Tractable Bayesian teaching

Baxter S. Eaves Jr.*[1], April M. Schweinhart[1], and Patrick Shafto[1]

[1]Department of Mathematics and Computer Science, Rutgers University–Newark

October 27, 2015

Abstract

The goal of cognitive science is to understand human cognition in the real world, however, Bayesian theories of cognition are often unable to account for anything beyond schematic situations whose simplicity is typical only of experiments in psychology labs. For example, teaching to others is commonplace, but under recent Bayesian accounts of human social learning, teaching is, in all but the simplest of scenarios, intractable because teaching requires considering all choices of data and how each choice of data will affect learners' inferences about each possible hypothesis. In practice, teaching often involves computing quantities that are either combinatorially implausible or that have no closed-form solution. In this chapter we integrate recent advances in Markov chain Monte Carlo approximation with recent computational work in teaching to develop a framework for tractable Bayesian teaching of arbitrary probabilistic models. We demonstrate the framework on two complex scenarios inspired by perceptual category learning: phonetic category models and visual scenes categorization. In both cases, we find that the predicted teaching data exhibit surprising behavior. In order to convey the number of categories, the data for teaching phonetic category models exhibit hypo-articulation and increased within-category variance. And in order to represent the range of scene categories, the optimal examples for teaching visual scenes are distant from the category means. This work offers the potential to scale computational models of teaching to situations that begin to approximate the richness of people's experience.

Pedagogy is arguably humankind's greatest adaptation and perhaps the reason for our success as a species (Gergely, Egyed, & Király, 2007). Teachers produce data to efficiently convey specific information to learners and learners learn with this in mind (Shafto & Goodman, 2008; Shafto, Goodman, & Frank, 2012; Shafto, Goodman, & Griffiths, 2014). This not only ensures that information lives on after its discoverer, but also ensures that information is disseminated quickly and effectively. Shafto and Goodman (2008) introduced a Bayesian model of pedagogical data selection and learning, and used a simple teaching game to demonstrate that human teachers choose data consistently with the model and that human learners make stronger inferences from pedagogically-sampled data than from randomly-sampled data (data generated according to the true distribution). Subsequent work, using the same model, demonstrated that preschoolers learn differently from pedagogically-selected data (Bonawitz et al., 2011).

Under the model, a teacher, T, chooses data, x^*, to induce a specific belief (hypothesis, θ^*) in the learner, L. Mathematically, this means choosing data with probability in proportion with the induced posterior probability of the target hypothesis,

$$\begin{aligned} p_T(x^* \mid \theta^*) &= \frac{p_L(\theta^* \mid x^*)}{\int_x p_L(\theta^* \mid x) dx} && (1)\\ &= \frac{\frac{p_L(x^*|\theta^*)p_L(\theta^*)}{p_L(x^*)}}{\int_x \frac{p_L(x|\theta^*)p_L(\theta^*)}{p_L(x)} dx} && (2)\\ &\propto \frac{p_L(x^* \mid \theta^*)p_L(\theta^*)}{p_L(x^*)}. && (3)\end{aligned}$$

*email: baxter.eaves@rutgers.edu. This work was supported in part by NSF award DRL-1149116 to P.S.

Bayesian teaching includes Bayesian learning as a sub-problem because it requires considering all possible inferences given all possible data. At the outer layer (Equation 1) the teacher considers (integrates; marginalizes) over all possible alternative data choices, $\int_x p_L(\theta^* \mid x)dx$; at the inner layer (Equation 3), the learner considers all alternative hypotheses in the marginal likelihood, $p_L(x^*)$. The teacher considers how each possible dataset will affect learning of the target hypothesis, and the learner considers how well the data chosen by the teacher communicate each possible hypothesis. Pedagogy works because learners and teachers have an implicit understanding of each other's behavior. A learner can quickly dismiss many alternatives by reasoning that the teacher would have chose data differently had she meant to convey one of those alternatives. The teacher chooses data with this in mind.

Computationally, Bayesian teaching is a complex problem. Producing data that lead a learner to a specific inference about the world requires the teacher to make choices between different data. Choosing requires weighing one choice against all others, which requires computing large, often intractable sums or integrals (Luce, 1977). The complexity of the teacher's marginalization over alternative data can, to some extent, be mitigated by standard approximation methods (e.g. Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; S. Geman & Geman, 1984), but for teaching, this is not enough. For each choice of data, the teacher must consider how the learner will weigh the target hypothesis against alternative hypotheses. As we shall see, this inner marginalization is not one that we can easily make go away. And as the hypothesis becomes more complex, the marginalization becomes more complex; often, as is the case in categorization, the size of the set of alternative hypotheses increases ever faster as the number of data increases. For example, if a category learner does not know the number of categories, she must assume there can be as few as one category or as many categories as there are data. Learning complex concepts that are reminiscent of real-world scenarios often introduce marginalizations that have no closed form solution or that are combinatorially intractable. Because of this, existing work that models teaching has done so using necessarily simple, typically discrete, hypothesis spaces.

A Bayesian method of eliciting a specific inference in learners has applications beyond furthering our understanding of social learning, to education, perception, and machine learning; thus it is in our interest to make Bayesian teaching tractable. It is our goal in this chapter to leverage approximation methods that allow us to scale beyond the simple scenarios used in previous research.

We employ recent advances in Monte Carlo approximation to facilitate tractable Bayesian teaching. We proceed as follows: In section 1 we discuss some of the sources of complexity that arise in Bayesian statistics, such as marginalized probabilities, and discuss standard methods of combating complexity. In section 2 we briefly discuss new methods from the *Bayesian big data* literature, paying special attention to one particular method, *psuedo-marginal sampling*, which affords the same theoretical guarantees of standard Monte Carlo approximation methods while mitigating the effects of model complexity through further approximation, and outline the procedure for simulating teaching data. In section 3 we apply the teaching model to a debate within developmental psychology, whether infant-directed speech is for teaching, which amounts to teaching category models. Lastly, in section 4 we apply the teaching model to a more complex problem of teaching natural scene categories, which we model as categories of category models. We conclude with a brief recapitulation and meditation on future directions.

1 Complexity in Bayesian statistics

Complexity is the nemesis of the Bayesian modeler. It is a problem from the outset. Bayes' Theorem states that the *posterior probability*, $\pi(\theta'|x)$ of a hypothesis, θ', given some data, x is equal to the *likelihood*, $f(x|\theta')$, of the data given the hypothesis multiplied by the *prior probability*, $\pi(\theta')$, of the hypothesis divided by the *marginal likelihood*, $m(x)$, of the data:

$$\pi(\theta' \mid x) = \frac{f(x \mid \theta')\pi(\theta')}{m(x)} \tag{4}$$

$$= \frac{f(x \mid \theta')\pi(\theta')}{\sum_{\theta \in \Theta} f(x \mid \theta)\pi(\theta)}. \tag{5}$$

The immediate problem is this marginal likelihood (lurking menacingly below the likelihood and prior).

Often, the sum (or integral) over Θ involves computing a large, or infinite, number of terms or may have no closed-form solution, rendering it analytically intractable. Thus much of the focus of Bayesian statistical research involves approximating inference by either approximating certain intractable quantities or avoiding their calculation altogether.

1.1 Importance sampling

Importance sampling is a Monte Carlo method used to approximate integrals that are analytically intractable or not suitable for quadrature (numerical integration).[*] Importance sampling involves re-framing the integral of a function p with respect to θ as an expectation with respect to an *importance weight*, $w(\cdot) = p(\cdot)/q(\cdot)$, under q, such that $q(\cdot) > 0$ whenever $p(\cdot) > 0$. One draws a number, M, of independent samples $\bar{\theta}_1, \dots \bar{\theta}_M$ from q, and takes the arithmetic mean of $w(\bar{\theta_1}), \dots, w(\bar{\theta_M})$. By the Law of Large Numbers,

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} w(\bar{\theta}_i) = \int_\theta p(\theta) d\theta, \tag{6}$$

as $M \to \infty$, the average approaches the true value of the target expectation, which means that importance sampling produces an *unbiased estimate* (the expected value of the estimate is the true value).

If we wish to estimate $m(x)$, we set $w(\theta) = f(x \mid \theta)\pi(\theta)/q(\theta)$,

$$m(x) = \int_\theta f(x \mid \theta)\pi(\theta) d\theta = \int_\theta \left(\frac{f(x \mid \theta)\pi(\theta)}{q(\theta)} \right) q(\theta) d\theta = \mathbb{E}_q \left[\frac{f(x \mid \theta)\pi(\theta)}{q(\theta)} \right] \approx \frac{1}{M} \sum_{i=1}^{M} \frac{f(x \mid \bar{\theta}_i)\pi(\bar{\theta}_i)}{q(\bar{\theta}_i)}. \tag{7}$$

When we approximate the integral by using a sum, we no longer consider the differential, $d\theta$, but consider only individual realizations, $\bar{\theta}$, drawn from q. As well shall see in section 3, the choice of q influences the efficiency of the importance sampler. A straight-forward, though usually inefficient, choice is to draw $\bar{\theta}$ from the prior, $q(\theta) = \pi(\theta)$, in which case,

$$m(x) \approx \frac{1}{M} \sum_{i=1}^{M} f(x \mid \bar{\theta}_i). \tag{8}$$

We will use importance sampling to generate unbiased estimates of $p_L(x*)$ (the marginal probability for the learner; Equation 3). If the number of data is finite (and computationally tractable), the teaching probability, $p_T(x* \mid \theta^*)$, can be approximated directly by normalizing the approximations for each choice of data. If the number of data is infinite or intractable, we must adopt an approximate simulation scheme.

