5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–17

Approximate inference by broadening the support of the likelihood

Michael T. Wojnowicz[1,3] MICHAEL.WOJNOWICZ@TUFTS.EDU
Martin Buck[1,2] MARTIN.BUCK@TUFTS.EDU
Michael C. Hughes[3] MICHAEL.HUGHES@TUFTS.EDU
[1] *Data Intensive Studies Center, Tufts University, Medford, MA, USA*
[2] *Dept. of Mathematics, Tufts University, Medford, MA, USA*
[3] *Dept. of Computer Science, Tufts University, Medford, MA, USA*

Abstract

Here we present a framework for approximate statistical inference on a target observation model F via inference on an observation model H with broader support which gives relatively easy and efficient inference. For example, inference is typically easier to derive and implement, and quicker to compute, for an independent binary model than a categorical model, or for an unconstrained model than a model truncated to some possibly exotic region. If the pair (F, H) is chosen such that the likelihood of F dominates that of H, then our framework gives a simple recipe for approximate inference. In the frequentist paradigm, we can substitute the maximum likelihood parameters for H into F. In the Bayesian paradigm, we can use the posterior under likelihood H as an approximate posterior under likelihood F. We show that this *dominated likelihood approximation* provably minimizes an upper bound on an error term between the true data generating distribution and the now tractable model. Experiments on real datasets fitting a Gaussian mixture model truncated to a union of rectangular regions and fitting a categorical Generalized Linear Model (GLM) via an independent binary approximation demonstrate the utility of our approach.

1. Introduction

Suppose we observe the random variables $Y_i \overset{\text{i.i.d.}}{\sim} G$ for $i = 1, \dots, n$, where G is an unknown probability distribution on $\mathcal{Y}$. Suppose further that we wish to model our observations as $Y_i \overset{\text{i.i.d.}}{\sim} F_\theta$, where $\{F_\theta : \theta \in \Theta\}$ is a family of probability distributions on $\mathcal{Y}$ indexed by parameter θ, and where each F_θ has density f_θ w.r.t. some σ-finite measure μ on the measurable space $(\mathcal{Y}, \mathcal{F})$ which is continuous w.r.t θ for each $y \in \mathcal{Y}$.[1] When inference with the desired model F_θ is intractable (often due to intractable normalizing constants; see Sec. 3 for examples), it can be convenient to do approximate inference by artificially broadening the support of the likelihoods. In particular, at inference time, we assume that $Y_i \overset{\text{i.i.d.}}{\sim} H_\phi$, where $\{H_\phi : \phi \in \Phi\}$ is a family of probability distributions indexed by parameter ϕ on a space $(\mathcal{Y}*, \mathcal{F}*)$ where $\mathcal{Y}* \supsetneq \mathcal{Y}, \mathcal{F}* \supsetneq \mathcal{F}$, where each H_ϕ has density h_ϕ w.r.t. some σ-finite measure ν which is continuous w.r.t ϕ for each $y* \in \mathcal{Y}*$, and

[1] In particular if μ is Lebesgue measure, then f becomes a probability density function (pdf) of an absolutely continuous random variable. If μ is the counting measure, then f becomes a probability mass function (pmf) of a discrete random variable.

where

$$\texttt{supp}(H_\phi) \supsetneq \texttt{supp}(F_\theta), \qquad \forall \phi \in \Phi, \theta \in \Theta. \qquad \textit{(Assumption 1)} \qquad (1.0.1)$$

In a typical application, we hope that the observation model is well-specified (i.e. that $F_\theta = G$ for some θ), but we recognize that this is unlikely. However, we can typically have more confidence in the more innocuous assumption of *well-specified support*

$$\texttt{supp}(F_\theta) = \texttt{supp}(G), \qquad \forall \theta \in \Theta \qquad \textit{(Assumption 2)} \qquad (1.0.2)$$

In tandem, Assumptions 1 and 2 guarantee that $H_\phi \neq G$, and in particular that $\texttt{supp}(H_\phi) \supsetneq \texttt{supp}(G)$, for each ϕ. Hence, by using H to model the random variable Y, we are introducing *intentional model misspecification*; we intentionally do inference with a model that has inflated support. In order to construct a valid approximate inference procedure, we require the pair (F_θ, H_ϕ) to satisfy the *dominated likelihood* assumption

$$h_\phi(y) \leq f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi. \qquad \textit{(Assumption 3)} \qquad (1.0.3)$$

In particular, $\Phi \subseteq \Theta$, so we may substitute parameters from Φ into Θ. Since we do inference with the dominated approximate likelihood H to get approximate inference for the target likelihood F, we refer to our strategy as *dominated likelihood approximation* (DLA).

2. Methods

Here we detail and justify doing approximate statistical inference on a target observation model F via inference on a dominated observation model H with broader support. In the frequentist paradigm (Sec. 2.1), we can substitute the maximum likelihood parameters for H into F. In the Bayesian paradigm (Sec. 2.2), we can treat the posterior under likelihood H as an approximate posterior under likelihood F. We can also treat a variational approximate posterior under likelihood H as a doubly approximate posterior under likelihood F.

2.1. Maximum Likelihood

Although the true distribution governing Y is G, we can define the **quasi-maximum likelihood estimator** (QMLE) (White, 1982) for both our target observation model F and our support-broadened observation model H via:

$$\widehat{\theta}_n \triangleq \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{\sum_{i=1}^n \log f_\theta(Y_i)}{n}, \qquad \widehat{\phi}_n \triangleq \underset{\phi \in \Phi}{\operatorname{argmax}} \frac{\sum_{i=1}^n \log h_\phi(Y_i)}{n}$$

In the following, we justify the utility of both (1) $H_{\widehat{\phi}_n}$ and (2) $F_{\widehat{\phi}_n}$ as reasonable models for the random variable Y, even though ϕ indexes a distribution H that has artificially and incorrectly broadened support, i.e. $\texttt{supp}(H) \supsetneq \texttt{supp}(G)$. Model (2) is especially useful; we do inference (QMLE) with a model H that has artificially inflated support, and then substitute the learned parameter into the model F which generates observations with the correct support.

1. Justification for $H_{\widehat{\phi}_n}$. Here we sketch the justification; see Sec. B for details. The justification follows from known properties of the QMLE (White, 1982). By definition, we have that the quasi-MLE $\widehat{\phi}_n$ maximizes $L_n(h) \triangleq \frac{1}{n}\sum_{i=1}^n \log h_\phi(Y_i)$. By the Strong Law of Large Numbers, $\lim_{n\to\infty}\ L_n(h) = \mathbb{E}_G[\log h_\phi(Y)]$ almost surely with respect to G. Hence, with probability 1, $\widehat{\phi}_n$ asymptotically maximizes $\mathbb{E}_G[\log h_\phi(Y)]$, and so asymptotically minimizes $\texttt{KL}\big[G \,\big|\big|\, H_\phi\big]$. In our case, $\texttt{KL}\big[G \,\big|\big|\, H_\phi\big]$ is well-defined even though $\texttt{supp}(H) \supsetneq \texttt{supp}(G)$, because the KL-divergence assigns zero probability mass to any points in $\mathtt{supp}(H) \setminus \mathtt{supp}(G)$. Thus, performing maximum likelihood with the support-broadened model H is reasonable because the procedure asymptotically minimizes $\mathtt{KL}\big[G \;||\; H_\phi\big]$.

