_{-i}, \mathbf{x}) \leq 2r_{\boldsymbol{\omega}} \exp\left(-\left[\frac{9 - 40r_\delta}{20}\right]\frac{u^2}{\sigma^2}\right)$$

The rest of the proof mirrors the original case, with the inclusion of this $r_{\boldsymbol{\omega}}$ scaling factor. □

Appendix C

Simulation Methodology

C.1 Assessing Markov Chain Convergence

MCMC sampling confronts a fundamental challenge—the initial distribution of the chain is far from the stationary target, and we must wait for approximate convergence before we can generate viable samples. In this dissertation, the primary object of interest is the mixing time, which is the number of iterations until an approximate convergence criterion is satisfied. In our theoretical analysis, we define this convergence criterion using a fixed total variation distance, but in our empirical simulations, this cannot be directly computed. However, while total variation is a common choice in the literature, it is not the only viable definition. In practice, a variety of convergence criteria have similar properties—the key is to follow a *consistent* standard when making any comparisons.

For our computationally challenging target distribution, it is difficult to reliably assess convergence using a single chain (in fact, the very premise of the underlying challenge is the fact that local behavior can be misleading). For our empirical simulations, we draw inspiration from the literature and instead estimate convergence using the observed properties of *multiple* independent chains. The multi-chain diagnostic criterion we follow is the *potential scale reduction factor* (PSRF), originally proposed by Gelman & Rubin [31] (with

a recent introduction provided by Gelman et al. [48]). In short summary, the PSRF assesses convergence through a comparison of the *between-chain* and *within-chain* variation. As this criterion is approximate, we will supplement it with other measures and sanity checks to ensure it behaves as we would expect. Our goal is to compare the relative mixing times under different data settings, and thus as long as we follow a *consistent* benchmark, this multi-chain diagnostic provides a viable substitute for total variation distance.

Let $\hat{R}_\theta$ denote the PSRF, which we define as follows. We run J chains with independent initialization, discarding the first half of each chain (the "burn-in" phase), and compare the within-chain and between-chain variation of the remaining S samples for each chain. Our criterion is computed separately for each dimension of the parameter θ, but for notational simplicity, we omit the subscript on dimension in this definition (i.e. it is implicitly specific to some dimension d). If $\theta_{s,j}$ denotes the sth element of the jth chain (for the dth dimension), we define W as the within-chain variation, constructed using the mean of the empirical variances among the chains.

$$\begin{aligned}
\bar{\theta}_{\cdot,j} &:= \frac{1}{S}\sum_{s=1}^{S}\theta_{s,j} \\
s_j^2 &:= \frac{1}{S}\sum_{s=1}^{S}(\theta_{s,j} - \bar{\theta}_{\cdot,j})^2 \\
W &:= \frac{1}{J}\sum_{j=1}^{J} s_j^2
\end{aligned}$$

Then, let B denote the between-chain variation.

$$\begin{aligned}
\bar{\theta}_{\cdot,\cdot} &:= \frac{1}{JS}\sum_{j=1}^{J}\sum_{s=1}^{S}\theta_{s,j} \\
B &:= \frac{S}{J-1}\sum_{j=1}^{J}(\bar{\theta}_{\cdot,j} - \bar{\theta}_{\cdot,\cdot})^2
\end{aligned}$$

These form our sample estimate of the posterior variance,

$$\widehat{\text{Var}}(\theta \mid \mathbf{x}) := \frac{S-1}{S}W + \frac{1}{S}B.$$

Finally, we define the *potential scale reduction factor* (PSRF) as

$$\hat{R}_\theta := \sqrt{\widehat{\text{Var}}(\theta \mid \mathbf{x})/W}.$$

As the chain converges, the PSRF approaches 1, as the within-chain variance will be unbiased for the true posterior variance (before convergence, it is an underestimate). We follow the typical recommendation in the literature (e.g. Gelman et al. [48]), and require $\hat{R}_\theta < 1.10$ as our convergence criterion. This formula defines the PSRF for a *single* dimension in the parameter space, and thus our full convergence criterion is that we require $\hat{R}_\theta < 1.10$ for *each* dimension.

