DIFFLTM: CONTINUOUS DIFFUSIONS LARGE TABULAR MODELS FOR CONDITIONAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generation is a key capability to be supported by Large Tabular Models, enabling a number of applications. Accurate conditional generation has to contend with two key challenges: (i) Data comprise a mixture of numerical and categorical entries, the latter often with high cardinality and heavily unbalanced proportions; (ii) Conditional generation becomes increasingly difficult for rare sub-populations. The best models in this class use a hybrid of continuous and discrete diffusions. In this paper, we design a continuous diffusion model with strong generation quality. An important element is the use of cross-entropy loss for categoricals, for which we provide a first theoretical justification. We introduce metrics that focus on rare subpopulations and extend conditional generation techniques. Finally, we show that our model (trained on labor data) can capture classical insights from econometrics.

1 INTRODUCTION

The term Large Tabular Model (LTM) was coined by Van Breugel & Van Der Schaar (2024) to denote generative models for tabular data with cross-dataset modeling capabilities, enabling a variety of tasks. Challenges in developing such models (already highlighted in the cited paper) include:

(i) Tabular data present a mixture of categorical and numerical data, and categoricals are often high-cardinality or heavily skewed.

(ii) Standard evaluation metrics (e.g., FID, precision-recall) are often hard to interpret or misleading in the tabular setting, and synthetic samples are hard to interpret.

(iii) Conditional generation is arguably the most important capability for such models, as it enables a number of applications: forecasting; scenario analysis; imputation, and so on.

In this paper, we address these challenges. As a result, we design a generative tabular model with state of the art conditional generation behavior (Figure 1). We train the model on sizable census datasets (up to 7M rows), and compare it with alternative approaches using newly designed faithfulness metrics. In the rest of this introduction we summarize some of our innovations, contrasting them with earlier literature and providing pointers to the rest of the paper.

Metrics. An important concern with tabular data is that humans do not have good intuition of when the generated data is 'good.' This is in sharp contrast with images or language models. We address this in two ways. *First,* we define interpretable fidelity metrics that emphasize the accuracy on rare subpopulations (Section 2.2). While in hindsight this is a natural idea, previous work uses metrics that are dominated by majority classes (Shi et al., 2025; Mueller et al., 2025).

Second, we argue that an LTM must correctly reproduce *non-obvious statistical patterns* in the target population. To implement this idea, we consider two major discoveries in econometrics: the counterfactual analysis using synthetic controls and Mincer equation on returns to education. In a sense, checking such patterns is the analogue of the '6-fingers test' for image generation. We demonstrate that our LTM reproduces these patterns (Figure 2).

Architecture. We develop a diffusion LTM (DiffLTM). Given the mixture of categorical and numerical data in tabular datasets, an important choice is whether to use discrete or continuous dif-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

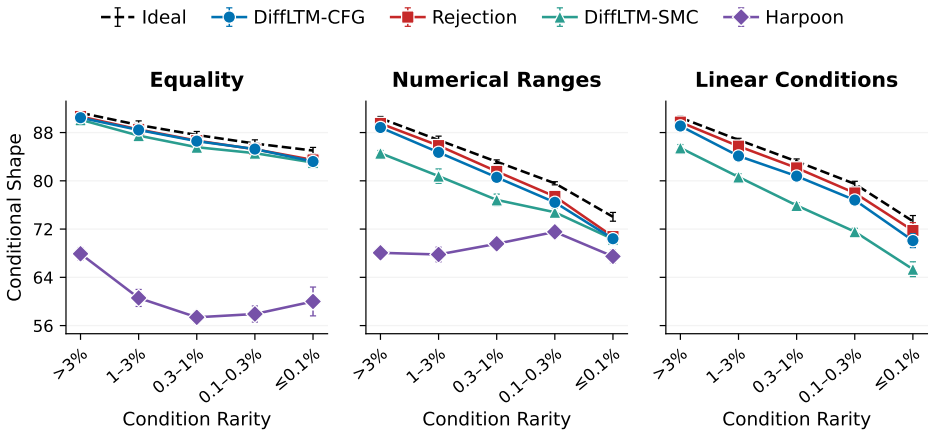


Figure 1: DiffLTM conditional generation is near-optimal across condition types and rarity levels on ACSLabour. *DiffLTM-CFG* tracks *Rejection* sampling (exact conditional samples from DiffLTM) and stays close to *Ideal* (the score between two independent real samples). *DiffLTM-SMC* trails on numerical conditions, and the *Harpoon* baseline is worse throughout (Section 5; Appendix D.3).