2. Justification for $F_{\widehat{\phi}_n}$. We have

$$\begin{aligned}
& h_\phi(y) \leq f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi && \text{Assumption 3} \\
\implies & \mathbb{E}_G[\log h_\phi(Y)] < \mathbb{E}_G[\log f_\phi(Y)] && \text{Monotonicity, Assumption 1} \\
\iff & \mathbb{E}_G[\log g(Y)] - \mathbb{E}_G[\log f(Y)] < \mathbb{E}_G[\log g(Y)] - \mathbb{E}[\log h_\phi(Y)] && \text{algebra} \\
\iff & \mathtt{KL}\big[G \;||\; F_\phi\big] < \mathtt{KL}\big[G \;||\; H_\phi\big] && \text{def. KL}
\end{aligned} \tag{2.1.1}$$

where for simplicity we have assumed that G has density g. Now by item 1, we have that $\widehat{\phi}_n$ asymptotically minimizes $\mathtt{KL}\big[G \;||\; H_\phi\big]$, which by Eq. (2.1.1) is an upper bound on $\mathtt{KL}\big[G \;||\; F_\phi\big]$. Hence, substituting the quasi-MLE $\widehat{\phi}_n$ from family H into family F to obtain model $F_{\widehat{\phi}_n}$ can be justified since $\widehat{\phi}_n$ is the parameter in Φ which (asymptotically) minimizes an upper bound on $\mathtt{KL}\big[G \;||\; F_\phi\big]$.

Discrepancy induced by the dominated likelihood approximation. From Eq. (2.1.1), we see that a discrepancy induced by support-broadening – more specifically, an asymptotic discrepancy between the objective functions whose optimizations produce the approximate ($\mathtt{DLA}$) model $F_{\widehat{\phi}_n}$ rather than the target (quasi) MLE model $F_{\widehat{\theta}_n}$ – is given by

$$\mathcal{D}(F_\phi, H_\phi) \triangleq \mathtt{KL}\big[G \;||\; H_\phi\big] - \mathtt{KL}\big[G \;||\; F_\phi\big] = \mathbb{E}_G[\log f_\phi(Y)] - \mathbb{E}_G[\log h_\phi(Y)] \geq 0 \tag{2.1.2}$$

If the target model is well-specified (i.e. $G = F$), then $\mathcal{D}(F_\phi, H_\phi) = \mathtt{KL}\big[F_\phi \;||\; H_\phi\big]$. If F_ϕ is a truncation of H_ϕ (i.e. $f_\phi = h_\phi / Z_\phi$, where the normalizing constant Z_ϕ gives the probability mass that H_ϕ assigns to some truncation region), then $\mathcal{D}(F_\phi, H_\phi) = -\log Z_\phi$. For an application of this discrepancy, see Sec. D.2.

2.2. Bayesian inference

We begin with a non-asymptotic justification for Bayesian inference with $\mathtt{DLA}$. We then asymptotically relate Bayesian $\mathtt{DLA}$ to frequentist $\mathtt{DLA}$.

Justification for approximate Bayesian inference via dominated likelihoods. Given a prior distribution π on Φ, we obtain the following marginal density relationship from Assumption 3:

$$p_F(y) \triangleq \int_\Phi f_\phi(y)\, \pi(d\phi) \geq \int_\Phi h_\phi(y)\, \pi(d\phi) \triangleq p_H(y)$$

where we have used the same prior π on both Φ and Θ, using the implication from Assumption 3 that $\Phi \subseteq \Theta$. Hence, for any probability distribution Q on Φ within some chosen family $\mathcal{Q}$, we have

$$\log p_F(y) \geq \log p_H(y) \geq \mathtt{ELBO}_H(Q) \tag{2.2.1}$$

where $\mathtt{ELBO}_H(Q) \triangleq \mathbb{E}_Q[\log h(y \mid \phi)] - \mathtt{KL}\big[Q \;||\; \pi\big]$ is the *evidence lower bound*, a traditional lower bound on the log marginal likelihood of a Bayesian model (Blei et al., 2017). In other words, variational inference using the support-broadened likelihood H, which finds $Q \in \mathcal{Q}$ to maximize $\mathtt{ELBO}_H(Q)$, can be understood to maximize both a lower bound on (the logarithm of) the marginal density of the *support-broadened* observation model p_H as well as the *target* observation model p_F. This justifies $\mathtt{DLA}$ in the context of variational Bayesian inference. Moreover, as is well-known, when the family $\mathcal{Q}$ is unconstrained, $\texttt{ELBO}_H(Q)$ is optimized by the true posterior π_H^n, defined as the posterior distribution on Φ after observing $Y_1, \ldots, Y_n$ when assuming the observation model H. Thus, Eq. (2.2.1) also says that exact Bayesian inference for computing the posterior on Φ using H can be seen as producing the probability distribution on Φ which maximizes a lower bound on p_F. This observation justifies Markov chain Monte Carlo (MCMC) sampling under likelihood H, because its estimated posterior converges in distribution to the exact posterior under likelihood H.

Perturbation to posterior concentration. Define π_D^n as the posterior distribution on Φ after observing $Y_1, \ldots, Y_n$ when assuming a generic observation model D with density d. Then $\Phi_0 \subset \Phi$ is called an *asymptotic carrier* for Φ if for any open set U containing Φ_0, $\lim_{n\to\infty} \pi_D^n(U) = 1$ almost surely with respect to G. Berk (1966) shows that under mild conditions, an *asymptotic carrier* for Φ is given by

$$\Phi_0 = \{\phi^* \in \Phi : \phi^* = \operatorname*{argmax}_{\phi\in\Phi} \mathbb{E}_G[\log d(Y \mid \phi)]\} \tag{2.2.2}$$

Since $\mathbb{E}_G[\log d(Y|\phi)]$ is maximized when $D = G$, the set Φ_0 is precisely those $\phi \in \Phi$ that minimize $\texttt{KL}\big[G \;\big|\big|\; D\big]$. Therefore, we can also write

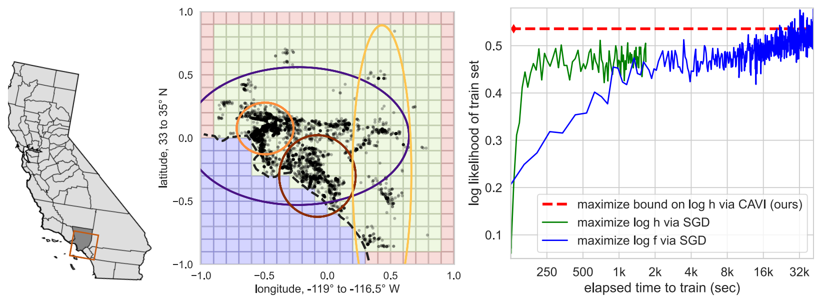
$$\Phi_0 = \{\phi^* \in \Phi : \phi^* = \operatorname*{argmin}_{\phi\in\Phi} \texttt{KL}\big[G \;\big|\big|\; D\big]\}. \tag{2.2.3}$$

So as n increases, the Bayesian posterior concentrates its support on regions of parameter space given by the limiting values of the QMLE. Thus, when choosing to use a dominated observation model $D = H_\phi$, we can apply the argument of Eq. (2.1.1) to find that the posterior π_H^n concentrates on regions of parameter space Φ that minimize an upper bound on $\texttt{KL}\big[G \;\big|\big|\; F_\phi\big]$. This observation justifies the continued relevance in the Bayesian setting of the discrepancy $\mathcal{D}(F, H)$ in Eq. (2.1.2) to quantify the error induced by $\texttt{DLA}$; $\mathcal{D}(F, H)$ gives the amount of perturbation that $\texttt{DLA}$ imposes upon the objective function which is maximized by the asymptotic carrier.