This approximation is highly effective (and matches our intuition for mixing) as long as no relevant region of the parameter space is omitted from the full set of chains. In practical applications, this could be difficult to verify (the challenge in simply finding the many isolated modes can be significant), but for our empirical simulations, we have prior knowledge of the setting. Thus, it is typically straightforward to initialize these chains so that the representation of each relevant region of the state space is guaranteed.

C.2 Empirical Experiment Specification

C.2.1 General Methodology

In this appendix, we outline the empirical simulation methodology that is shared among various experiments, and in subsequent appendices we provide the concrete details on the set-up for each individual experiment.

We assess the convergence of our Markov chains using the multi-chain diagnostic criterion

described in Appendix C.1. First, we note that our chains are defined in the discrete space of the binary labels, while the definition of $\hat{R}_\theta$ above is defined for a continuous θ. Thus, we assess convergence on the conditional draws, $\theta \sim p(\cdot \mid \mathbf{z}, \mathbf{x})$ for each state $\mathbf{z}$ which we generate. While there are many reasons we find it advantageous to operate directly on the discrete space of the labels, our ultimate target is the parameter θ, and thus it is natural to use these draws in our approximate convergence criterion.

As mentioned above, the primary potential weakness of the PSRF convergence criterion is that if relevant regions of the state space are not represented among the chains, the results can be misleading. We can (and will) manually inspect the results of our simulations to verify that this is not causing any issues. But more importantly, we can ensure that all isolated regions are properly represented through our choice of initialization for the chains (leveraging our prior knowledge of the setting). The datasets used for our experiments are constructed using pre-defined clusters of data (which imply isolated labelings). We will initialize one of the chains at *each* of these clusters. Thus, all isolated regions of the data are guaranteed representation. As we will have more chains than data clusters, we initialize the rest through a convenient data-dependent strategy. We simply select a datum at random, and generate a labeling conditional on that datum (which provides a more reasonable starting estimate than choosing entirely at random).

It is impractical to assess convergence with this multi-chain convergence criterion $\hat{R}_\theta$ for *every* chain iteration. Thus, we run batches of iterations (with a size of 10^4) until the convergence criterion is satisfied, and then we scan through the latest batch to determine if convergence was first reached at an earlier point. While this is technically a slight approximation, it is trivial relative to the general noise that arises from the use of $\hat{R}_\theta$.

For convenience, we set $\sigma^2 = 1$ and $\alpha = 1/5$ for all empirical simulations.

C.2.2 Simulations in Chapter 2

Three-cluster: u^2 Separation

We consider observed data $\mathbf{x}$ formed by three clusters of data, each with a sample size of 10 ($N = 30$ in total) drawn from a multivariate Gaussian (with $d = 2$). The first cluster center (with label $\mathbf{w}$) is placed at the origin $\bar{x}_{\mathbf{w}} = (0, 0)$, the second cluster ($\mathbf{z}_1$) center is distance u from the origin, with $\bar{x}_{\mathbf{z}_1} = (u/\sqrt{2}, u/\sqrt{2})$, and the third cluster ($\mathbf{z}_2$) center is its reflection about the origin, $\bar{x}_{\mathbf{z}_2} = (-u/\sqrt(2), -u/\sqrt{2})$. For each level of $u^2 \in \{3, 3.5, \ldots, 5.5\}$, we generate 50 datasets following this pattern. For each dataset, we initialize 5 independent chains, with the first three initialized at the three cluster labels, and the final two initialized from the data (as described above), and run the chains until our convergence criterion is satisfied (as described above, requiring $\hat{R}_\theta < 1.10$ for each dimension).

Three-cluster: Δ^2 Separation

We follow the exact same specifications as the "Three-cluster: u^2 Separation" experiment above, only differing on the cluster centers $\bar{x}_{\mathbf{z}_1}$ and $\bar{x}_{\mathbf{z}_2}$. We fix $\bar{x}_{\mathbf{z}_1} = (4, 0)$, and for each chosen value of $\Delta^2 \in \{25, 25.5, \ldots, 29.5\}$, we set the third cluster center as

$$\bar{x}_{\mathbf{z}_2} = (4\cos(\arcsin(\Delta)/8), 4\sin(\arcsin(\Delta)/8)).$$

In short, this ensures $\|\bar{x}_{\mathbf{z}_2} - \bar{x}_{\mathbf{w}}\| = \|\bar{x}_{\mathbf{z}_1} - \bar{x}_{\mathbf{w}}\| = 4$, and $\|\bar{x}_{\mathbf{z}_2} - \bar{x}_{\mathbf{z}_1}\| = \Delta$.