1.2 The Metropolis-Hastings algorithm

If we do not explicitly need the quantity $m(x)$, we can avoid calculating it altogether using the Metropolis-Hastings algorithm (MH; Metropolis et al., 1953; Hastings, 1970). MH is a Markov-chain Monte Carlo (MCMC) algorithm that is used to construct a Markov chain with $p(y)$ as its stationary distribution. This means that in the limit of state transitions, y will occur in the induced Markov chain with probability $p(y)$. MH requires a function g that is proportional to p, $g(y) = cp(y)$, and a proposal function $q(y \to y')$ that proposes moves to new states, y', from the current state, y. MH works by repeatedly proposing samples from q and accepting samples (setting $y := y'$) with probability min$[1, A]$, where

$$A := \frac{g(y')q(y' \to y)}{g(y)q(y \to y')}. \tag{9}$$

If q is symmetric, that is $q(y \to y') = q(y' \to y)$ for all y, y', then q cancels from the equation. For example, if $y \in \mathbb{R}$, then proposing y' from a normal distribution centered at y, $q(y \to y') := \mathcal{N}(y, \sigma)$, is a symmetric proposal density.[†]

[*] In general, quadrature is a more precise, computationally efficient solution than Monte Carlo integration in the situations in which it can be applied.

[†] Proposals that generate local moves result in *random walk* behavior.

To sample from the posterior distribution, set $g = f(x \mid \theta)\pi(\theta)$ and notice that $m(x)$ is a constant,

$$\begin{aligned} A &:= \frac{\pi(\theta' \mid x)}{\pi(\theta \mid x)} && (10) \\ &= \frac{f(x \mid \theta')\pi(\theta')m(x)}{f(x \mid \theta)\pi(\theta)m(x)} && (11) \\ &= \frac{f(x \mid \theta')\pi(\theta')}{f(x \mid \theta)\pi(\theta)} && (12) \end{aligned}$$

Thus, to draw posterior samples using MH, one need only evaluate the likelihood and prior.

The Metropolis-Hastings algorithm allows us to simulate data from the teaching model, regardless of the number of choices of data, if we can calculate the learner's posterior exactly. This allows us to generate data to teach in continuous data spaces. For example, we could generate pairs of continuous data to teach the mean of a univariate Gaussian distribution by performing a random walk on $\mathbb{R}^2$, accepting and rejecting moves along the random walk according to $A := p_L(\mu|x_1', x_2')/p_L(\mu|x_1, x_2)$. The difficulty is that in most cases we cannot calculate the learner's posterior exactly because $p_L(x*)$ is intractable.

2 Recent advances in Monte Carlo approximation

Thanks to algorithms like those mentioned in the previous section, the marginal likelihood is rarely a problem for standard Bayesian inference. It is so immaterial that modelers rarely acknowledge it, substituting '$\propto$' for '=' to avoid even writing it, for they shall not be bothering to calculate it anyway. These days, complex likelihoods pose a greater problem. For example, in Bayesian teaching, one is interested in the likelihood of the learner's inference given data, which is the learner's posterior. The complexity of the likelihood increases as the number of data increases and as the complexity of the learning model increases.

Large amounts of data directly affect computation time. Assuming that data are not reduced to a summary statistic, computation of the likelihood requires $\mathcal{O}(N)$ function evaluations, $f(x \mid \theta) = \prod_{i=1}^{N} \ell(x_i \mid \theta)$. If N is very large and ℓ is expensive to compute, then computing f is infeasible, which renders standard Monte Carlo inference infeasible. Methods exist for approximate (biased) MCMC using random subsets of the data, such as adaptive subsampling (Bardenet, Doucet, & Holmes, 2014) and stochastic gradient methods (Patterson & Teh, 2013). Firefly Monte Carlo (Maclaurin & Adams, 2014), which uses a clever proposal density to activate (light up) certain data points, is the first exact MCMC algorithm to use subsets of data. Other proposals employ multiprocessing strategies such as averaging results from independent Monte Carlo simulations run on subsets of data (Scott et al., 2013) and dividing computations and computing parts of the Metropolis-Hastings acceptance ratio on multiple processors (Banterle, Grazian, & Robert, 2014).

2.1 Pseudo-marginal Markov chain Monte Carlo

Teaching is more complicated than learning because it contains learning a sub-problem. The learner's posterior can be thought of as the likelihood of the teaching data. Thus, the Bayesian Teacher faces a problem of intractable likelihoods, which is exacerbated by the number of data used to teach. Increasing the number of data used to teach increases the complexity of the learner's posterior and increases the dimensionality of the teaching model (as opposed to the model being taught). Because in most cases the teaching likelihood (the learner's posterior) cannot be calculated exactly we would like an algorithm that allows us to sample teaching data using approximations; and because in most cases the number of choices of possible data does not permit enumeration, we would like an algorithm that allows us to simulate data using unnormalized approximations. We would like an algorithm that combines importance sampling approximation with Metropolis-Hasting (MH) simulation.

We focus on a technique referred to as *pseudo-marginal MCMC* (PM-MCMC; Andrieu & Roberts, 2009; Andrieu & Vihola, 2012), which allows exact MH to be performed using approximated functions. Assume that g in Equation 9 is difficult to evaluate but that we can compute an estimate $\hat{g}(y) = Wg(y)$, where $W = \hat{g}(y)/g(y)$ is a non-deterministic error term (weight) associated with, but not dependent on, y such that W is a positive random variable, $W \sim \psi$,[*] with expected value equal to some constant, $\mathbb{E}[W] = k$. The target distribution is then a joint distribution over y and W; each time we propose y, we implicitly propose a new weight. The approximate acceptance ratio is then

$$A := \frac{w' \, \psi(w') g(y') q(y' \to y)}{w \, \psi(w) g(y) q(y \to y')}. \tag{13}$$

Because $\int p(y, W) dW \propto \mathbb{E}[W] g(y) = k \, g(y)$, the Markov Chain induced by Equation 13 has a target distribution proportional to g. Thus we achieve exact MH simply by substituting g with $\hat{g}$ in the original acceptance ratio,

$$A := \frac{\hat{g}(y') q(y' \to y)}{\hat{g}(y) q(y \to y')}. \tag{14}$$

And to simulate from the posterior of a density with an intractable likelihood:

$$A := \frac{\hat{f}(x \mid \theta') \pi(\theta') q(\theta' \to \theta)}{\hat{f}(x \mid \theta) \pi(\theta) q(\theta \to \theta')}, \tag{15}$$

where $\hat{f}(x \mid \theta)$ is a Monte Carlo estimate of the likelihood, $f(x \mid \theta)$. The stability and convergence properties of PM-MCMC have been rigorously characterized (Andrieu & Vihola, 2012; Sherlock, Thiery, Roberts, & Rosenthal, 2013). In practice, the user need only ensure that $\hat{g}(y)$ has a constant bias—for example, importance sampling is *unbiased*, which implies that $\mathbb{E}[W] = 1$—and that each $\hat{g}(y)$ is never recomputed for any y.

2.2 Teaching using PM-MCMC

The purpose of teaching, from the teacher's perspective, is to choose one specific dataset from the collection of all possible datasets, to convey one specific hypothesis to a learner who considers all hypotheses,

$$p_T(x^* \mid \theta^*) = \frac{p_L(\theta^* \mid x^*)}{m(\theta^*)} \propto \frac{p_L(x^* \mid \theta^*)}{p_L(x^*)}.$$

The teacher marginalizes over datasets, $m(\theta^*) = \int_x p_L(\theta^* \mid x) dx$, and for each dataset marginalizes over all possible inferences, $p_L(x^*) = \int_\theta p_L(x^* \mid \theta) p(\theta) d\theta$. To generate teaching data, we must simulate data according to this probability distribution while navigating nested marginalizations.

We use PM-MCMC to simulate teaching data by embedding importance sampling within the Metropolis-Hastings algorithm. We use MH to avoid calculating the integral over alternative data, $\int_x p_L(\theta^* \mid x) dx$, leaving the acceptance ratio,

$$A = \frac{p_L(x' \mid \theta) p_L(x^*)}{p_L(x^* \mid \theta) p_L(x')}, \tag{16}$$

where x' is the proposed (continuous) data and it is assumed that the proposal density, q, is a symmetric, Gaussian perturbation of the data (random walk). Equation 16 indicates that we must calculate the marginal likelihoods of data in order to use MH for teaching. This marginalization is inescapable, so we replace it with an importance sampling approximation, $\hat{p}_L(x)$.