3. Case Study 1: Truncated mixtures of Gaussians for geolocation data

Geolocation data often have constrained support due to water obstacles, political boundaries, or other issues. While using truncated distributions is possible, fitting these in practice may not be easy. Consider a bivariate Normal distribution with truncated support to one rectangular region $\mathcal{Y}$ with bounds a, b such that $y_1 \in [a_1, b_1], y_2 \in [a_2, b_2]$. Estimating the Gaussian parameters μ, Σ of such truncated Gaussians is considered quite difficult (Wilhelm and Manjunath, 2010), solvable via Newton-Raphson steps (Zeng and Gui, 2021) or other iterative methods. In contrast, untruncated Gaussian parameters can be estimated in closed-form by well-known textbook equations.

In this case study, we apply our DLA framework to exploit fast estimation of unconstrained Gaussians for density modeling of human geolocation data. We study Cho et al. (2011)'s open dataset of user "check-ins" to a website called Gowalla, focusing on Southern California as in Lichman and Smyth (2014). Each observation y_i represents the latitude, longitude location of a check-in event. Our large training set (N=67891) is illustrated in Fig. 1. Given such data, we wish to fit a density model in order to predict the locations of new events from new users. We divide the study area into a coarse 20x20 grid of rectangular regions, marking each as either in-bounds (green) or out-of-bounds (blue or red). Ideally, we could use truncated Gaussians to focus on land (green) and ignore water (blue). We thus set our target likelihood f to a mixture of union-of-rectangles truncated Gaussians



method	train log f
CAVI on h	0.536
SGD on h	0.514
SGD on f	0.523
	test log f
CAVI on h	0.515
SGD on h	0.489
SGD on f	0.469

Figure 1: **Application: GMMs for geolocations in southern CA**. *Left:* Ideal model f truncates to the union of green rectangles (land area). Tractable model h (unconstrained mixture of Gaussians) allocates mass to water (blue) or out-of-bounds (red). Ellipses show the 99% high-density-areas of 4 Gaussian clusters fit to data using our CAVI approach. *Right:* Comparison of our approach to directly maximizing ideal likelihood f: DLA yields comparable models in far less time. Table reports each method's mean $\log f$ over all examples.

with K clusters (for details, see App. C), and set tractable h to an *untruncated* Gaussian mixture:

$$f(y) = \textstyle\sum_{k=1}^{K} \pi_k \text{TruncNormPDF}_{a,b}(y|\mu_k, \Sigma_k) \qquad h(y) = \textstyle\sum_{k=1}^{K} \pi_k \text{NormPDF}(y|\mu_k, \Sigma_k). \quad (3.0.1)$$

For fixed K, Assumption 3 holds: $f(y|\phi) > h(y|\phi)$ for all parameters ϕ (all valid frequencies π, means μ, and covariances Σ) as well as all geolocations y in our truncated region $\mathcal{Y}$, because by construction $f(y|\phi) = \frac{1}{Z_{a,b}(\phi)} h(y|\phi)[y \in \mathcal{Y}]$, where $Z_{a,b}(\phi) < 1$ (defined in Eq. (C.0.4)) is the probability mass that h allocates to the union-of-rectangular regions defined by bound vectors a, b.

We consider 3 strategies to estimate ϕ (details in App. C.2). First, maximum likelihood estimation that optimizes either f or h directly via stochastic gradient descent (SGD) using JAX for automatic differentiation (Bradbury et al., 2018). SGD on f requires an expensive bespoke implementation of $Z_{a,b}(\phi)$, which we must call at each iteration. Second, we fit ϕ by maximizing a lower bound on $\log h$ via coordinate ascent variational inference (CAVI). We use well-known closed-form procedures for GMMs available in an off-the-shelf package (Hughes and Sudderth, 2014). Fig. 1 shows that our DLA approach via either SGD or CAVI delivers parameters ϕ that reach competitive likelihood values far faster than direct pursuit of f (less than 30 seconds for CAVI vs. over an hour for SGD on f). We emphasize CAVI's gains require *no customized code* for training, because our theory justifies using existing fast routines for unconstrained GMMs. In a heldout likelihood assessment, our "train on h then plug ϕ into f" strategies do slightly better than direct SGD on f.

4. Case Study 2: Categorical Generalized Linear Models for computer process starts

Bayesian inference with categorical generalized linear models (GLMs) is surprisingly difficult to scale to large datasets (Wojnowicz et al., 2022). For instance, coordinate ascent variational inference (CAVI) (Blei et al., 2017) with multi-logit (a.k.a. softmax) regression faces an expected log-sum-exp term, a notorious blocker to closed-form CAVI (Braun and McAuliffe, 2010; Wang and Blei, 2013). The problematic term is distinct to the multi-class case; it does not appear for binary regression models (e.g. logistic regression). This raises the question: is it possible to use binary regression models to approximate the posterior from a categorical regression model?

Motivated by this question, Wojnowicz et al. (2022) define *categorical-from-binary* (CB) models, a new class of categorical GLMs that are constructed from independent binary (IB) models (products of binary-outcome regression models, such as logistic regressions). Interpreting categorical variables as one-hot encoded vectors, IB models have broadened support (Assumption 1) relative to CB models: IB models have support on K-bit space, a proper superset of one-hot space where CB models are supported. Moreover, CB models dominate IB models (in the strict sense of Assumption 3'; see Eq. (A.0.1) and Sec. D.1). Hence, the DLA framework applies. In particular, the results of Sec. 2 suggest that CAVI with IB models (IB-CAVI), an inference approach which is straightforward and fast, provides a coherent approximation to the true posterior of the CB models.

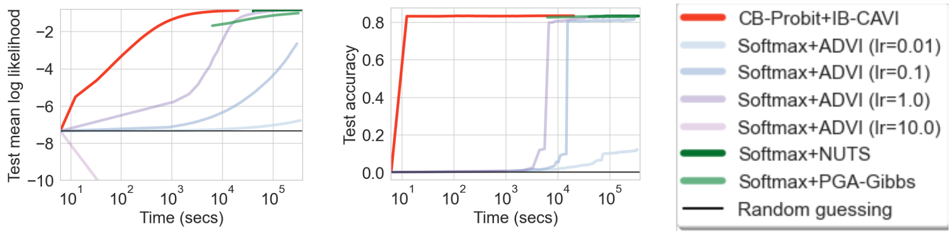


Figure 2: **Application: Scalable Bayesian Categorical GLM for predicting computer process starts.** Bayesian inference methods are compared on a real dataset with $K = 1,553$ categories, $1,553$ covariates, and $17,724$ instances. Prediction quality is measured by holdout log likelihood (left) and accuracy (middle). For ADVI, we try the learning rates $\{0.01, 0.1, 1.0, 10, 100\}$ recommended by Kucukelbir et al. (2017). Figure reproduced from Wojnowicz et al. (2022).