C.2.3 Simulations in Chapter 3

Simulated Tempering, for Three-Cluster u^2 Separation

We follow the data setting from the "Three-cluster: u^2 Separation" experiment (Appendix C.2.2), and use the collapsed Gibbs sampler results from that experiment (originally shown in Figure 2.3a).

We additionally implement a simulated tempering chain (Algorithm 3) via internal annealing (Section 3.3). We follow a linear inverse temperature schedule with $L = 5$, i.e. $\beta_\ell = \frac{1}{4}(\ell - 1)$ for $\ell \in \{1, \ldots, 5\}$. We estimate normalizing constants through the methodology described in Appendix C.3, and apply $M = 5$ collapsed Gibbs transitions between each temperature index update. As these results tend to be noisier (and the runtime is not prohibitive), we increase the number of datasets generated per level of u^2 to 150, and otherwise follow the simulation methodology used for the collapsed Gibbs sampler. In the plot (Figure 3.2), we count the temperature index transitions as equivalent "iterations", but more broadly, we should not assume that these results are directly comparable.

C.2.4 Simulations in Chapter 4

Subsample Annealing: Data Ordering Comparison

We consider observed data formed by three clusters drawn from a multivariate Gaussian (with dimension $d = 2$), with equal sample sizes of 18 each (for $N = 54$ in total). The three cluster centers form an equilateral triangle, with each center placed so that it is distance $u = 1.65$ from the origin. We generate 50 such datasets, and for each dataset, apply the four different MCMC sampling techniques described below. For each technique, we initialize 5 independent Markov chains following the same strategy used in earlier simulations, and track the relationship between the PSRF and the iteration count (we note that the experiments described in Section 2.2.3 simply compute the number of iterations until convergence is reached, while here we track the evolution of the convergence diagnostic directly).

The first MCMC technique is the collapsed Gibbs sampler, defined the same as before. The second technique is temperature annealing, following a linearly spaced inverse temperature schedule with $L = 9$. Specifically, this implies fractional annealing where $\boldsymbol{\beta}_\ell := \frac{1}{8}(\ell - 1, \ell - 1, \ldots, \ell - 1)$ for $\ell \in \{1, \ldots, 9\}$ (we implement this through fractional annealing, but this is equivalent to the internal annealing described in Section 3.3).

The final two techniques are both implementations of subsample annealing, only dif-

fering on the ordering of the data. The schedule of subsample sizes is given by $n = 0, 3, 9, 18, 27, 36, 45, 54$. However, in tuning this technique, the initial addition of data has a dramatic impact on the posterior. Thus, it is convenient to include a single *ramp-up* step. To be specific, we follow a fractional annealing schedule where $\boldsymbol{\beta}_1 := (0, 0, \ldots, 0)$ includes none of the data, $\boldsymbol{\beta}_2$ is the ramp-up step, $\boldsymbol{\beta}_3$ corresponds with observed sample size $n = 3$, $\boldsymbol{\beta}_4$ corresponds with $n = 9$, and so on down the subsample size schedule. Then, the ramp-up step is defined as $\boldsymbol{\beta}_2 := \boldsymbol{\beta}_3/2$ (i.e. each of the $n = 3$ data are only "half included"). This is a useful tool when implementing subsample annealing on unruly datasets, and it is a convenient advantage offered by fractional annealing. In total, this subsample size schedule (and the single ramp-up step) implies a length $L = 9$ fractional annealing schedule.

Our two subsample annealing implementations share this sequence of subsample sizes, but diverge in their ordering of the data (i.e. the composition of the subsamples). The first follows a *random* ordering of the data, and the second follows a *pre-set* ordering ("SSA Shuffled" and "SSA Pre-set" in Figure 4.3, respectively). The pre-set ordering is chosen so that in each subsample, the count of data from each of the three components is balanced (i.e. when $n = 3$, the subsample includes one datum from each cluster, and when $n = 9$, it includes three data from each cluster, and so on). For the three simulated tempering implementations, the normalizing constants are estimated using the methodology described in Appendix C.3.