Teaching necessarily depends on the content to be taught, and different problems require different formalizations of learning. In the following two sections we employ the teaching model to generate data to teach in two distinct perceptual learning problems involving categorization: phonetics and visual scenes. Categorization is a well-studied psychologically (see Anderson, 1991; J. Feldman, 1997; Markman & Ross, 2003) and computationally (see Jain, Murty, & Flynn, 1999; Radford M Neal, 2000; Rasmussen, 2000), and presents a particularly challenging marginalization problem, and is thus an ideal testbed.

[*] The notation $X \sim F$ denotes that the random variable X is distributed according to F

3 Example: Infant-directed speech (infinite mixture models)

Infant-directed speech (IDS; motherese) has distinct properties such as a slower speed, higher pitch, and singsong prosody. Kuhl et al. (1997) discovered that IDS has unique phonetic properties that might indicate that IDS is for teaching.

Phonemes are defined by their formants, which are peaks in the spectral envelope. The first formant, F_1, is the lowest frequency peak; the second formant, F_2, is the second lowest frequency peak; and so on. The first two formants are usually sufficient to distinguish phonemes. When examples of phonemes are plotted in $F_1 \times F_2$ formant space they form bivariate Gaussian clusters. Kuhl et al. (1997) observed that the clusters of IDS corner vowel examples, (/ɑ/, as in pot; /i/, as in beet; /u/, as in boot) are hyper-articulated (farther apart), resulting in an increased vowel space. This led to the proposition that IDS is for teaching because clusters that are farther apart should be easier to discriminate.

The IDS research currently lacks a formal account of teaching vowel phonemes to infants; rather arguments are built around intuitions, which is conceivably the source of much of the contention regarding this topic.* This sort of question has been unapproachable because languages contain many phonemes, and the set of possible categorizations of even a small number of examples rapidly results in an intractable number of terms in the marginalization likelihood. Here we show how the teaching model can be applied to such a problem. We first describe a model of learning Gaussian category models and then describe the teaching framework. We then generate teaching data and explore their qualitatively properties.

3.1 Learning phonetic category models

Infants must learn how many phonemes there are in their native language as well as the shapes, sizes, and location of each phoneme in formant space, all while inferring to which phoneme each datum belongs. For this task, we employ a Bayesian nonparametric learning framework for learning category models with an unknown number of categories (Rasmussen, 2000).†

Following other work, we formalize phonetic category models as mixtures of Gaussians (Vallabha, McClelland, Pons, Werker, & Amano, 2007; N. H. Feldman, Griffiths, Goldwater, & Morgan, 2013). Each phoneme, $\phi_1, \dots, \phi_K \in \Phi$, is a Gaussian with mean μ_k and covariance matrix, Σ_k,

$$\Phi = \{\phi_1, \dots, \phi_K\} = \{\{\mu_1, \Sigma_1\}, \dots, \{\mu_K, \Sigma_K\}\}. \tag{17}$$

The likelihood of some data under a finite Gaussian mixture model is

$$f(x|\Phi, \Omega) = \prod_{i=1}^{N} \sum_{k=1}^{K} \omega_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k), \tag{18}$$

where $\mathcal{N}(x \mid \mu, \Sigma)$ is the multivariate Normal likelihood of x given μ and Σ, and each ω_k in Ω is a nonnegative real number such that $\sum_{k=1}^{K} \omega_k = 1$. The above model assumes that the learner knows the number of categories. We are interested in the case where in addition to the means and covariance matrices of each phoneme, the learner must infer the assignment of examples to an unknown number of phonemes. The assignment is represented as an N-length vector $z = [z_1, \dots, z_N]$. Each entry $z_i \in 1, \dots, K$. In this case the likelihood is,

$$f(x|\Phi, z) = \prod_{i=1}^{N} \sum_{k=1}^{K} \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \delta_{z_i,k}, \tag{19}$$

where $\delta_{i,j}$ is the Kronecker delta function which assumes value 1 if $i = j$ and value 0 otherwise. $\delta_{z_i,k}$ equals 1 if, and only if the i^{th} datum is a member of phoneme k.

[*] We refer those interested in reading more about this debate to Burnham, Kitamura, and Vollmer-Conna (2002), de Boer and Kuhl (2003), Uther, Knoll, and Burnham (2007), McMurray, Kovack-Lesh, Goodwin, and McEchron (2013), and Cristia and Seidl (2013).

[†] The term *nonparametric* is used to indicate that the number of parameters is unknown (that we must infer the number of parameters), not that there are no parameters.

We employ the Dirichlet process Gaussian mixture model (DPGMM) framework. Learners must infer Φ and z. We assume the following generative model:

$$\begin{aligned} G_k &\sim \mathrm{DP}(\alpha H), && (20)\\ \phi_k &\sim G_k, && (21)\\ x_k &\sim \mathcal{N}(x_k \mid \phi_k). && (22) \end{aligned}$$

where $\mathrm{DP}(\alpha H)$ is a *Dirichlet process* with concentration parameter α that emits probability distributions (components; categories) of type H, where H is the prior distribution on $\boldsymbol{\phi}$. Here H is the Normal Inverse-Wishart (NIW) prior (Murphy, 2007),

$$\mu_k, \Sigma_k \sim \mathrm{NIW}(\mu_0, \Lambda_0, \kappa_0, \nu_0) \tag{23}$$

which implies

$$\begin{aligned} \Sigma_k &\sim \text{Inverse-Wishart}_{\nu_0}(\Lambda_0^{-1}), && (24)\\ \mu_k|\Sigma_k &\sim \mathcal{N}(\mu_0, \Sigma_k/\kappa_0), && (25) \end{aligned}$$

where Λ_0 is the prior scale matrix, μ_0 is the prior mean, $\nu_0 \geq d$ is the prior degrees of freedom, and κ_0 is the number of prior observations.

To formalize inference over z, we introduce a prior, $\pi(z \mid \alpha)$, via the Chinese Restaurant Process (Teh, Jordan, Beal, & Blei, 2006), denoted $\mathrm{CRP}(\alpha)$, where the parameter α affects the probability of new components. Higher α creates a higher bias toward new components. Data points are assigned to components as follows:

$$P(z_i = k|z^{(-i)}, \alpha) = \begin{cases} \frac{n_k}{N-1+\alpha} & \text{if } k \in 1 \ldots K \\ \frac{\alpha}{N-1+\alpha} & \text{if } k = K+1 \end{cases}, \quad n_k = \sum_{i=1}^{N} \delta_{z_i,k}, \tag{26}$$

where $z^{(-i)} = z \setminus z_i$.

3.2 Teaching DPGMMs

Recall that the probability of the teacher choosing data is proportional to the induced posterior. The posterior for the DPGMM is,

$$\pi(\Phi, z|x) = \frac{f(x \mid \Phi, z)\pi(\Phi \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)\pi(z \mid \alpha)}{m(x \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}. \tag{27}$$

Our choice of prior allows us to calculate the marginal likelihood exactly,

$$\begin{aligned} m(x \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) &= \sum_{z \in \mathfrak{Z}} \pi(z \mid \alpha) \prod_{k=1}^{K_z} \iint_{\phi_k} f(x_k \mid \phi_k, z)\pi(\phi_k \mid \mu_0, \Lambda_0, \kappa_0, \nu_0) d\phi_k && (28)\\ &= \sum_{z \in \mathfrak{Z}} \pi(z \mid \alpha) \prod_{k=1}^{K_z} f(x_k \mid \mu_0, \Lambda_0, \kappa_0, \nu_0), && (29) \end{aligned}$$

where

$$\pi(z \mid \alpha) = \frac{\Gamma(\alpha) \prod_{k=1}^{K_z} \Gamma(n_k)}{\Gamma(N + \alpha)} \alpha^{K_z}, \tag{30}$$

$\mathfrak{Z}$ is the set of all possible partitions of N data points into 1 to N categories, K_z is the number of categories in assignment vector z, and $f(x_k \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)$ is the marginal likelihood of the data assigned to category k

under NIW (which can be calculated analytically). The size of $\mathfrak{Z}$ has its own named combinatorial quantity: the *Bell number*, or B_n. If we have sufficiently little data or ample patience, we can calculate the quantity in Equation 29 by enumerating $\mathfrak{Z}$. However, Bell numbers grow quickly, $B_1 = 1$, $B_5 = 25$, $B_{12} = 4213597$, and so on. We can produce an importance sampling approximation by setting $q(z) := \pi(z|\alpha)$,

$$\hat{m}(x \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha) = \frac{1}{M} \sum_{i=1}^{M} \prod_{k=1}^{K_{z_i}} f(x_k \mid \mu_0, \Lambda_0, \kappa_0, \nu_0). \tag{31}$$