In support of this claim, Fig. 2 compares various Bayesian methods for fitting categorical regressions w.r.t. performance on heldout test data as a function of training time. In particular, a DLA approach (a CB-Probit model, estimated with IB-CAVI) is compared to various well-established variational inference procedures for estimating the most conventional categorical GLM, softmax regression. In particular, the softmax regression is estimated using automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) as well as two "gold standard" MCMC samplers: the No U-Turn Sampler (NUTS) (Hoffman et al., 2014) and a Gibbs sampler available via Pòlya-Gamma augmentation (Polson et al., 2013). Overall, Fig. 2 shows that the DLA approach (CB-Probit+IB-CAVI, plotted in red) delivers *indistinguishable accuracy* and *little-to-no cost in log likelihood* compared to alternative methods for categorical data, while requiring *far less time* to get there. Moreover, DLA gives updates which are exact and optimal, and unlike alternatives does not require correctly choosing a learning rate (as with ADVI) or tuning period length (as with NUTS).

5. Discussion

Our support-broadening procedure applies in principle to any modeling problem and to most common inference procedures, so long as a dominated/dominating likelihood pair can be found. We emphasize that two different use cases exist. First, given a fixed target model, one may specify a dominated approximation. Alternatively, given a fixed tractable model, one may specify a dominating target model. When the target model is a truncated model, a dominated likelihood can be found simply by removing the truncation, as we saw in Sec. 3. However, the pair need not be related through truncation. For instance, the CBM categorical regression models described Sec. D or in Wojnowicz et al. (2022) dominate tractable independent binary models, but are not truncations of them. Although specialized approaches can likely provide better approximations in specific settings, our procedure yields a simple tool for quickly and easily obtaining approximate inference in a wide range of settings, including when scalability is a concern.

Acknowledgments

Author M.B. was supported in part by the Data Intensive Studies Center (DISC) and the Department of Mathematics at Tufts University.

References

Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016, 1973.

R.R. Bahadur. Some limit theorems in statistics. *SIAM*, 1971.

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, and Dougal Maclaurin. JAX: Composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Michael C Hughes and Erik B Sudderth. BNPy: Reliable and scalable variational inference for bayesian nonparametric models. In *Proceedings of the NIPS Probabilistic Programimming Workshop*, 2014. URL https://github.com/bnpy/bnpy.

Robert Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, pages 633–643, 1969.

J Kiefer and J Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, pages 887–906, 1956.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

Moshe Lichman and Padhraic Smyth. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

M.R. Mickey. Test criteria for pearson type iii distributions. *Aerospace Research Laboratories*, pages 40–41, 1963.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, pages 225–228, 1981.

Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(4), 2013.

Halbert White. Nonlinear regression on cross-section data. *Econometrica: Journal of the econometric society*, pages 721–746, 1980.

Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, pages 419–433, 1981.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Stefan Wilhelm and B G Manjunath. tmvtnorm: Truncated multivariate normal and student t distribution. *The R Journal*, 2(1), 2010. URL `https://journal.r-project.org/archive/2010/RJ-2010-005/index.html`.

Michael T Wojnowicz, Shuchin Aeron, Eric L Miller, and Michael Hughes. Easy variational inference for categorical models via an independent binary approximation. In *International Conference on Machine Learning*, pages 23857–23896. PMLR, 2022.

Xinyi Zeng and Wenhao Gui. Statistical inference of truncated normal distribution based on the generalized progressive hybrid censoring. *Entropy*, 23(2), 2021.

Appendix A. Strict Dominated Likelihood Assumption

Note that equality cannot be attained in Eq. (1.0.3) for all $y \in \mathcal{Y}$; by Assumption 1, for any parameters $(\phi, \theta) \in \Phi \times \Theta$, there must be some $A \in \mathcal{F}$ such that $0 < H_\phi(A) < F_\theta(A)$. Hence, in the most straightforward application of the framework, one imposes a *strict dominated likelihood* assumption

$$h_\phi(y) < f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi. \qquad \textit{(Assumption 3')} \qquad \text{(A.0.1)}$$

Indeed, since probability densities must satisfy $\int_{\mathcal{Y}} f_\theta(y)\, \mu(dy) = \int_{\mathcal{Y}*} h_\phi(y)\, \nu(dy) = 1$, Assumption 3' implies Assumption 1.

Appendix B. Quasi Maximum Likelihood Estimation

B.1. Existence of the quasi maximum likelihood estimator (QMLE)

Here we state and verify the assumptions given by (White, 1982) pertaining to the existence of the QMLE $\hat{\phi}_n$. These assumptions are standard in statistical theory when establishing properties such as consistency of estimators.

Assumption B1.1: The independent random variables Y_i for $i \in \{1, \dots, N\}$ have a common joint distribution function G on Ω, a measurable Euclidean space, with Radon-Nikodym derivative $g = \frac{dG}{d\nu}$.

Verification of B1.1: The first assumption simply defines the space Ω and $g = \frac{dG}{d\nu}$ as the Radon-Nikodym derivative of a distribution function G on Ω. In most scenarios, we do not know the underlying true distribution G. In Section 1, the setting is a measurable space $(\mathcal{Y}, \mathcal{H})$ where G is a probability distribution on $\mathcal{H}$ with Radon-Nikodym derivative $g = \frac{dG}{d\nu}$. Thus A1 is met. When we have access to the underlying true distribution such as when generating data for the categorical GLM experiments in Section 3, this assumption is verified as any probability mass function gives a Radon-Nikodym derivative with respect to the counting measure.

Assumption B2.2: The family of distribution functions $\{F_\theta : \theta \in \Theta\}$ and $\{H_\phi : \phi \in \Phi\}$ has Radon-Nikodym derivatives $f_\theta = \frac{dF_\theta}{d\nu}$ and $g_\phi = \frac{dH_\phi}{d\nu}$ which are measurable in x for every $\theta \in \Theta$ and $\phi \in \Phi$ and continuous in θ and ϕ for every $x \in \Omega$. The parameter spaces Θ and Φ are compact subsets of Euclidean space.

Verification of B2.2: The second assumption requires the existence of Radon-Nikodym derivatives for the distributions F_θ and H_ψ and that the Radon-Nikodym derivatives were continuous with respect their parameters for all $y \in \mathcal{Y}$. We assumed this was the case in Section 1. This is met in practice and in our experiments: any probability measure that is absolutely continuous (has a probability density function) or discrete (has a probability mass function) automatically has a Radon-Nikodym derivative with respect to the Lebesgue and counting measure, respectively. This Radon-Nikodym derivative is automatically measurable.

This assumption also requires that the parameter space Θ is compact, which on its face does not hold in practice. For example, we generally do not restrict the space of the mean parameter $\mu \in \mathbb{R}$ of a normal distribution $N(\mu, \sigma^2)$ when estimating it using maximum likelihood; yet $\mathbb{R}$ is not compact. To address this technical issue, it is standard practice to *compactify* the parameter spaces Θ and Ψ if necessary, wherein points are added to the spaces so as to make them compact. See Bahadur (1971), Kiefer and Wolfowitz (1956), and White (1981) for a discussion.

Theorem (White, 1982) - Existence of a QMLE: Given assumptions A1 and A2, there exists a measurable QMLE.

Verification of (White, 1982) - Existence of a QMLE: Since assumptions A1 and A2 are met, we conclude that the QMLEs $\hat{\theta}_n$ and $\hat{\phi}_n$ for both our observation model F and our support-broadened observation model H from Sec. 2.1 exist.