C.3 Normalizing Constant Estimation

Throughout our study of simulated tempering (in particular, Theorems 3.4.4 & 4.3.2), we are interested in the fundamental mixing behavior of the joint stationary distribution, $\pi(\cdot, \cdot)$. Thus, we assume its marginal distribution is set to be uniform on the annealing indices, $\pi(\ell) = 1/L$. However, a practical implementation of the algorithm will typically need to estimate the relative normalizing constants, and imprecise estimates lead to a non-uniform

marginal distribution. In this appendix, we briefly discuss this process, and describe how we compute these estimates for our empirical simulations.

For the generic state space $y \in \mathcal{Y}$, let the *unnormalized* interpolating distributions for the simulated tempering chain be given by $\tilde{p}_\ell$, for $\ell \in [L]$. Starting from the state (ℓ, y), the annealing index transition proposes an adjacent index $\ell' = \ell \pm 1$, and accepts or rejects the proposal following the Metropolis-Hastings ratio $p_{\ell'}(y)/p_\ell(y)$. However, this is the ratio of *normalized* densities, and we may only be able to query the unnormalized form. To address this, consider constants $C_1, \ldots, C_L$ such that

$$\frac{C_{\ell'}}{C_\ell} = \frac{\int \tilde{p}_{\ell'}(y)dy}{\int \tilde{p}_\ell(y)dy}. \tag{C.1}$$

In the literature (and this dissertation), the C_ℓ are often referred to as "normalizing constants", but this is shorthand for *relative normalizing constants*—they need only preserve the correct *ratio* between the interpolating distributions. Access to these relative normalizing constants allows us to compute the desired Metropolis-Hastings ratio,

$$\begin{aligned}\frac{\tilde{p}_{\ell'}(y)/C_{\ell'}}{\tilde{p}_\ell(y)/C_\ell} &= \frac{\frac{\tilde{p}_{\ell'}(y)}{\int \tilde{p}_{\ell'}(y)dy}}{\frac{\tilde{p}_\ell(y)}{\int \tilde{p}_\ell(y)dy}} \\ &= \frac{p_{\ell'}(y)}{p_\ell(y)},\end{aligned}$$

and thus we can apply a transition rule that preserves the desired marginal distribution, $\pi(\ell) = 1/L$.

In practical settings, the relative normalizing constants are unknown, and we must instead substitute the *estimates* $\hat{C}_1, \ldots, \hat{C}_L$. If these estimates are imprecise, the distribution of the output samples on the target state space $\mathcal{Y}$ are unaffected, but it will distort the marginal distribution $\pi(\ell)$. For example, if we anneal the target through direct exponentiation (and inverse temperature β_ℓ), we observe $\pi(\ell) = \int \pi(\ell, y)dy \propto \frac{1}{\hat{C}_\ell} \int p(y)^{\beta_\ell} dy$. Clearly, if our relative normalizing constants satisfy Equation C.1, the resulting $\pi(\ell)$ is uniform. However,

if we neglect to estimate the normalizing constants (perhaps setting $\hat{C}_\ell = 1$ for all ℓ), then we observe that the marginal distribution of $\pi(\ell)$ can vary with the index. As our output samples require $\ell = L$, poor representation could slow down the generation of the samples.

Our theoretical analysis assumes that π has uniform marginals, but in order to *implement* our empirical simulations we will need to estimate these normalizing constants. Below, we introduce the technique we use for this estimation, but first we provide a brief note of context. While this estimation task can be difficult in unknown applications, our simulations involve known toy examples, and it is relatively straightforward to ensure that the speed of mixing reflects the computational challenge for the idealized π. In particular, the most pressing concern would be if the index L was underrepresented in the joint states of the simulated tempering chain (as this is our target for sampling), but we can verify that we are generating a sufficient number of these target samples. Thus, while this estimation task is notable in practical applications, we can construct our simulations such that they capture the desired behavior.