The approach of drawing from the prior by setting $q(\theta) := \pi(\theta)$ is usually inefficient. Areas of high posterior density contribute most to the marginal likelihood, thus the optimal q is close to the posterior. Several approaches have been proposed for estimating the marginal likelihood in finite mixture models (Chib, 1995; Marin & Robert, 2008; Rufo, Martín, & Pérez, 2010; Fiorentini, Planas, & Rossi, 2012), here we use a Gibbs initialization importance sampling scheme suited to the infinite case using sequential importance sampling (Maceachern, Clyde, & Liu, 1999).* Each sample, $\bar{z}_1, \dots, \bar{z}_M$, is drawn by sequentially assigning the data to categories based on the standard collapsed Gibbs sampling scheme (Algorithm 1),

Algorithm 1 Partial Gibbs importance sampling proposal

```
function PGibbs($x$, $\mu_0$, $\Lambda_0$, $\kappa_0$, $\nu_0$, $\alpha$)
    $q \leftarrow 0$
    $Z \leftarrow [1]$
    $K \leftarrow 1$
    $n \leftarrow [1]$
    for $i \in 2, \dots, |x|$ do
        $P \leftarrow$ empty array of length $K + 1$
        for $k \in 1, \dots, K$ do
            $y \leftarrow \{x_j \in x_1, \dots, x_{i-1} : Z_j = k\}$
            $P[k] \leftarrow n[k] \times f(x_i \mid y, \mu_0, \Lambda_0, \kappa_0, \nu_0)$
        end for
        $P[K+1] \leftarrow \alpha \times f(x_i \mid \mu_0, \Lambda_0, \kappa_0, \nu_0)$
        $P \leftarrow P / \sum_{p \in P} p$
        $z \sim$ Discrete($P$)
        $Z$.append($z$)
        $q \leftarrow q + P[z]$
        if $z \leq K$ then
            $n[z] \leftarrow n[z] + 1$
        else
            $n$.append(1)
            $K \leftarrow K + 1$
        end if
    end for
    return $q$
end function
```

$$q(z) = \prod_{i=2}^{N} p(z_i | \{z_1, \dots, z_{i-1}\}, \{x_1, \dots, x_{i-1}\}, \Lambda_0, \mu_0, \kappa_0, \nu_0, \alpha), \tag{32}$$

$$p(z_i | \{z_1, \dots, z_{i-1}\}, \{x_1, \dots, x_{i-1}\}, \Lambda_0, \mu_0, \kappa_0, \nu_0, \alpha) \propto \begin{cases} n_k \, f(x_i \mid x_k, \mu_0, \Lambda_0, \kappa_0, \nu_0) & \text{if } k \in 1 \dots K \\ \alpha \, f(x_i \mid \mu_0, \Lambda_0, \kappa_0, \nu_0) & \text{if } k = K + 1 \end{cases}. \tag{33}$$

[*] For an overview of methods for controlling Monte Carlo variance, see Robert and Casella (2013, Chapter 4).

Because each sample is independent, we can ignore the label-switching problem, which produces unpredictable estimator biases (see Chib, 1995 and Radford M. Neal, 1999). From here, we simulate teaching data for the target model, (Φ, z), according to, pseudo-marginal MH acceptance ratio,

$$\hat{A} = \frac{f(x' \mid \Phi, z)\hat{m}(x \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}{f(x \mid \Phi, z)\hat{m}(x' \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)}. \tag{34}$$

To ensure that the importance sampling estimate of the marginal likelihood (Equation 31) converges to the exact quantity, we generated 2500 random datasets for $N = 6, 8, 10$ from $\mathcal{N}([0,0], I_2)$ and calculated the importance sampling estimate for up to 10000 samples. For the calculation, the NIW parameters were set to $\mu_0 = [0,0]$, $\lambda_0 = I_2$, $\kappa_0 = 1$, and $\nu_0 = 2$; and the CRP parameter was $\alpha = 1$. Figure 1A shows the average relative error as a function of the number of samples. The results demonstrate that the relative error of the IS estimate decreases as the number of samples increases, that there is generally more error for higher N, and that the Gibbs importance sampling scheme produces a third of the error of the prior importance sampling scheme. We compared the runtime performance of C++ implementations of the exact calculation via enumeration and importance sampling (using $M = 1000$ samples) for $n = 1, \ldots, 13$. The results can be seen in Figure 1B and C. Enumeration is faster than IS until $N = 10$ after which the infeasibility of enumeration becomes apparent. For $N = 13$ importance sampling with $M = 1000$ is 400 times faster than enumeration.

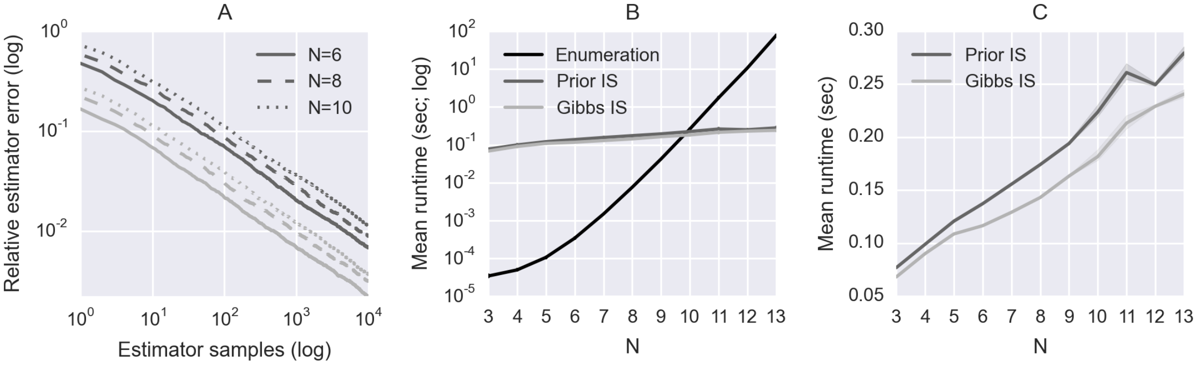

Figure 1: Performance comparison between prior and Gibbs importance sampling. A) Mean relative error, over 2500 random datasets, (y-axis) of the prior importance sampling approximation (dark) and the Gibbs importance sampling approximation (light; Equation 32) by number of samples (x-axis) for 6 (solid), 8 (dashed), and 10 datapoints (dotted). B) Runtime performance (seconds; y-axis) of algorithms for calculating/approximating $m(x \mid \mu_0, \Lambda_0, \kappa_0, \nu_0, \alpha)$ by number of data points (N; x-axis): exact calculation via enumeration (black), 1000 samples of prior importance sampling (dark gray), and 1000 samples of Gibbs importance sampling (light gray). C) Separate view of runtime of the importance samplers.

3.3 Experiments

We first conducted small-scale experiments to demonstrate that $\hat{x}$ simulated using $\hat{A}$ (the pseudo-margienal acceptance ratio) is equivalent to x simulated using A (the exact acceptance ratio) while demonstrating the basic behavior of the model. We then scaled and conducted experiments to determine what type of behavior (e.g. hyper- or hypo-articulation, and variance increase) can be expected in data designed to teach complex Gaussian category models to naive learners.

To ensure the exact MH samples and pseudo-marginal MH samples are identically distributed we used a three-category model taught with two data points assigned to each category. We collected 1000 samples across 5 independent Markov chains, ignoring the first 200 samples from each chain and thereafter collecting every 20th sample.* The prior parameters were set as in the previous section. Figure 2A and B show the result. Both data sets exhibited similar behavior including hyper-articulation, denoted by the increased distance between the category means of the teaching data, and within-category variance increase. A two-sample, permutation-based, Gaussian Kernel test (Gretton, Fukumizu, Harchaoui, & Sriperumbudur, 2009, 2012) using 10000 permutations indicated that the exact and pseudo-marginal data are identically distributed ($p = 0.9990$).

From a pedagogical standpoint, hyper-articulation is intuitively sensible. A learner cannot accurately learn the shapes and locations of categories if she has not learned how many categories there are. Learning the number of categories is made considerably easier by accentuating the differences between phonemes. Thus, much of the effort of teaching category models should be allocated toward teaching the number of categories, perhaps at the expense of accurately conveying their other attributes.

To further demonstrate this behavior in the teaching model, we designed a two-category model where both categories had identical means ($[0, 0]$), but had opposite x and y variances $(3.5, 1)$ and $(1, 3.5)$. Each category was taught with two data points. We repeated the same sampling procedure used in the previous experiment and used the same NIW parameters. We see in Figure 2C that the samples from the target model (gray) appear to form a single cluster but that the teaching model elongated the categories perpendicularly to form a cross, which makes the number of categories clear.