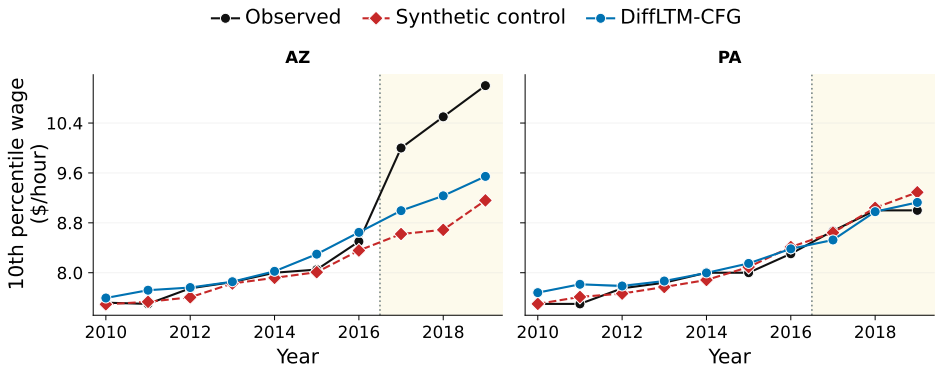


Figure 2: DiffLTM generates counterfactuals consistent with synthetic control for unseen subpopulations. Trained without each state's post-2016 rows (shaded), *DiffLTM-CFG* predicts the held-out Pennsylvania wages (placebo, right), and in Arizona (left) produces a no-policy counterfactual well below the observed post-2017 wages. (Section 5.3; Appendix E).

fusion, or a hybrid. State of the art methods such as TabDiff (Shi et al., 2025) are hybrid, while examples of purely continuous methods are Mueller et al. (2025; 2026); Si et al. (2025).

We demonstrate that excellent performance can be achieved by a purely continuous model, provided embeddings and loss function are chosen carefully. We compare our model with continuous diffusion baselines and demonstrate state of the art unconditional generation performance (Table 1).

While cross-entropy (CE) loss has been chosen in applications in the past (Mueller et al., 2025; 2026; Eijkelboom et al., 2024), theoretical arguments appear to support score matching (a.k.a. MSE) loss (Chewi et al., 2026). We present empirical evidence that CE with learned embeddings is preferable (Table 2), and show mathematically that –under a natural misspecification scenario– this is indeed the optimal choice (Theorem 1).

Conditional generation. We study three types of conditions: (i) *equality conditions*: a subset of columns take specified values; (ii) *range conditions*: numerical column i lies in a range $[L_i, U_i]$; (iii) *linear conditions*: A linear combination of the numerical columns x_i satisfies $L \leq \sum_i w_i x_i \leq U$. A number of important applications can be modeled using conditions of these three types. We describe how to adapt classifier free guidance (CFG) (Ho & Salimans, 2022)

Table 1: Comparison of DiffLTM, TabRep (Si et al., 2025), and TabCascade (Mueller et al., 2026) on unconditional generation. Higher values indicate better performance for all metrics except the DCR Ratio (mean ± SE over three training seeds; **Best**, second-best).