The actual estimation technique we leverage for our empirical simulations combines ratio importance sampling and a simulated tempering "outer loop". While a wide range of schemes can estimate normalizing constants, this versatile framework is naturally adapted to our planned use of simulated tempering, and requires minimal further work. The iterative scheme is built with the following steps. We assume that for the index ℓ, we have already computed our normalizing constant estimates $\hat{C}_1, \ldots, \hat{C}_\ell$, and we wish to estimate $\hat{C}_{\ell+1}$. We run simulated tempering using just the annealing indices $1, \ldots, \ell$ (and these normalizing constant estimates), until we generate S samples $y_1, \ldots, y_S$ from the target. Then, we estimate the next normalizing constant ratio using ratio importance sampling,

$$r_\ell = \frac{1}{S} \sum_{s=1}^{S} \frac{\tilde{p}_{\ell+1}(y_s)}{\tilde{p}_\ell(y_s)},$$

and thus our estimate for the subsequent normalizing constant is given by

$$\hat{C}_{\ell+1} = r_\ell \hat{C}_\ell.$$

For a base case, we simply set $\hat{C}_1 = 1$. The use of this technique is common, and for an example (where its properties are used in the theoretical proof itself) one can examine the pseudocode of Ge et al. [25]. For clarity, we write out the full "outer loop" process in the following boxed instructions.

Normalizing Constant Estimation ("Outer Loop"):

1. Initialization.
   - Let $\tilde{p}_1(\cdot), \ldots, \tilde{p}_L(\cdot)$ denote a sequence of unnormalized interpolating distributions.
   - Let $T_1(\cdot \mid \cdot), \ldots, T_L(\cdot \mid \cdot)$ denote our state space transition kernels, where T_ℓ preserves stationarity for p_ℓ.
   - Initialize $\hat{C}_1 \leftarrow 1$, and $\ell \leftarrow 1$.
2. Estimate the subsequent normalizing constant.
   - Generate S samples $\{y_1, \ldots, y_S\}$ using simulated tempering (Algorithm 3) on interpolating distributions $\tilde{p}_1(\cdot), \ldots, \tilde{p}_\ell(\cdot)$, with $\hat{C}_1, \ldots, \hat{C}_\ell$ as the input normalizing constant estimates.
   - Compute $r_\ell \leftarrow \frac{1}{S} \sum_{s=1}^{S} \frac{\tilde{p}_{\ell+1}(y_s)}{\tilde{p}_\ell(y_s)}$.
   - Set $\hat{C}_{\ell+1} \leftarrow \hat{C}_\ell r_\ell$.
3. If $\ell = L$, return the estimated normalizing constants, $\{\hat{C}_1, \ldots, \hat{C}_L\}$. Otherwise, increment $\ell \leftarrow \ell + 1$, and return to Step # 2.

Bibliography

[1] G. E. Box, N. R. Draper, *et al.*, *Empirical model-building and response surfaces*, vol. 424. Wiley New York, 1987.

[2] K. Roeder, "Density estimation with confidence sets exemplified by superclusters and voids in the galaxies," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 617–624, 1990.

[3] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[5] G. O. Roberts, R. L. Tweedie, *et al.*, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.

[6] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid monte carlo," *Physics letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[7] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through bayesian sampling," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 2, pp. 363–375, 1994.

[8] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[9] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[10] J. Q. Li and A. R. Barron, "Mixture density estimation," in *Advances in neural information processing systems*, pp. 279–285, 2000.

[11] A. L. Gibbs, "Bounding the convergence time of the gibbs sampler in bayesian image restoration," *Biometrika*, vol. 87, no. 4, pp. 749–766, 2000.

[12] J. S. Rosenthal, "Minorization conditions and convergence rates for markov chain monte carlo," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 558–566, 1995.

[13] N. Madras and D. Randall, "Markov chain decomposition for convergence rate analysis," *Annals of Applied Probability*, pp. 581–606, 2002.

[14] D. B. Woodard, J. S. Rosenthal, *et al.*, "Convergence rate of markov chain methods for genomic motif discovery," *The Annals of Statistics*, vol. 41, no. 1, pp. 91–124, 2013.

[15] D. B. Woodard, S. C. Schmidler, M. Huber, *et al.*, "Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions," *The Annals of Applied Probability*, vol. 19, no. 2, pp. 617–640, 2009.

[16] S. Frühwirth-Schnatter, *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[17] G. J. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2004.

[18] J.-M. Marin, K. Mengersen, and C. P. Robert, "Bayesian modelling and inference on mixtures of distributions," *Handbook of statistics*, vol. 25, pp. 459–507, 2005.

[19] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 957–970, 2000.