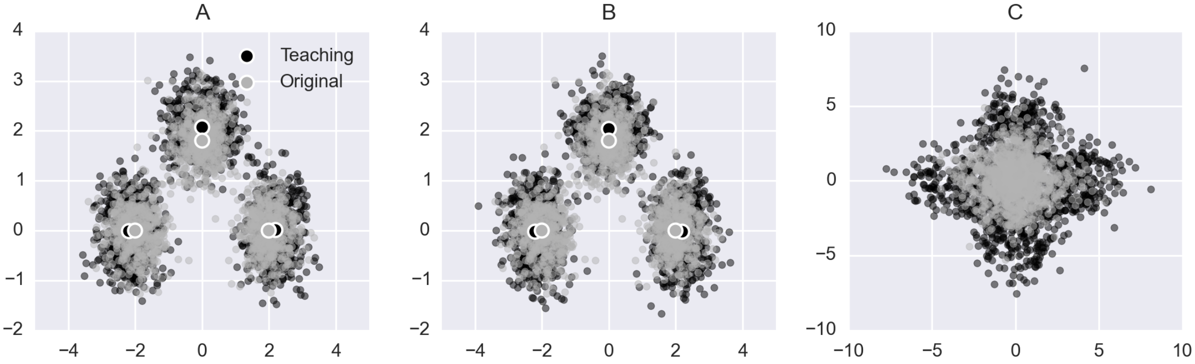

Figure 2: Behavior of the teaching model with exact and pseudo-marginal samples. A) Three-category Gaussian mixture model. The gray points are drawn directly from the target model and the black points are drawn from the teaching model using the exact acceptance ratio with $N = 3$. B) Three-category Gaussian mixture model. The gray points are drawn directly from the target model and the black points are drawn from the teaching model using the pseudo-marginal acceptance ratio with $N = 3$. C) Pseudo-marginal samples for a two category model where both categories have the same mean.

We increased the complexity of the target model to better represent the complexity of speech and to determine whether more complex target models necessitate more nuanced teaching manipulations. Dialects of the English language have around 20 vowel phonemes. Hypo-articulation may result when two near clusters move apart; if one pair of clusters moves apart, the resulting movement may force other pairs of clusters closer together.

To ensure the simulation results were interpretable, we did not increase the number of dimensions of the phonetic categories. The complexity of the teaching model is mainly a function of the number of data; recall that computing the posterior of a DPGMM grows with the number of data according to the Bell number. If the number of dimensions is increased, the number of terms in the marginal likelihood remains the same. The dimensionality of the data affects the time to complete matrix operations and the convergence time of PM-MCMC.

We randomly generated a twenty-category target model for which we simulated teaching data (one datum per category). After discarding the first 1000 samples, we collected every 40th sample until 500 samples had

[*] When the target distribution is multi-modal, Markov chain samplers often become stuck in a single mode. To mitigate this, it is common practice to sample from multiple independent Markov chains.

been collected from each of 8 independent chains of PM-MCMC. We aggregated the samples across chains and calculated the category means of the teaching data using the aggregated data. The teaching samples and their category means are plotted along with random samples from the target model in Figure 3 (top). The change in distance (hypo- and hyper-articulation) between all 120 pairs can be seen in Figure 3 (bottom). The resulting teaching data exhibited a general tendency to hyper-articulate, though a number of category pairs seemed to hypo-articulate. This hypo-articulation did not occur only in pairs that were well-separated in the target model. Among the hypo-articulated category pairs were adjacent pairs 0-8, 1-6, 4-14, and 11-14. Qualitatively, it appeared the teaching model used hypo-articulation in conjunction with variance increases to disambiguate cluster boundaries. For example, clusters 0, 6, 8, and 12 were close together in the original data; in the teaching data, clusters 6 and 8 moved away from each other but moved closer to clusters 1 and 0. This movement created a clearer separation between clusters 1 and 6, and clusters 0, 8, and 12. Variance increases then helped to disambiguate the hypo-articulated clusters as in Figure 2. These results demonstrate that hypo-articulation is consistent with teaching.

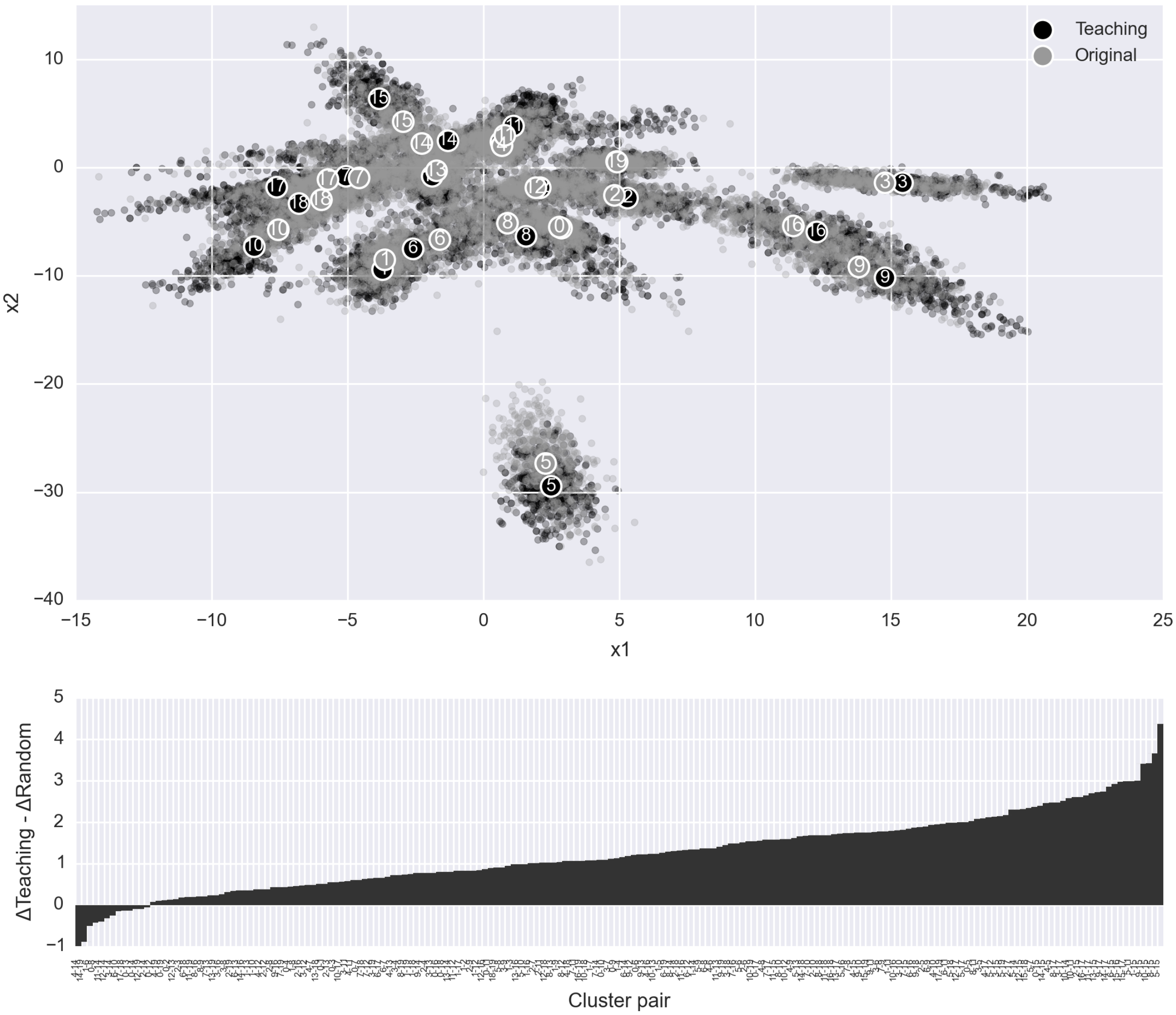

Figure 3: Twenty-phoneme experiment. Top) Scatter plot of random samples from the target model (gray) and the teaching data (black). The numbered circles represent the means of each of the 20 categories. Bottom) Change in distance between category pairs from random to teaching samples. Negative values indicate hypo-articulation and positive values indicate hyper-articulation.

3.4 Discussion

Inspired by a debate from the infant-directed speech literature, we sought to teach Gaussian category models to naive learners using a non-parametric categorization framework. We demonstrated how standard MH sampling in the teaching model becomes intractable at a small number of datapoints/categories (Figure 1B) and showed how PM-MCMC allows for tractable teaching in complex models. We then conducted experiments verifying that PM-MCMC produces results indistinguishable from standard MH, while demonstrating that, like IDS, the teaching model produces hyper-articulation and within-category variance increase (Figure 2). We then scaled up and created a random target model with roughly the same complexity as an English phonetic category model finding that hypo-articulation, hyper-articulation, and variance increase are all features consistent with teaching.

The results suggest that these features are consistent with teaching in general, but do not indicate that they are consistent specifically with teaching phonetic category models. One may speculate that the movement trends in the teaching data are mainly a result of the format space being reduced to two dimensions. Doubtless, the number of dimensions does affect teaching manipulations because how a model is taught depends on the model that is being taught. Phonetic categories are an interesting in this respect because two formants are usually sufficient to identify vowels excepting *rhotic* or *r-colored* vowels (e.g., bur, bar, and bore), which have lower third formants. Teaching manipulations in the third formant may be more severe for rhotic vowels than other vowels because the third formant is important to identifying rhotic vowels but not other vowels. Determining whether results similar to those above hold for real phonetic category models and models with higher dimension is a direction for future research. We have demonstrated that, using PM-MCMC, the teaching model is capable of contributing to this and other theoretical debates in teaching complex categories models such as those of natural language.