Dataset	Method	Shape	Trend	CShape	MLE	DCR ratio
ACSLabour	TabRep	93.53 ± 0.05	87.00 ± 0.09	88.57 ± 0.05	98.43 ± 0.03	1.000 ± 0.000
	TabCascade	**99.00** ± 0.01	97.56 ± 0.01	94.10 ± 0.04	98.81 ± 0.03	1.000 ± 0.000
	DiffLTM	**99.01** ± 0.03	**97.67** ± 0.03	**95.40** ± 0.02	**99.43** ± 0.02	1.000 ± 0.000
ACSPubCov	TabRep	98.52 ± 0.08	97.60 ± 0.11	97.96 ± 0.04	99.14 ± 0.00	1.000 ± 0.000
	TabCascade	97.79 ± 0.03	95.75 ± 0.07	94.06 ± 0.16	86.25 ± 0.15	0.999 ± 0.000
	DiffLTM	**99.51** ± 0.05	**99.10** ± 0.07	**98.91** ± 0.03	**99.51** ± 0.00	1.000 ± 0.000
ACSMobility	TabRep	98.92 ± 0.03	98.21 ± 0.04	97.68 ± 0.01	98.91 ± 0.05	1.033 ± 0.035
	TabCascade	98.76 ± 0.00	97.86 ± 0.01	96.62 ± 0.11	97.30 ± 0.11	0.847 ± 0.013
	DiffLTM	**99.37** ± 0.04	**98.84** ± 0.12	**98.07** ± 0.04	**99.46** ± 0.02	1.114 ± 0.037

to enable these conditions, and show state-of-the art (and near-optimal) performances. We also adapt inference-time interventions (in particular Sequential Monte Carlo (SMC) (Doucet et al., 2001)) and compare the results with CFG, showing that both our methods significantly outperform baselines in tabular conditional generation (Shankar et al., 2026) (Figure 1).

2 METRICS

2.1 DATASETS

All datasets considered in this work are derived from the American Community Survey Public Use Microdata Sample (ACS PUMS), using Folktables (Ding et al., 2021). We use two datasets from this suite: ACSPubCov (originally constructed to build predictive models for public health-insurance coverage), and ACSMobility (for predictive models of residential mobility.)

In addition, we construct *ACSLabour*, a dataset which retains continuous log wage, as well as a broad set of labor and demographic columns, including high-cardinality occupation, industry, geographic, and birthplace variables. (See Appendix B.1 for detailed descriptions.)

2.2 FIDELITY METRICS

We assess the quality of generated data by comparing two samples: n^{test} hold-out rows from the target table; n^{synth} rows generated by our model (often $n^{\text{test}} = n^{\text{synth}}$). All fidelity metrics can be thought of as statistical estimators of the corresponding distributions P^{test}, P^{synth}. Typically, the 'plug-in' estimator that replaces P^{test}, P^{synth} by their empirical versions $\hat{P}^{\text{test}}$, $\hat{P}^{\text{synth}}$ is used. We provide full definitions of newly defined metrics, while deferring to Appendix B.2 for standard ones. Unless otherwise stated, the normal range of these metrics is $[0, 1]$, with 1 corresponding to the maximum possible value.

Shape. Shape (SDMetrics, 2023) measures fidelity of univariate marginal distributions, averaged uniformly over the column index.

Trend. Trend (SDMetrics, 2023) measures the fidelity of pairwise marginals (for categorical columns) or pairwise correlations (for numerical ones), averaged uniformly over column indices.

Conditional Shape. We introduce a new metric named *conditional shape*, which is the mean shape *within subpopulations*. For fixed categorical conditioning column j^* and categorical value a (the case of a numerical column is analogous), let $\hat{P}^{\text{real}}_{j|j^*,a}$ and $\hat{P}^{\text{synth}}_{j|j^*,a}$ be the empirical distributions of column j (respectively, real and synthetic) for rows with $x_{j^*} = a$. The shape score of column j is $S_{j|x^*_j=a} = 1 - \text{TV}(\hat{P}^{\text{test}}_{j|j^*,a}, \hat{P}^{\text{synth}}_{j|j^*,a})$ and we define the shape score of subpopulation (j^*, a) by averaging over all columns except j^*:

$$S(j^*, a) := \frac{1}{m-1} \sum_{j \neq j^*} S_{j|j^*,a}. \tag{1}$$

Table 2: Comparison across architectural variants, including MSE versus CE loss, representation (Repr) versus logits parameterization, and the effect of learned embeddings. Higher values indicate better performance for all metrics (mean ± SE over three training seeds; **Best**, second-best).