B.2. Convergence of QMLE and Justification for $H_{\hat{\phi}_n}$

Here we detail the argument laid out in Section 2.1 to justify $H_{\hat{\phi}_n}$ as a reasonable model despite the support-inflation. The argument in three steps is as follows:

1. $\hat{\phi}_n$ maximizes $L_n(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \log h_\phi(Y_i)$ is the definition of the QMLE.
2. $\lim_{n \to \infty} L_n(h) = E_G[\log h_\phi(Y)]$ is known as *Mickey's Theorem* (Theorem 2 in Jennrich (1969) and p. 40 in Mickey (1963) contain two proofs) and is a Law of Large Numbers-type result. In fact, we have the stronger result that this convergence is uniform in ϕ.

Theorem (Jennrich, 1969) - Convergence of Quasi-Likelihoods: Let q be a function on $\Omega \times \Theta$ such that $q(z,\theta)$ is a continuous function of θ for each z and a measurable function of z for each θ. Suppose also that $|q(z,\theta)| \leq m(z)$ for all z and θ, where m is integrable with respect to distribution G on Ω. If $Z_1, Z_2, \dots$ is a random sample from G, then $Q_n(z,\theta) \triangleq \frac{1}{n}\sum_{i=1}^n q(z_i,\theta)$ converges to $Q(\theta) \triangleq \int q(z,\theta)dG(z)$ uniformly for all $\theta \in \Theta$ and almost every sequence $\{Z_i\}$.

Verification of Theorem (Jennrich, 1969) - Convergence of Quasi-Likelihoods: The function $q(z,\theta)$ in our application is the quasi log-likelihood for the support-broadened model so that we have the notational correspondence $z \leftrightarrow y$, $\theta \leftrightarrow \phi$, and therefore the correspondence $q(z,\theta) \leftrightarrow \log h_\phi(y)$ and $Q_n(z,\theta) \leftrightarrow L_n(h)$. The only item to verify here is the condition that $q(z,\theta)$ is bounded by a function $m(z)$ that is integrable with respect to the true data-generating distribution G, which is usually unknown in practice. Therefore, this type of condition is often added as an assumption from the outset as in White (1981). However, in our application to categorical GLMs via an independent binary model, we can at least verify this assumption in the case that the covariates x_i are uniformly distributed. Recall from Wojnowicz et al. (2022), that the support-broadened model is an independent binary model whose likelihood for a K-bit observation $\widehat{\boldsymbol{y}}_i \in \{0,1\}^K$ for $i = 1, \dots, n$ is:

$$h_{\text{IB}}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{B}}) = \prod_{k=1}^{K} C(\widehat{\eta}_{ik})^{\widehat{y}_{ik}} \left(1 - C(\widehat{\eta}_{ik})\right)^{1-\widehat{y}_{ik}}$$

In the above, C is an arbitrary cumulative distribution function, where each linear predictor $\widehat{\eta}_{ik} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k$ is formed from known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\widehat{\boldsymbol{\beta}}_k \in \mathbb{R}^M$, and where $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_K) \in \mathbb{R}^{M \times K}$ is a matrix of weights for each combination of covariate and category. We drop the hat accent and use the notation y_i for categorical $y_i \in \{1, \dots, K\}$ so that $\mathbf{e}_{y_i}$ is a one-hot vector corresponding to a unique category. See also Sec. D for a discussion of the IB and CB models. In this context, we have the correspondence $z \leftrightarrow y \leftrightarrow \widehat{\boldsymbol{y}}$ and $\theta \leftrightarrow \phi \leftrightarrow \widehat{\boldsymbol{B}}$ and $q(z,\theta) \leftrightarrow \log h_\phi(y) \leftrightarrow \log h_{\text{IB}}(\widehat{\boldsymbol{y}} = \mathbf{e}_y \mid \widehat{\boldsymbol{B}})$. Therefore, we look for an integrable function $m(\cdot)$ that bounds $|\log h_{\text{IB}}(\widehat{\boldsymbol{y}} = \mathbf{e}_y \mid \widehat{\boldsymbol{B}})|$. We have:

$$\begin{aligned}
|\log h_{\text{IB}}(\widehat{\boldsymbol{y}} = \mathbf{e}_y \mid \widehat{\boldsymbol{B}})| &= |\sum_{k=1}^{K} \log[C(x^T \widehat{\boldsymbol{\beta}}_k)(1 - C(x^T\, \widehat{\boldsymbol{\beta}}_k))]| \\
&\leq \sum_{k=1}^{K} |\log[C(x^T \widehat{\boldsymbol{\beta}}_k)(1 - C(x^T \widehat{\boldsymbol{\beta}}_k))]| \\
&\leq \sum_{k=1}^{K} |\log[C(x^T \widehat{\boldsymbol{\beta}}_k)]|
\end{aligned}$$

Now, the inverse link function $C(\cdot) \in (0,1)$ so that $|\log C(\cdot)| \in (0,\infty)$. Now, $|\log C(\cdot)| < 1$ on $(e^{-1}, 1)$ so we are only concerned with bounding it by an integrable function on $(0, e^{-1})$. This is accomplished with the function $m(\cdot) = |\log C^2(\cdot)|$, which bounds $|\log C(\cdot)|$ from above and is integrable with respect to Lebesgue despite growing rapidly close to zero.

3. With probability 1, $\hat{\phi}_n$ asymptotically maximizes $E_G[\log h_\phi(Y)]$ and so asymptotically minimizes $\texttt{KL}\big[G \;||\; H_\phi\big]$ is *Lemma 3* from Amemiya (1973) or in a slightly more general form as

Lemma 2.2 from White (1980). We state *Lemma 3* from Amemiya (1973) and for completeness reproduce the proof.

Theorem (Amemiya, 1973) - Convergence of QMLE to a Minimizer of Cross-Entropy: Let $Q_n(z, \theta)$ be a measurable function on a measurable space Ω and for each $z \in \Omega$ a continuous function for θ in a compact set Θ. Then there exists a measurable function $\hat{\theta}_n$ such that for all $z \in \Omega$:

$$Q_n(z, \hat{\theta}_n) = \sup_{\theta \in \Theta} Q_n(z, \theta)$$

If $Q_n(z, \theta)$ converges to $Q(\theta)$ a.e. uniformly for all $\theta \in \Theta$, and if $Q(\theta)$ has a unique maximum at $\theta_0 \in \Theta$, then $\hat{\theta}_n$ converges to θ_0 a.e.

Proof: That there exists a measurable function $\hat{\theta}_n$ is *Lemma 2* from (Jennrich, 1969). Let O be an open neighborhood around θ_0. Then $\bar{O}$, the complement of O in Θ, is compact. Therefore, $\max_{\theta \in \bar{O}} Q(\theta)$ exists. Denote $\epsilon = Q(\theta_0) - \max_{\theta \in \bar{O}} Q(\theta)$. Then $|Q_n(z, \theta) - Q(\theta)| < \epsilon/2$ implies $\hat{\theta}_n \in O$. Therefore, $\hat{\theta}_n$ converges to θ_0 a.e.