4 Example: Natural scene categories (infinite mixtures of infinite mixtures)

The visual environment has regular properties including a predictable anisotropic distribution of oriented contours. While these properties are not explicitly taught, there is evidence to indicate that the visual system learns and takes advantage of these properties through experience in the visual world (Hansen & Essock, 2004; Schweinhart & Essock, 2013; Wainwright, 1999), and the ability to automatically teach people's perception would be useful.

The distribution of orientations in the visual environment is bimodal, peaking at the cardinal orientations (horizontal and vertical: Coppola, Purves, McCoy, & Purves, 1998; Hansen & Essock, 2004; Switkes, Mayer, & Sloan, 1978) In carpentered (man-made) environments, this makes sense as buildings and walls tend to have both horizontal and vertical contours. However, even in the natural environment (i.e. an outdoor rural scene), there tends to be more structural information at the cardinal orientations due to the horizon, foreshortening, and phototropic/gravitropic growth. The average scene contains most of its structural content around horizontal, second most around vertical and least near the 45-degree obliques. The human visual system's processing of oriented structure is biased in the opposite way, thus neutralizing this anisotropy in natural scenes by suppressing the perceptual magnitude of content at orientations near horizontal most, least at oblique orientations, and intermediate suppression at vertical orientations—termed the horizontal effect (Essock, DeFord, Hansen, & Sinai, 2003, 2009).

4.1 Sensory learning of orientation distributions

While the general pattern of anisotropy present in natural scenes has been found to be a good match to perceptual biases (Essock et al., 2009), there are substantial differences between the distributions for carpentered and non-carpentered environments (Girshick, Landy, & Simoncelli, 2011). The distribution of oriented contours in an office environment has substantially greater peaks at the cardinal orientations than the distribution in a national park, for instance. Here we generalize the teaching model described in the previous section to determine optimal examples for 'teaching' the visual system the distribution of natural perceptual scenes from different categories (e.g. nature versus "carpentered" environments) based solely on their oriented structure.

Given data in the form of the amplitudes of various, discrete orientations, scene categories can themselves be multi-modal. For example, the oriented content in forest scenes is different from the oriented content in desert scenes, but both desert and forest scenes fall into the category of natural scenes. In order to begin quantifying different types of scene categories, we employ a nested categorization model in which outer categories are composed of inner categories (For a similar but more restrictive model see Yerebakan, Rajwa, & Dundar, 2014). More specifically, we implement a Dirichlet process mixture model where the outer Dirichlet process emits a Dirichlet process that emits Gaussians according to NIW. This is a generalization of the DPGMM model outlined in the previous section.

The generative process of this Dirichlet Process mixture model of Dirichlet process Gaussian mixture models (DP-DPGMM) is outlined in Algorithm 2. A CRP parameter for the outer categories, γ, is drawn from H; and the assignment of data to outer categories, z, is drawn from $CRP_N(\gamma)$. For each outer category, $k = 1, \ldots, K$, an inner CRP parameter, α_k, is drawn from Λ; a set of NIW parameters, G_k, is drawn from G; and an assignment of data in outer category k to inner categories, v_k, is drawn from $CRP_{n_k}(\alpha_k)$. For each inner category, $j = 1, \ldots, J_k$, a mean and covariance, μ_{kj} and Σ_{kj}, are drawn from G_k; and data points are drawn from those μ_{kj} and Σ_{kj}. The full joint density is,

$$p(\gamma \mid H)p(z \mid \gamma) \prod_{k=1}^{K} \left(p(\alpha_k \mid \Lambda)p(v_k \mid \alpha_k)p(G_k \mid G) \prod_{j=1}^{J_k} \left(p(\mu_{kj}, \Sigma_{kj} \mid G_k) \prod_{x \in x_{kj}} p(x \mid \mu_{kj}, \Sigma_{kj}) \right) \right). \tag{35}$$

Algorithm 2 Generative process of the DP-DPGMM

```
procedure DP-DPGMM($G$, $H$, $\Lambda$, the number of data $N$)
    $\gamma \sim H$
    $z \sim CRP_N(\gamma)$
    for $k \in 1, \ldots, K_z$ do
        $\alpha_k \sim \Lambda$
        $G_k \sim G$
        $v_k \sim CRP_{n_k}(\alpha_k)$
        for $j \in 1, \ldots, J_k$ do
            $\mu_{kj}, \Sigma_{kj} \sim G_k$
        end for
        for $i \in 1, \ldots, n_k$ do
            $x_{ki} \sim \mathcal{N}(\mu_{k,v_i}, \Sigma_{k,v_i})$
        end for
    end for
end procedure
```

4.2 Teaching DP-DPGMMs

Given data $x = x_1, \ldots, x_N$ we wish to teach the assignment of data to outer categories, z, the assignment of data to inner categories, v, and the means and covariance matrices that comprise the inner categories. The DP-DPGMM framework assumes that G, H, and Λ (the base distributions on G_k, and the outer and inner CRP parameters) are known and that all other quantities are unknown. To compute the marginal likelihood $m(x \mid G, H, \Lambda)$, we must integrate and sum over all unknowns. The resulting quantity is far more complex than the DPGMM marginal likelihood (Equation 29). We approximate $m(x \mid G, H, \Lambda)$ via importance sampling by drawing parameters from the generative process and calculating the likelihood of the data,

$$m(x \mid G, H, \Lambda) \approx \frac{1}{M} \sum_{i=1}^{M} \prod_{k=1}^{K_{\bar{z}_i}} \prod_{j=1}^{J_{\bar{v}_{ki}}} f(x_{k,j} \mid \bar{G}_k), \tag{36}$$

where $K_{\bar{z}_i}$ is the number of outer categories in the ith outer category assignment, $\bar{z}_i$, and $J_{\bar{v}_{ki}}$ is the number of inner categories in the kth outer category according to the ith inner category assignment, $\bar{v}_i$. The MH acceptance ratio is then,

$$A = \frac{\hat{m}(x \mid G, H, \Lambda) \prod_{k=1}^{K_{z^*}} \prod_{j=1}^{J_{v^*_k}} \mathcal{N}(x'_{kj} \mid \mu^*_{kj}, \Sigma^*_{kj})}{\hat{m}(x' \mid G, H, \Lambda) \prod_{k=1}^{K_{z^*}} \prod_{j=1}^{J_{v^*_k}} \mathcal{N}(x_{kj} \mid \mu^*_{kj}, \Sigma^*_{kj})}. \tag{37}$$

Notice that all factors of the full joint distribution that do not rely on the data cancel from A, leaving only the likelihood of the data under the inner-category parameters $(\mu^*_{kj}, \Sigma^*_{kj})$ and the marginal likelihood.

4.3 Experiments

We used the teaching model to choose images from a set of existing images that are best for teaching scene categories given specific reference points from their orientation distribution.

Different types of visual experience were collected by wearing a head mounted camera (NET CMOS iCube USB 3.0; 54.9o X 37.0o FOV) which sent an outgoing video feed to a laptop that was stored in a backpack. A sample of images was taken from such videos that were recorded during typical human environmental interaction as observers walked around different types of environments for variable amounts of time (a nature preserve, inside a house, downtown in a city, around a University, etc). The observers were generally taking part in a separate experiment that also involved wearing a head mounted camera (Schweinhart, Shafto, & Essock, submitted). Subsequently, every 1000th frame of the videos was taken as a representative sample of a given video and sample images were sorted into two outer categories by an expert observer: purely natural (no man-made structure) or outdoor, but containing carpentered content. A sub-sample of 200 images from each outer category was used to generate the model. The structural information was extracted using a previously developed image rotation method (see Schweinhart & Essock, 2013). Briefly, each frame was fast Fourier transformed, rotated to the orientation of interest and the amplitude of the cardinal orientations (horizontal and vertical) was extracted and stored. Repeating this process every 15 degrees allowed each video frame to be condensed into a series of 12 data points representing the amount of oriented structure in the image (see Figure 4 for an illustration). In this work, we teach natural and carpentered scenes using using four-dimensional data using the amplitudes at 0, 45, 90, and 135 degrees.

To derive a target distribution (means and covariance matrices of inner categories), we applied expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) to the orientation data from each setting.[*] To facilitate cross-referencing existing images, rather than generating a distribution over datasets, we searched for the single best dataset, x_{opt}, for teaching the scene categories by searching for the dataset that maximized the quantity in Equation 3, i.e.,

$$x_{\text{opt}} = \text{argmax}_x \left[p_T(x \mid \theta^*) \right]. \tag{38}$$

When modeling aggregate human behavior, such as infant-directed speech, modelers make the assumption that humans, as a group, choose data in a way that can be modeled with probabilities. Thus choosing the single best dataset for teaching is not an appropriate strategy for modeling the teaching behavior of groups, but is appropriate for doing teaching. Finding the best teaching example is akin to modeling the optimal teacher.