Dataset	Architecture	Shape	Trend	CShape	MLE
ACSLabour	MSE+Repr	89.93 ± 0.24	81.40 ± 0.37	71.25 ± 0.35	95.59 ± 0.22
	MSE+Logits	94.48 ± 0.21	89.89 ± 0.38	83.40 ± 0.52	98.30 ± 0.18
	CE+Logits	98.61 ± 0.02	96.75 ± 0.05	94.78 ± 0.04	99.11 ± 0.03
	CE+Logits+Learnt	**99.01** ± 0.03	**97.67** ± 0.03	**95.40** ± 0.02	**99.43** ± 0.02
ACSPubCov	MSE+Repr	99.32 ± 0.10	98.74 ± 0.15	97.76 ± 0.29	99.09 ± 0.17
	MSE+Logits	99.42 ± 0.11	98.89 ± 0.18	98.08 ± 0.16	99.28 ± 0.12
	CE+Logits	**99.62** ± 0.02	**99.26** ± 0.04	**98.89** ± 0.02	99.34 ± 0.07
	CE+Logits+Learnt	99.51 ± 0.05	99.10 ± 0.07	**98.91** ± 0.03	**99.51** ± 0.00
ACSMobility	MSE+Repr	99.25 ± 0.04	98.61 ± 0.09	94.88 ± 0.24	99.01 ± 0.08
	MSE+Logits	99.27 ± 0.05	98.61 ± 0.10	97.08 ± 0.67	99.25 ± 0.06
	CE+Logits	**99.54** ± 0.02	**99.16** ± 0.02	**98.10** ± 0.04	99.23 ± 0.05
	CE+Logits+Learnt	99.37 ± 0.04	98.84 ± 0.12	**98.07** ± 0.04	**99.46** ± 0.02

If $\mathcal{A}_j$ is the set of values appearing in column j, we report[1]

$$\mathrm{CShape} := \frac{1}{m}\sum_{j=1}^{m}\frac{1}{|\mathcal{A}_j|}\sum_{a\in\mathcal{A}_j} S(j,a). \tag{2}$$

Note that subpopulations receives equal weight regardless of their frequency, hence emphasizing the importance of rare subpopulations with respect to standard Shape. In the appendix, we also define **Conditional Shape (Rare)**, (cf. Appendix B.2) which further emphasizes rare subpopulations.

MLE Ratio. (Zhang et al., 2024) We train an XGBoost model on synthetic data and evaluate it on held-out real data (*Train Synthetic, Test Real*; TSTR). We divide this score by that of an identically configured XGBoost model trained on real data and evaluated on the same held-out set (*Train Real, Test Real*; TRTR). See Appendix B.2.

DCR median ratio. (Zhang et al., 2024) Let $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$ denote the real training and held-out datasets, respectively, and let $\mathcal{D}_{\mathrm{syn}}$ denote the synthetic dataset. Let $d(\boldsymbol{z}, \mathcal{D}_{\mathrm{train}}) := \min_{\boldsymbol{z'}\in\mathcal{D}_{\mathrm{train}}} d(\boldsymbol{z}, \boldsymbol{z'})$, where $d(\cdot,\cdot)$ is the Euclidean distance between rows $\boldsymbol{z}$ and $\boldsymbol{z'}$ after categorical columns are one-hot encoded. Defining $A_{\text{synth}} := \mathrm{median}_{\boldsymbol{z}\in\mathcal{D}_{\text{synth}}} d(\boldsymbol{z}, \mathcal{D}_{\mathrm{train}})$ and $A_{\text{test}} := \mathrm{median}_{\boldsymbol{z}\in\mathcal{D}_{\text{test}}} d(\boldsymbol{z}, \mathcal{D}_{\mathrm{train}})$, we let $\mathrm{DCR\ Ratio} := A_{\text{synth}}/A_{\text{test}}$. Values substantially below one are indicative of memorization.