Verification of (Amemiya, 1973) - Convergence of QMLE to a Minimizer of Cross-Entropy:

Like above, the function $Q_n(z, \theta)$ corresponds to the quasi-likelihood function via the correspondence $L_n(h)$. The first statement of the theorem guarantees the existence of a QMLE, which is a restatement of B.1. By 2, we have that $L_n(h)$ converges to $E_G[\log h_\phi]$ uniformly in ϕ. This limit is the negative *cross-entropy* $H(\cdot, \cdot)$ of H_ϕ relative to G: $H(G, H_\phi) \triangleq -E_G[\log h_\phi(Y)]$. Clearly, maximizing the negative cross-entropy minimizes the cross-entropy, which in turn minimizes the $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$ via the relation: $H(G, H_\phi) = H(G, G) + \texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$. Therefore, if the cross-entropy has an identifiably unique minimum $\phi^* \triangleq \operatorname{argmin}_{\phi \in \Phi} E_G[\log h_\phi(Y)] = \operatorname{argmin}_{\phi \in \Phi} \texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$, we have convergence of the QMLE to a minimizer of $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$:

$$\hat{\phi}_n \xrightarrow{a.s.} \phi^*$$

It is shown in Wojnowicz et al. (2022) that the CB models are non-identifiable at least in the intercepts-only setting. One route to rectify the issue of non-identifiability is a technical topological argument that involves passing to the quotient topology of the parameter space Redner (1981) wherein the set of identifiable parameters are gathered into an equivalence class. In this setting, the maximum likelihood estimator is indeed consistent estimator for the true parameter of interest. However, Wojnowicz et al. (2022) also showed that the IB models are *globally identifiable*, at least in the intercepts-only setting. Thus, since the IB model is the support-inflated model on which we do QMLE, the identifiability condition on the minimizer ϕ^* is satisfied in the intercepts-only setting.

B.3. Bayesian Inference and Perturbation to Posterior Concentration

The justification for $F_{\hat{\phi}_n}$ laid out in the second half of Section 2.2 is based on a classic Bayesian consistency result laid out in (Berk, 1966). This result guarantees a Bayesian formulation of the

$\texttt{KL}\big[G \;\big|\big|\; F_\phi\big]$ upper-bound minimization and asymptotic discrepancy results in 2.1: when using a dominated observation model the posterior π_H^n concentrates on regions of parameter space that minimize $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$. Recall the definition of the asymptotic carrier: $\Phi_0 \subset \Phi$ is called an *asymptotic carrier* for Φ if for any open set U containing Φ_0, $\lim_{n\to\infty} \pi_D^n(U) = 1\,[G]$.

Notation: In the notation below from Berk (1966), $f(\cdot|\theta)$ is a family of densities that are used to model i.i.d $\{Z_i\}$ which are drawn from the distribution G. It is not assumed that G correspond to any of the densities $f(\cdot|\theta)$. The parameter θ belongs to parameter space Θ, a Borel subset of a complete metric space. The densities $f(\cdot|\theta)$ are with respect to some σ-finite measure on range Z. Furthermore, π denote a prior distribution on the Borel subsets of Θ and π_k denotes the posterior distribution of the parameter given $Z_1, \dots, Z_n$. Therefore, the probability that the parameter $\theta \in A$, where A is a Borel subset of Θ is given by:

$$\pi_k A = \frac{\int_A \prod_{i=1}^n f(Z_i \mid \theta) d\pi(\theta)}{\int_\Theta \prod_{i=1}^n f(Z_i \mid \theta) d\pi(\theta)}$$

Berk (1966) applies a modification if necessary to $\pi_k A$, noting that it remains valid if the density $f(\cdot|\theta)$ is substituted with $u(z|\theta) \triangleq g(z)f(z|\theta)$, where g is some positive function almost surely with respect to G. For example, the assumptions below can be less restrictive if $u(\cdot|\theta) = f(\cdot|\theta)/f(\cdot|\theta_0)$ where θ_0 denotes the "true value" of the parameter of interest. Berk (1966) assumes that an appropriate $u(\cdot|\theta)$ has been chosen. Lastly, we let $\bar{H}_n = \frac{1}{n}\sum_{i=1}^n H(Z_i|\theta)$, where $H(\cdot|\theta) = \log u(\cdot|\theta)$. In our $\texttt{DLA}$ framework, $f(\cdot|\theta)$ above could be either of the misspecified models f_θ or h_ϕ. Our interest is in its application to the support-broadened model h_ϕ which is easier to do inference on.

Assumption B3.1: $f(z|\theta)$ is measurable jointly in z and θ; for almost every z, $f(z|\cdot)$ is continuous in θ, at all $\theta \in \Theta$.

Interpretation and Verification of B3.1: This is a standard measurability and continuity in parameters assumption that we have assumed from the outset in Section 1.

Assumption B3.2: For all $\theta \in \Theta$, $G\{z : f(z|\theta) > 0\} = 1$.

Interpretation and Verification of B3.2: (Berk, 1966) states that this assumption avoids situations where we obtain realizations of the random variables $\{Z_i\}$ for which the posterior may be undefined on certain subsets $A \subset \Theta$ and that this is not a restrictive assumption. In the case where it does not hold, it is possible to rectify this by constructing a sequence of sets A_i such that $\bigcup_{i\in\mathcal{I}} A_i = A$ where $G[\lim_{k\to\infty} \pi_k A_i = 0] = 1$. The true data generating distribution G is unknown, so this can added as an assumption from the outset.

Assumption B3.3: For every $\theta \in \Theta$, there is an open neighborhood U of θ such that:

$$\int_U ||H(Z|\cdot)||_\infty \pi(d\theta) < \infty$$

Interpretation and Verification of B3.3: This together with B4 are boundedness conditions so that the dominated convergence theorem can be applied in Berk's proof of the theorem below. Given that we have the freedom to choose $u(\cdot|\theta)$ and the prior distribution π, we can control $||H(Z|\cdot)||_\infty$ by setting $u(\cdot|\theta) = f(\cdot|\theta)/f(\cdot|\theta_0)$ and adjusting π to de-emphasize regions of parameter space where $||H(Z|\cdot)||_\infty$ may be unbounded.

Assumption B3.4: There is an integer $p > 0$ such that for every real number r there is a compact subset $D \subset \Theta$ such that:

$$\int_D \sup \bar{H}_p \leq r$$

Interpretation and Verification of B3.4: Again, this together with B3 are boundendess conditions so that the dominated convergence theorem can be applied. As stated in Berk (1966), in the case of a univariate normal with mean θ and unit variance and setting $u(\cdot|\theta) = f(\cdot|\theta)$, this assumption requires a second finite moment $EZ^2 < \infty$. But if one chooses $u(z|\theta) = f(z|\theta)/f(z|0)$, this can be shown to reduce to $E|Z| < \infty$.

Theorem (Berk, 1966) - Concentration of Posterior under Misspecification: Suppose a model for the random variables $\{Z_i\}$ specifies they are i.i.d. with one of densities $f(\cdot|\theta)$, where the range, Θ, is a Borel subset of a complete separable metric space and $f(\cdot|\theta)$ are densities with respect to a fixed σ-finite measure on range Z. Let π be a prior distribution on the Borel subsets of Θ and let π_n be the posterior distribution of θ given $Z_1, \ldots, Z_n$. If the $\{Z_i\}$ are in fact distributed according to a distribution G, and assumptions (B3.1)-(B3.4) hold, then π_k is almost surely $[G]$ asymptotically carried on the asymptotic carrier.

Appendix C. Supplementary material on Union-of-Rectangle Truncated GMMs

We define a Gaussian mixture model truncated to a union of rectangular regions. Assume the entire space $\mathbb{R}^2$ is divided into an infinite grid of non-overlapping rectangles. A finite number of rectangles R is selected as the truncation region. Each one is indexed by integer r, and has lower bounds $a_r = [a_{1r}, a_{2r}]$ and upper bound $b_r = [b_{1r}, b_{2r}]$ such that an observed 2-dimensional vector y is in the rectangle if $y_1 \in [a_{1r}, b_{1r}]$ and $y_2 \in [a_{2r}, b_{2r}]$. We represent the entire truncation region as the union of the selected rectangles, with bounds $a = \{a_1, \ldots a_R\}$ and $b = \{b_1, b_2, \ldots b_R\}$.