We found the approximate argmax via *simulated annealing* (Metropolis et al., 1953). Simulated annealing applies a temperature, T, to the Metropolis-Hastings (MH) acceptance ratio, $A^{1/T}$. A higher temperature has the effect of softening the distribution, which prevents MH from becoming stuck in a local maxima, allowing it to more easily find the global maximum. The temperature is reduced as the MH run progresses and ends with T=1. We adopted a annealing schedule such that on transition t of t_{max} total transitions, $T^{-1} = t/t_{max}$. We ran 16 independent Markov chains for 3000 iterations and chose the dataset that produced the maximum score under the teaching model. With the intention of being sufficiently vague, we set the DP-DPGMM hyper-parameters as follows,

[*] We used the implementation of EM in the *scikit-learn* (Pedregosa et al., 2011) python module's `DPGMM` class.

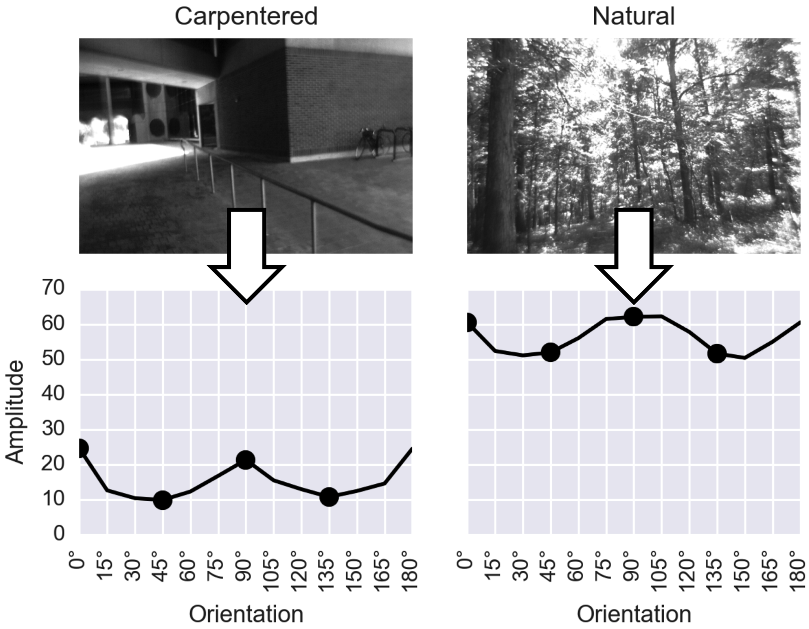

Figure 4: Illustration of the results of the image processing processing procedure. The line represents the amplitudes across orientations for each image; the black dots denote the points used for modeling.

$$\begin{aligned}
\mu_k, \lambda_k, \kappa_k, \nu_k &\sim G, \\
\mu_k &\sim \mathcal{N}(\bar{x}, \mathrm{cov}(x)), \\
\lambda_k &\sim \text{Inverse-Wishart}_{d+1}(\mathrm{cov}(x)), \\
\mu_k &\sim \mathcal{N}(\bar{x}, \mathrm{cov}(x),) \\
\kappa_k &\sim \mathrm{Gamma}(2, 2), \\
\nu_k &\sim \mathrm{Gamma}_d(2, 2),
\end{aligned}$$

where $\mathrm{Gamma}_d(\cdot, \cdot)$ denotes the gamma distribution with lower bound d, $\bar{x}$ is the mean of x, and $\mathrm{cov}(x)$ is the covariance of x.[*] All α and γ were drawn from $\text{Inverse-Gamma}(1, 1)$.

We note that PM-MCMC does not offer the same theoretical guarantees for optimization problems that it does for simulation because PM-MCMC relies on approximate scores; thus the maximum score may be inflated to some degree by estimator error. Pilot simulations revealed that at 1000 IS samples, the variance of $\hat{m}$ for this problem was acceptable. If estimator error is a concern, one may verify the top few optimal datasets post-hoc by re-evaluating their scores a number of times.

The optimal teaching data are plotted along with the data from the original model, with the target model means superimposed in Figure 5. The images closest to the teaching data and the empirical means in Euclidean space are displayed in Figure 6. The images closest to the mean in terms of their primary orientation content were not the best examples to teach the inner categories; the algorithm instead chose images that contrasted the category distributions. This was especially true for the natural images and when the distribution of the inner category had higher variance (Figure 4, bottom row; gray data).

Although the teaching model was only given information about the amplitude of structure at 4 specific orientations in the global image, there were qualitative visual implications of the choice of images used for teaching. Whereas images near the mean for both 'Natural' categories had predominant horizon lines and ground planes, the teaching model made a clearer distinction between the two categories by choosing images

[*] The degrees of freedom of NIW cannot be less than the number of dimensions, thus the lower bound on ν_k must be d.

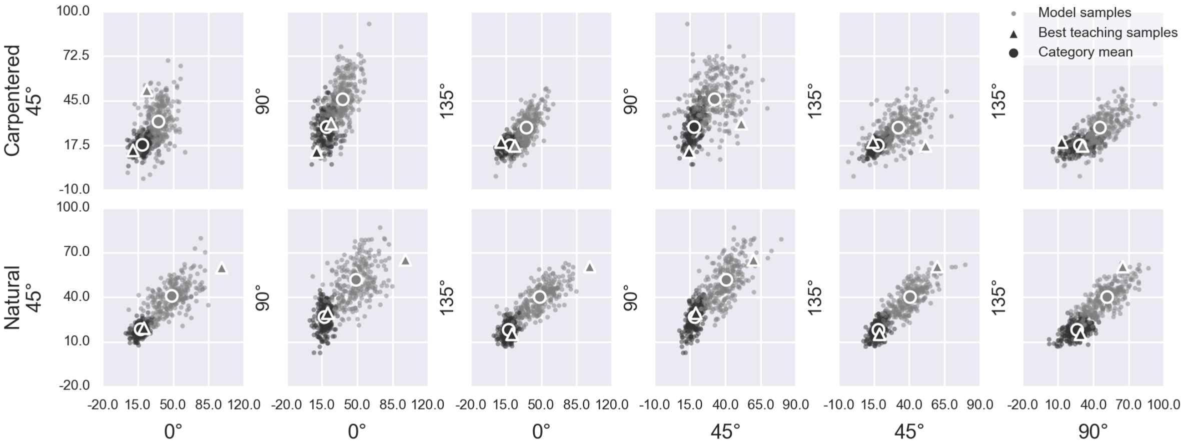

Figure 5: Scene category teaching results. Orientation-orientation scatter plots of random samples from the target model. Different marker colors denote difference inner categories. Circles represent the target model category means and triangles represent the optimal teaching data. The top row shows data from carpentered scenes; the bottom shows data from natural scenes.

with and without a strong horizontal gradient (see Natural Inner A vs Inner B for Teaching Figure 6). The teaching model also more readily distinguished urban (Inner A) from rural-type (Inner B) environments for the carpentered scenes as indicated by the inclusion of cars and buildings in inner category A (see Figure 6). Overall, the teaching model included a wider variety of vantage points (including looking at the ground) for teaching images of all categories, better capturing the variability of the image set. This is opposed to the relatively equal height in the visual field of the centers of the mean images. Again, these observations are purely qualitative, but suggest that the model based on only four dimensions of the orientation distribution captured something meaningful about scene categories.

4.4 Discussion

In this section, we sought to select optimal images for teaching categories of natural scenes. We employed a nested categories model to generalize the DPGMM model used in IDS categorization. Unlike the DPGMM, the DP-DPGMM had no closed-form posterior (due to use of non-conjugate models) and therefore computing the MH acceptance ratio required approximation. The results of the simulation indicate that the best examples for teaching the inner categories of purely natural and carpentered scenes are not the means of the respective categories.

The images that are best for teaching different visual categories under the model exhibit surprising features; the teaching model emphasizes data away from the mean in order to contrast the categories and represent the variation. Although we have not directly tested the effectiveness of the teaching images in visual category learning, the results of this model have potential implications for fields in which visual training and image categorization are important (i.e. medical images, airport scanners, and target detection).

5 Conclusion

The goal of cognitive science is to understand human cognition in common scenarios, however a valid complaint against Bayesian theoretical accounts of cognition is that they are often unable to account for anything more than schematic scenarios. Though we have focused on the problem of teaching categories, we have demonstrated how recent advances in the so-called Bayesian Big Data literature allow Bayesian cognitive modelers to build more compelling models that are applicable to real-world problems of interest to experimentalists.