Results: unconditional generation Table 1 compares DiffLTM on multiple metrics defined above with two state-of-the-art methods (Si et al., 2025; Mueller et al., 2026) and illustrates that DiffLTM outperforms these on *all* considered metrics on *all* considered datasets. We point out that earlier benchmarking of these models was carried out on significantly smaller datasets than our study. All datasets considered in Si et al. (2025) have less than 100K rows; and all datasets but one in Mueller et al. (2026) have less than 102K rows, with one exception comprising 540K rows. In contrast our study focuses on moderate-size datasets with ~ 5M rows.

3 Diffusion architecture and empirical results

In this section, we outline the diffusion architecture, emphasizing treatment of categoricals. We defer to Appendices C and D for precise details regarding the architecture used in our experiments.

Continuous embedding. Consider a table with (without loss of generality) the first m_{num} columns being numerical, and the latter m_{cat} columns being categorical. Let categorical column j take values in $\{1,\dots,C_j\}$. For each such column j, we associate an embedding

$$\phi_j : \{1,\dots,C_j\} \to \mathbb{R}^{d_j}.$$

[1] We sometimes exclude values whose frequency is below a small threshold.

A widely used choice is the *one-hot* embedding $\phi_j(i) = e_i \in \mathbb{R}^{C_j}$. Another choice is *learned embeddings* where $\phi_j(i) \in \mathbb{R}^{d_{\text{emb}}}$ is a vector of learnt parameters. We enforce the normalization $\|\phi_j(i)\| = \sqrt{d_{\text{emb}}}$. For a data point $\boldsymbol{x} = (x_1, \ldots, x_{m_{\text{num}}+m_{\text{cat}}})$, we thus have a continuous representation $\boldsymbol{r} \in \mathbb{R}^D$

$$\boldsymbol{r} = (x_1, \ldots, x_{m_{\text{num}}}, \phi_1(x_{m_{\text{num}}+1}), \ldots, \phi_{m_{\text{cat}}}(x_m)) \,, \tag{3}$$

where $D = m_{\text{num}} + \sum_{j=1}^{m_{\text{cat}}} d_j$.

Diffusion noising process. For each training example, we sample $t \sim \text{Unif}[0, 1]$, which corresponds to a "noise level". This level is then mapped to a variance via a monotonically increasing noise schedule $\sigma : [0, 1] \to [\sigma_{\min}, \sigma_{\max}]$. In our experiments, we use the *power-mean schedule* of Karras et al. (2022). The forward corruption occurs in the continuous representation space, so that for a row's representation $\boldsymbol{r} \in \mathbb{R}^D$ and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I_D)$ the noising process is $\boldsymbol{r}_t := \boldsymbol{r} + \sigma_t \boldsymbol{z}$.

Denoiser output head. The denoiser f_θ receives the noisy representation $\boldsymbol{r}_t$ of target row $\boldsymbol{x}$ and a sinusoidal embedding of the diffusion time t. Numerical columns are always predicted directly; for categorical columns, we consider two parameterizations:

1. Representation. The denoiser predicts the clean continuous representation,

$$f_\theta(r_t, t) = (\widehat{\boldsymbol{r}}_{\text{num}}, \widehat{\boldsymbol{r}}_{\text{cat},1}, \ldots, \widehat{\boldsymbol{r}}_{\text{cat},m_{\text{cat}}}) \,,$$

where $\widehat{\boldsymbol{r}}_{\text{num}} = (\widehat{x}_1, \ldots, \widehat{x}_{m_{\text{num}}}) \in \mathbb{R}^{m_{\text{num}}}$ and $\widehat{\boldsymbol{r}}_{\text{cat},j} \in \mathbb{R}^{d_j}$. For categorical column j, the target is its clean embedding $\phi_j(x_j)$.