The probability mass Z_r that a GMM allocates to one rectangle r can be defined as

$$\begin{aligned}
Z_{a_r,b_r} &= \int_{y \in [a_r, b_r]} \sum_{k=1}^{K} \pi_k \text{NormPDF}(y|\mu_k, \Sigma_k) dy && \text{(C.0.1)} \\
&= \sum_{k=1}^{K} \pi_k \int_{y \in [a_r, b_r]} \text{NormPDF}(y|\mu_k, \Sigma_k) dy && \text{(C.0.2)} \\
&= \sum_{k=1}^{K} \pi_k (F_k(b_{1r}, b_{2r}) - F_k(b_{1r}, a_{2r}) - F_k(a_{1r}, b_{2r}) + F_k(a_{1r}, a_{2r})) && \text{(C.0.3)}
\end{aligned}$$

where F_k is the multivariate normal CDF under mean μ_k and covariance Σ_k. Each CDF evaluation computes the mass assigned to the rectangle whose lower corner is $(-\infty, -\infty)$ and whose upper right corner has the provided coordinates. The addition and subtraction handles taking the differences of such larger rectangles to compute the area of the desired rectangle. CDF evaluations for fixed parameters are rapid, and can be cached if needed for multiple computations (multiple rectangles that share some boundary corners).

The overall probability mass assigned to the union of R rectangles is simply

$$Z_{a,b}(\phi) = \sum_{r=1}^{R} Z_{a_r,b_r} \tag{C.0.4}$$

As long as the rectangles do not cover all of $\mathbb{R}^2$, we know that $Z < 1$.

C.1. Gowalla data preprocessing

We extract geolocation records from the Gowalla dataset that fall within a rectangular area of southern California, with longitude in (-118.825, -116.675) and latitude in (33.14, 34.86). We then keep only records for individual users that are spaced at least 6 hours apart (to avoid too much time dependence), then only keep users with at least 20 observations. We divide data by user so that each user's records belong to one of train/valid/test. 80% of users are allocated to training (at random), the rest split equally between validation and test. We have 67,925 records in train, 8,375 records in the test set. We did not use any validation set. Very few records had locations over water, likely due to boating, geolocation errors, or people visiting some of the islands off the coast. We removed these for simplicity, but note that our union of rectangle approach could easily handle islands.

C.2. Parameter estimation for Union-of-Rectangle Truncated GMMs

All code for parameter estimation is available here `https://github.com/tufts-ml/gmm_truncated_to_rectangles`.

To ease computation, for all experiments, we parameterized each component's covariance Σ_k as a *diagonal* matrix. While our CAVI procedure could have handled full-rank covariance matrices easily with readily-available off-the-shelf procedures, we chose to focus on diagonal covariance to avoid the expense of maintaining a valid positive definite matrix during stochastic gradient descent.

For maximum likelihood estimation applied directly to either the ideal target f or tractable h, we implement the calculation of h and f, including normalization term $Z_{a,b}$ defined above, in Python using the JAX automatic differentiation library (Bradbury et al., 2018). JAX allows computing the gradients with respect to parameters $\phi = (\pi, \mu, \Sigma)$ without needing to derive the gradient. We pursue stochastic gradient ascent (aka steepest ascent) to maximize each objective ($\log f$ or $\log h$). To handle constrained parameters (frequencies π that must live in the probability simplex, diagonal covariances Σ that must remain positive), we use methods described here[2]. We used $K = 4$ components, a batch size of 6000 and selected the best learning rate (in terms of training performance) from 0.0033, 0.0100, 0.0333, 0.1000, 0.3333. We spent about 10 hours of human effort on this implementation (verifying correctness, etc.). Each training run requires several hours of compute on a modern cluster (using 3 cores each at 3.0 GHz). We expect further effort could make this code more efficient, especially our code for computing Z but it represents an adequate prototype of what a capable researcher might do in a fast day or two of prototyping.

For an alternative estimation of ϕ using our DLA approach, we pursue off-the-shelf variational coordinate ascent (CAVI) code available in the Bayesian nonparametrics for Python (BNPy) package (Hughes and Sudderth, 2014). We fit a GMM with $K = 4$ components (technically a Dirichlet Process GMM, but the truncation level is held fixed at $K = 4$). Both the variational-E-step and

[2] https://www.cs.tufts.edu/cs/136/2023s/cp4.html#transform_to_unconstrained

variational M-step leverage closed-form updates to cluster assignment probabilities and parameter posteriors. Once fit, we use posterior means to get point estimates of $\phi = (\pi, \mu, \Sigma)$. We spent about an hour on this implementation, mostly writing the wrapper code to call the off-the-shelf routines on our dataset and feeding the learned parameters ϕ into our implementation of likelihood f to make results tables.

Appendix D. Supplementary material on Categorical-from-Binary (CB) models

Categorical-from-binary (CB) models (Wojnowicz et al., 2022) are categorical GLMs for which independent binary (IB) models (products of binary-outcome regression models) naturally provide a dominated likelihood (in the strict sense of Assumption 3').

The the likelihood for K-bit observation $\widehat{\boldsymbol{y}}_i = (\widehat{y}_{i1}, \dots \widehat{y}_{iK}) \in \{0,1\}^K$ under an *independent binary* (IB) model is

$$h_{\text{IB}}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{B}}) = \prod_{k=1}^{K} C(\widehat{\eta}_{ik})^{\widehat{y}_{ik}} \left(1 - C(\widehat{\eta}_{ik})\right)^{1-\widehat{y}_{ik}}, \tag{D.0.1}$$

where C is an arbitrary cumulative distribution function, where each linear predictor $\widehat{\eta}_{ik} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k$ is formed from known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\widehat{\boldsymbol{\beta}}_k \in \mathbb{R}^M$, and where $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_K) \in \mathbb{R}^{M \times K}$ is a matrix of weights for each combination of covariate and category.

A *categorical-from-binary-via-conditioning* (CBC) model assigns probabilities to categorical observations $y_i \in \{1, \dots, K\}$ by conditioning the IB model on the event that the vector has exactly one positive entry:

$$f_{\text{CBC}}(y_i{=}k \mid \boldsymbol{B}) = \frac{C(\eta_{ik}) \prod_{j \neq k} (1 - C(\eta_{ij}))}{\sum_{\ell=1}^{K} C(\eta_{i\ell}) \prod_{j \neq \ell} (1 - C(\eta_{ij}))} \tag{D.0.2}$$

for categories $k = 1, ..., K$. Here, we distinguish parameters $\boldsymbol{B}$ and observations y_i from the IB case by dropping the hat accent.

A *categorical-from-binary-via-marginalization* (CBM) model produces category probabilities by normalizing the marginal probabilities of success $\{C(\eta_{ik})\}_{k=1}^K$ from an IB model:

$$f_{\text{CBM}}(y_i = k \mid \boldsymbol{B}) = \frac{C(\eta_{ik})}{\sum_{\ell=1}^{K} C(\eta_{i\ell})}. \tag{D.0.3}$$

for all categories $k \in \{1, \dots, K\}$.