[20] M. Stephens, "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 4, pp. 795–809, 2000.

[21] S. Richardson and P. J. Green, "On bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society: series B (statistical methodology)*, vol. 59, no. 4, pp. 731–792, 1997.

[22] M. Stephens, "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods," *Annals of statistics*, pp. 40–74, 2000.

[23] W. Mou, N. Ho, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, "Sampling for bayesian mixture models: Mcmc with polynomial-time mixing," *arXiv preprint arXiv:1912.05153*, 2019.

[24] C. Tosh and S. Dasgupta, "Lower bounds for the gibbs sampler over mixtures of gaussians," in *International Conference on Machine Learning*, pp. 1467–1475, 2014.

[25] R. Ge, H. Lee, and A. Risteski, "Simulated tempering langevin monte carlo ii: An improved proof using soft markov chain decomposition," *arXiv preprint arXiv:1812.00793*, 2018.

[26] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, "Sharp convergence rates for langevin dynamics in the nonconvex setting," *arXiv preprint arXiv:1805.01648*, 2018.

[27] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, "Sampling can be faster than optimization," *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 20881–20885, 2019.

[28] J. S. Liu, "The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 958–966, 1994.

[29] D. A. Levin and Y. Peres, *Markov chains and mixing times*, vol. 107. American Mathematical Soc., 2017.

[30] A. Sinclair and M. Jerrum, "Approximate counting, uniform generation and rapidly mixing markov chains," *Information and Computation*, vol. 82, no. 1, pp. 93–133, 1989.

[31] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statist. Sci.*, vol. 7, pp. 457–472, 11 1992.

[32] J.-W. van de Meent, B. Paige, and F. Wood, "Tempering by subsampling," *arXiv preprint arXiv:1401.7145*, 2014.

[33] E. Marinari and G. Parisi, "Simulated tempering: a new monte carlo scheme," *EPL (Europhysics Letters)*, vol. 19, no. 6, p. 451, 1992.

[34] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[35] R. H. Swendsen and J.-S. Wang, "Replica monte carlo simulation of spin-glasses," *Physical review letters*, vol. 57, no. 21, p. 2607, 1986.

[36] R. M. Neal, "Sampling from multimodal distributions using tempered transitions," *Statistics and computing*, vol. 6, no. 4, pp. 353–366, 1996.

[37] C. J. Geyer and E. A. Thompson, "Annealing markov chain monte carlo with applications to ancestral inference," *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 909–920, 1995.

[38] N. Madras and Z. Zheng, "On the swapping algorithm," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 66–97, 2003.

[39] N. Bhatnagar and D. Randall, "Torpid mixing of simulated tempering on the potts model," Citeseer.

[40] D. Woodard, S. Schmidler, M. Huber, *et al.*, "Sufficient conditions for torpid mixing of parallel and simulated tempering," *Electronic Journal of Probability*, vol. 14, pp. 780–804, 2009.

[41] N. G. Tawn, G. O. Roberts, and J. S. Rosenthal, "Weight-preserving simulated tempering," *Statistics and Computing*, vol. 30, no. 1, pp. 27–41, 2020.

[42] A. Jasra, C. C. Holmes, and D. A. Stephens, "Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling," *Statistical Science*, pp. 50–67, 2005.

[43] M. Quiroz, M. Villani, R. Kohn, M.-N. Tran, and K.-D. Dang, "Subsampling mcmc-an introduction for the survey statistician," *Sankhya A*, vol. 80, no. 1, pp. 33–69, 2018.

[44] F. Obermeyer, J. Glidden, and E. Jonas, "Scaling nonparametric bayesian inference via subsample-annealing," in *Artificial Intelligence and Statistics*, pp. 696–705, 2014.

[45] G. Behrens, N. Friel, and M. Hurn, "Tuning tempered transitions," *Statistics and computing*, vol. 22, no. 1, pp. 65–78, 2012.

[46] P. Diaconis *et al.*, "Some things we've learned (about markov chain monte carlo)," *Bernoulli*, vol. 19, no. 4, pp. 1294–1305, 2013.

[47] B. D. He, C. M. De Sa, I. Mitliagkas, and C. Ré, "Scan order in gibbs sampling: Models in which it matters and bounds on how much," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, pp. 1–9, Curran Associates, Inc., 2016.

[48] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.