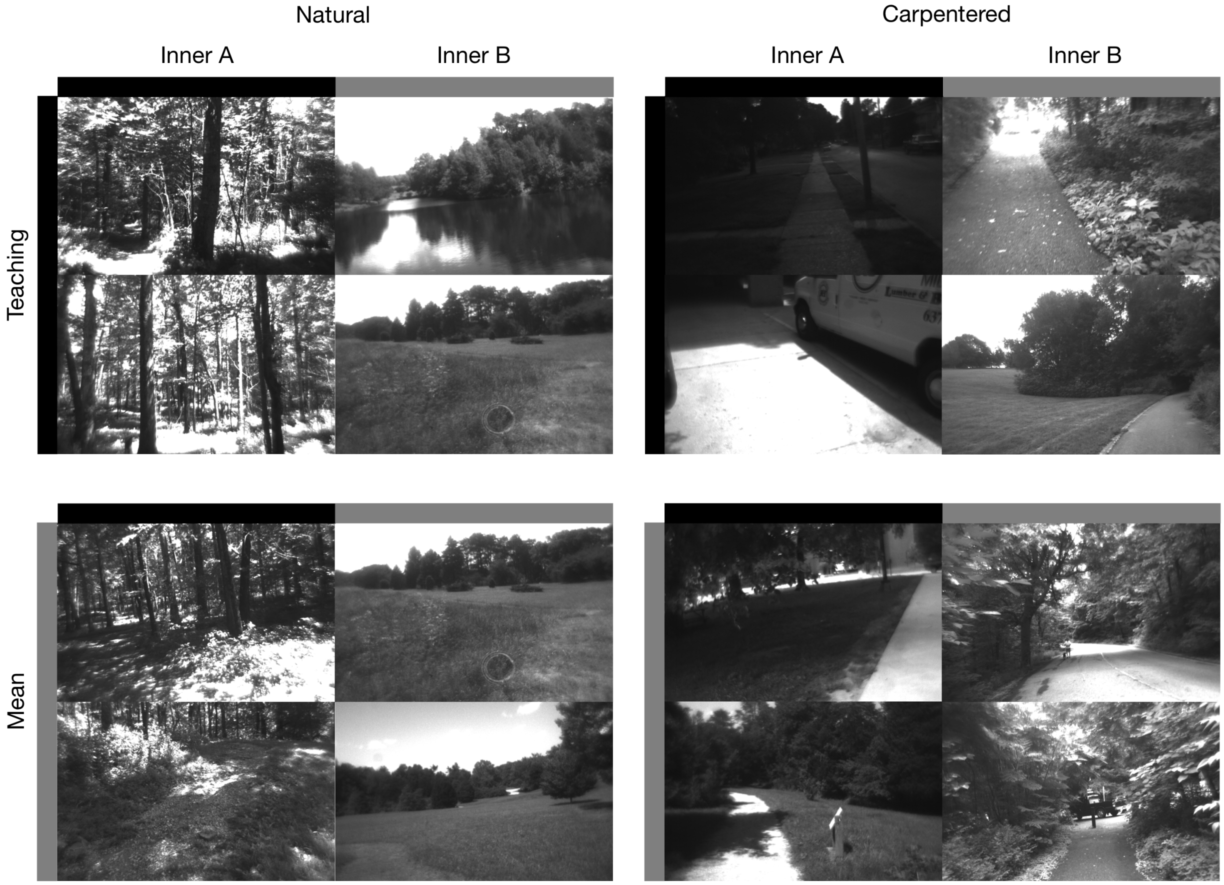

Figure 6: Image data associated with the mean empirical data and optimal teaching data. Top) The two closest images, in Euclidean space, to the optimal teaching datum for each inner category for natural (left) and carpentered (right) scenes. Bottom) The two closest images, in Euclidean space, to the empirical means for each inner category for natural (left) and carpentered (right) scenes.

We began the chapter by briefly discussing the complexity concerns of the Bayesian cognitive modeler, especially in the domain of teaching, and outlined some standard methods of dealing with it. We then discussed pseudo-marginal sampling and applied it to the problem of teaching complex concepts. We applied the PM-MCMC-augmented teaching model to teaching phonetic category models, demonstrating how the framework could be used to contribute to an active debate in linguistics: whether infant-directed speech is for teaching. The results suggested that some of the unintuitive properties of IDS are consistent with teaching—though further work is needed to be directly applicable to IDS. We then applied the teaching model to the far-more-complex problem of teaching nested category models. Specifically, we outlined a framework for learning and teaching scene categories from orientation spectrum data extracted from images. We found that the optimal images for teaching these categories captured a more descriptive picture of the nested category than the mean data; the teaching data seek to convey the ranges of the categories.

This work represents a first step toward a general framework for teaching arbitrary concepts. In the future, we hope to extend the model to teach in richer domains and under non-probabilistic learning frameworks by creating a symbiosis between Bayesian and non-Bayesian methods such as artificial neural networks and convex optimization.

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, *37*(2), 697–725. arXiv: 0903.5480

Andrieu, C. & Vihola, M. (2012). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *25*(2), 43. arXiv: 1210.1484

Banterle, M., Grazian, C., & Robert, C. P. (2014). Accelerating Metropolis-Hastings algorithms: Delayed acceptance with prefetching, 20. arXiv: 1406.2660

Bardenet, R., Doucet, A., & Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. *Proceedings of The 31st International Conference on Machine Learning*, (4), 405–413.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011, September). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–30.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002, May). What's new, pussycat? On talking to babies and animals. *Science (New York, N.Y.) 296*(5572), 1435.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321.

Coppola, D. M., Purves, H. R., McCoy, a. N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 4002–4006.

Cristia, A. & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, 1–22.

de Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the royal statistical society. Series B (methodological)*, *39*(1), 1–38.

Essock, E. A., DeFord, J. K., Hansen, B. C., & Sinai, M. J. (2003). Oblique stimuli are seen best (not worst!) in naturalistic broad-band stimuli: A horizontal effect. *Vision Research*, *43*(12), 1329–1335.

Essock, E. A., Haun, A. M., & Kim, Y. J. (2009). An anisotropy of orientation-tuned suppression that matches the anisotropy of typical natural scenes. *Journal of vision*, *9*(1), 35.1–15.

Feldman, J. (1997, June). The Structure of Perceptual Categories. *Journal of mathematical psychology*, *41*(2), 145–70.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–78.

Fiorentini, G., Planas, C., & Rossi, a. (2012, September). The marginal likelihood of dynamic mixture models. *Computational Statistics & Data Analysis*, *56*(9), 2650–2662.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721–741.

Gergely, G., Egyed, K., & Király, I. (2007, January). On pedagogy. *Developmental science*, *10*(1), 139–46.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, *14*(7), 926–932.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., & Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, *13*, 723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., & Sriperumbudur, B. K. (2009). A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems*, 673–681.

Hansen, B. C. & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of vision*, *4*(12), 1044–1060.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

Kuhl, P. K., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevinkova, E. V., Ryskina, V. L., . . . Lacerda, F. (1997, August). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*(5326), 684–686.

Luce, R. (1977, June). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233.

Maceachern, S. N., Clyde, M., & Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, *27*(2), 251–267.

Maclaurin, D. & Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with Subsets of Data. *arXiv: 1403.5693*, (2000), 1–13. arXiv: 1403.5693

Marin, J.-M. & Robert, C. P. (2008). Approximating the marginal likelihood using copula. *arXiv preprint arXiv:0804.2414*. arXiv: 0810.5474

Markman, A. B. & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613.

McMurray, B., Kovack-Lesh, K., Goodwin, D., & McEchron, W. (2013, November). Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–78.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. arXiv: 5744249209

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution*. University of British Columbia.

Neal, R. M. [Radford M.]. (1999). *Erroneous Results in "Marginal Likelihood from the Gibbs Output"*. University of Toronto.

Neal, R. M. [Radford M]. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249–265.

Patterson, S. & Teh, Y. W. (2013). Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. *Nips*, 1–10.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in neural information processing*, (11), 554–560.

Robert, C. P. & Casella, G. (2013). *Monte carlo statistical methods*. Springer Science & Business Media.

Rufo, M., Martín, J., & Pérez, C. (2010). New approaches to compute Bayes factor in finite mixture models. *Computational Statistics & Data Analysis*, (May 2010).

Schweinhart, A. M. & Essock, E. A. (2013). Structural content in paintings: Artists overregularize oriented content of paintings relative to the typical natural scene bias. *Perception*, *42*(12), 1311–1332.

Schweinhart, A. M., Shafto, P., & Essock, E. A. (submitted). Effects of recent exposure to atypical environmental statistics on orientation perception: analyzing the plasticity of the horizontal effect. *Journal of Vision*.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. a., George, E. I., & Mcculloch, R. E. (2013). Bayes and Big Data: The Consensus Monte Carlo Algorithm, 1–22.

Shafto, P. & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society*.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012, June). Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014, March). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71C*, 55–89.

Sherlock, C., Thiery, A., Roberts, G., & Rosenthal, J. (2013). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *arXiv preprint arXiv: . . . 43*(1), 238–275. arXiv: arXiv:1309.7209v1

Switkes, E., Mayer, M. J., & Sloan, J. A. (1978). Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Research*, *18*(10), 1393–1399.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006, December). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Uther, M., Knoll, M., & Burnham, D. (2007, January). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*(1), 2–7.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007, August). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13273–8.

Wainwright, M. J. (1999). Visual adaptation as optimal information transmission. *Vision Research*, *39*(23), 3960–3974.

Yerebakan, H. Z., Rajwa, B., & Dundar, M. (2014). The Infinite Mixture of Infinite Gaussian Mixtures. In *Advances in neural information processing systems 27* (pp. 28–36). Curran Associates, Inc.