D.1. Applicability of the *Dominated Likelihood Approximation* Framework

Here we defend that IB models (D.0.1) give dominated likelihood approximations (`DLA`) to both CBM (D.0.3) and CBC (D.0.2) models:

1. By interpreting categorical variables as one-hot encoded vectors, we see that IB models have broadened support (Assumption 1) relative to CB models.

$$G = \text{true distn}, \qquad F_\theta = \text{CB model} \qquad\qquad \text{one-hot space} \subset \text{K-bit space}$$
$$H_\phi = \text{IB model} \qquad\qquad \text{K-bit space}$$

2. CB models dominate IB models (in the strict sense of Assumption 3'; see Eq. (A.0.1)). CBC models are truncations (of IB models, to one-hot encoded space), and hence satisfy the dominated likelihood assumption trivially. CBM models satisfy the dominated likelihood assumption as well, although not by truncation. For a proof, see Wojnowicz et al. (2022, Sec. B.3).

Hence, so long as the true data generating process is supported by categorical outcomes (Assumption 2), the DLA framework essentially applies. There is, however, one caveat. In this paper, we have presented the framework in terms of an $i.i.d$ assumption on the observations, whereas GLMs require a relaxed assumption of independence. We defer formal generalization of our framework to the case of independent observations to a future development of this workshop paper.

D.2. Experiment: Assessing the quality of the dominated likelihood approximation

Eq. (2.1.2) provides what we might call an *asymptotic objective discrepancy*; that is, the asymptotic discrepancy in objective function induced by DLA. What would be of greater interest, however, is an expression of the *modeling discrepancy*, such as

$$\mathtt{KL}\big[F_{\widehat{\theta}} \;||\; F_{\widehat{\theta}_{\text{DLA}}}\big] \tag{D.2.1}$$

where $\widehat{\theta}$ is a parameter estimate obtained through some inference procedure (maximum likelihood, MCMC posterior expectation, variational posterior expectation, etc.) applied to the target model F and $\widehat{\theta}_{\text{DLA}}$ is the parameter obtained by applying the same inference procedure in the context of a dominated likelihood approximation. Eq. (D.2.1) is useful because it tells us the extent to which DLA provides a good approximation to inference with the target model.

Unfortunately, we do not know how to compute the modeling approximation discrepancy of Eq. (D.2.1) without performing inference with the target model F, for which inference is by assumption difficult to obtain. However, the asymptotic objective discrepancy of Eq. (2.1.2) is often much easier to compute, or at least approximate via Monte Carlo sampling.

In this experiment, we investigate the extent to which the objective discrepancy can provide a proxy for the modeling discrepancy. In particular, we focus on CBC models (Eq. (D.0.2)), which are truncated IB models. When we approximate a CBC likelihood with an IB likelihood (representing categories as one-hot vectors), the asymptotic discrepancy in the objective function for maximum likelihood can be estimated by

$$\widehat{\mathcal{D}}(F_{\theta=\phi}, H_\phi) \triangleq \frac{1}{n}\sum_{i=1}^{n}(-\log Z_i), \quad \text{where} \quad Z_i \triangleq \sum_{\ell=1}^{K} C(\eta_{i\ell}) \prod_{j\neq\ell}(1 - C(\eta_{ij})) \tag{D.2.2}$$

Here, Z_i is the probability mass assigned by the IB model to the event that the i-th observation is one-hot encoded.[3] We investigate the information provided by this formula in the context of variational inference, interpreting $\widehat{\theta}$ as the variational posterior expectation.

[3] In the limiting case where the IB model assigns all its probability mass to the event that the i-th observation is one-hot encoded, i.e. $Z_i = 1$, we have $-\log Z_i = 0$, and so the discrepancy for the i-th observation attains the minimum

Method. We simulated categorical regression datasets following the procedure of Wojnowicz et al. (2022, Sec G.1). In particular, we formed three collections of 36 simulated datasets generated from a softmax categorical GLM with $K = 5$ categorical outcomes and $M = 5$ covariates (which along with the intercept gives $P = K(M+1) = 30$ parameters). Each collection of datasets was defined by the number of observations (or examples), N, given as a multiple of the number of predictors $N \in \{1P, 10P, 100P\}$. For each collection, datasets were generated with different levels of category predictability (Wojnowicz et al., 2022, Sec. G.2).

The target model for inference F was a CBC-Probit likelihood with an isotropic Gaussian prior. The dominated approximation H was given by an IB-Probit model (the product of binary probit regressions). Inference for H was performed by a lightweight coordinate ascent variational inference (CAVI) algorithm given in Wojnowicz et al. (2022, Sec. D). Inference for F was performed by automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017). In both cases, the true parameter was approximated by the variational posterior expectation.

Results. Fig. 3 reveals that the computable estimate in Eq. (D.2.2) for the asymptotic objective discrepancy provides a potentially useful proxy for the typically unknown modeling discrepancy of Eq. (D.2.1). In particular, for a fixed number of observations, N, we find that the quality of the IB approximation improves as the IB model assigns higher probability to one-hot space. In this way, we can assess the quality of the approximation to the CBC model without ever fitting it. Moreover, we see that the modeling discrepancy (along the y-axis) decreases as N increases. This relationship is also predicted by the asymptotic objective discrepancy.

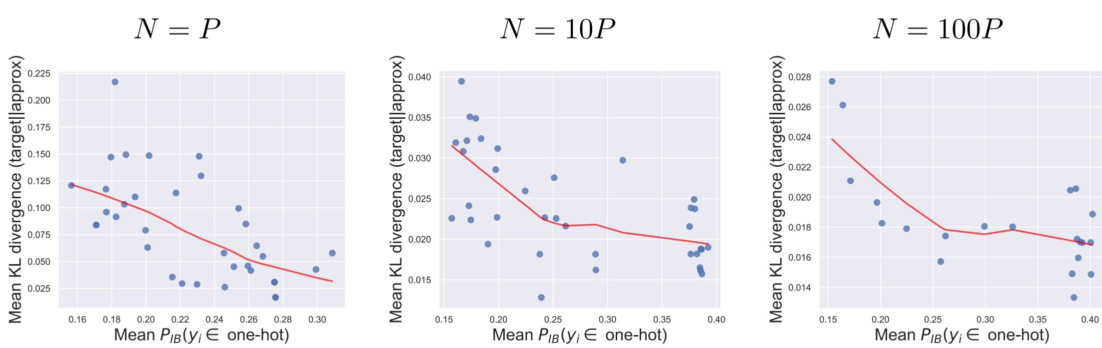


Figure 3: **Application: When do independent binary models give good approximations to categorical GLMs?** Each panel summarizes results on multiple simulated categorical regression datasets where the number of observations, N, is expressed as a multiple of the number of parameters P. The y-axis shows the typically unknown modeling discrepancy of Eq. (D.2.1); in this case, it is the empirical mean KL divergence from an independent binary (IB) approximation to a categorical-from-binary (CB) target model. The x-axis shows a quantity related to the analytically computable estimate from Eq. (D.2.2) of the asymptotic objective discrepancy; in this case, it is $\frac{1}{N}\sum_{i=1}^{N} Z_i$, the empirical mean probability that the IB model assigns to one-hot encoded space. We see that for fixed N, the estimated objective discrepancy serves as a proxy for the modeling discrepancy. Moreover, the modeling discrepancy (along the y-axis) decreases as N increases.

value of $\mathcal{D}_i = 0$. In the limiting case where the IB model assigns none of its probability mass to the event that the i-th observation is one-hot encoded, i.e. $Z_i = 0$, we have $-\log Z_i = \infty$, and so the discrepancy for the i-th observation attains attains the maximum value of $\mathcal{D}_i = \infty$.