2. Logit. Alternatively, the denoiser predicts one vector of logits per categorical column,

$$f_\theta(\boldsymbol{r}_t, t) = (\widehat{\boldsymbol{r}}_{\text{num}}, \boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{m_{\text{cat}}}) \,, \qquad \boldsymbol{\ell}_j \in \mathbb{R}^{C_j} .$$

The logits induce the posterior mean embedding

$$\widehat{\boldsymbol{r}}_{\text{cat},j} = \sum_{c=1}^{C_j} \widehat{p}_{j,c} \phi_j(c) \,, \qquad \widehat{p}_{j,c} = \frac{\exp(\ell_{j,c})}{\sum_{c'=1}^{C_j} \exp(\ell_{j,c'})} . \tag{4}$$

Loss function and training. For numerical columns, we use mean-squared error: $\ell_{\text{num}} := \frac{1}{m_{\text{num}}} \|\widehat{\boldsymbol{r}}_{\text{num}} - \boldsymbol{r}_{\text{num}}\|_2^2$. For categorical columns, we consider two choices, Square loss in embedding space, or cross-entropy loss on logits:

$$\ell_{\text{cat}}^{\text{MSE}} := \frac{1}{\sum_{j=1}^{m_{\text{cat}}} d_j} \sum_{j=1}^{m_{\text{cat}}} \|\widehat{\boldsymbol{r}}_{\text{cat},j} - \phi_j(x_j)\|_2^2 \,, \qquad \ell_{\text{cat}}^{\text{CE}} = \frac{1}{m_{\text{cat}}} \sum_{j=1}^{m_{\text{cat}}} \log \frac{1}{\widehat{p}_{j,x_j}} . \tag{5}$$

The total loss function is then $\ell = \ell_{\text{num}} + \ell_{\text{cat}}$. The optimizer then backpropagates the loss through the network and, if applicable, the learnt embedding lookup tables E_j.

Generation. Sampling is carried out entirely in the continuous representation space: obtain decreasing noise levels $\sigma_{t_0} > \sigma_{t_1} > \cdots > \sigma_{t_T}$ via the noise schedule and discretization of $[0, 1]$, and initialize $\boldsymbol{r}^{(0)} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{t_0}^2 I_D)$. Using the denoiser prediction, we construct the clean representation $\widehat{\boldsymbol{r}}^{(i)} := \widehat{\boldsymbol{r}}(\boldsymbol{r}^{(i)}, t_i)$ and perform the Euler update (Karras et al., 2022)

$$\boldsymbol{r}^{(i+1)} = \boldsymbol{r}^{(i)} + (\sigma_{t_{i+1}} - \sigma_{t_i}) \frac{\boldsymbol{r}^{(i)} - \widehat{\boldsymbol{r}}^{(i)}}{\sigma_{t_i}}, \qquad i = 0, \ldots, T-1. \tag{6}$$

Numerical coordinates are returned directly after the final step. For categorical column j, representation mode returns $\widehat{x}_j = \arg\min_{c \in \{1,\ldots,C_j\}} \|\widehat{\boldsymbol{r}}_{\text{cat},j} - \phi_j(c)\|_2$, while logit mode returns $\widehat{x}_j = \arg\max_{c \in \{1,\ldots,C_j\}} \widehat{p}_{j,c}$.

Empirical results. We note that the parameterization and the loss function together create three possible architecture variants: representation with MSE loss, logits with MSE loss, and logits with CE loss. The empirical results in Table 2 demonstrate the superiority of logits parametrization with CE loss; furthermore, they demonstrate that learnt embeddings enhance the performance for this model vis-a-vis the one-hot embedding.

4 OPTIMAL LOSS ON CATEGORICALS: ASYMPTOTIC THEORY

The superiority of CE loss over MSE loss is surprising: (i) At infinite sample size, minimizing MSE loss in denoising diffusions is equivalent to minimizing the KL divergence between the generated and target distribution (Song et al., 2021a; Chen et al., 2022); (ii) It was proven by Chewi et al. (2026) that —in a parametric setting— MSE loss results in the minimum parameter estimation error, under the classical asymptotics in which the sample size n diverges with the number of parameters fixed.

Our main result is that these conclusions should be revised if the denoisers learnt at different noise levels are not tied in the Bayes-optimal way (as is usually the case for neural denoisers). In other words a somewhat minimal form of misspecification leads to MSE becoming suboptimal.

For this analysis we consider a simple categorical variable taking k values, with probabilities $q(\sigma)$, $\sigma \in \{0, 1, \dots, q-1\}$. This variable is represented via embeddings $\boldsymbol{x}(\sigma) \in \mathbb{R}^d$, $\|\boldsymbol{x}(\sigma)\| = 1$. We learn a general diffusion indexed by the continuous time variable which increases from $t = 0$ (pure noise), to $t = T$ (sample from the target distribution):

$$\boldsymbol{X}_t = a(t)\boldsymbol{X} + \tau_0(t)\boldsymbol{B}_{s(t)}\,. \tag{7}$$

Here $\boldsymbol{X} \sim \sum_\sigma q(\sigma)\delta_{\boldsymbol{x}(\sigma)}$ is distributed according to the target, independent of the Brownian motion $\boldsymbol{B}$, $a(\,\cdot\,)$ increases from 0 to 1, while τ_0 goes from $\tau_0(0)$ to $\tau_0(T) = 0$. As usual, the process equation 7 is equivalently defined by an SDE of the form $\mathrm{d}\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)\,\mathrm{d}t + g(t)^{1/2}\,\mathrm{d}\boldsymbol{W}_t$ (for suitable coefficients $\boldsymbol{f}$ and g); one-time marginals are reproduced by the probability flow ODE $\mathrm{d}\boldsymbol{Z}_t = \boldsymbol{h}(\boldsymbol{Z}_t, t)\,\mathrm{d}t$ (for suitable $\boldsymbol{h}$). As standard in practice, we use probability flow for generation.

We are given i.i.d. data triples $\{(t_i, \sigma_i, \boldsymbol{Y}_i) : i \le n\}$, $\sigma_i \sim q$, $t_i \sim \lambda(t)\cdot \mathrm{d}t$, $\boldsymbol{Y}_i|_{(t_i,\sigma_i)} \sim \mathsf{N}\big(\boldsymbol{x}(\sigma_i)\,a(t_i), \tau(t_i)^2\boldsymbol{I}_d\big)$ and learn logits $\boldsymbol{\theta}_\sigma(t) = (\theta_{\sigma,0}(t), \boldsymbol{\theta}^+_\sigma(t)) \in \mathbb{R}^{d+1}$, $\sigma \in \{1, \dots, k-1\}$, parametrizng the posterior probabilities $\hat{q}_0(\boldsymbol{Y}; \boldsymbol{\theta}) \propto 1$, $\hat{q}_\sigma(\boldsymbol{Y}; \boldsymbol{\theta}) \propto \exp(\theta_{\sigma,0}(t) + \langle \boldsymbol{Y}, \boldsymbol{\theta}^+_\sigma(t)\rangle)$.

Crucially, the parameters $\boldsymbol{\theta}(t) = (\boldsymbol{\theta}_1(t), \dots, \boldsymbol{\theta}_{k-1}(t))$ are allowed to vary with t, as is usually the case for denoising diffusions, where a suitable encoding of t is used as input to the denoiser.

In our theoretical analysis, we consider a simplified setting whereby $\boldsymbol{\theta}(t)$ is learned via the following weighted empirical risk minimization (ERM):

$$\hat{\boldsymbol{\theta}}_b(t) := \arg\min_{\boldsymbol{\theta}} \hat{R}_{b,t}(\boldsymbol{\theta}), \qquad \hat{R}_{b,t}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} w_b(t_i - t)\,\ell_{\sigma_i}(\boldsymbol{Y}_i; \boldsymbol{\theta})\,. \tag{8}$$

where $w_b(t) = w_1(t/b)$ is an unnormalized weight function and b is a bandwidth parameter and $t \mapsto K(t) \subseteq E$ specifies the time-dependent optimization domain.

Denote by $\hat{q}_b(\sigma)$ the probability of generating class σ via the probability flow ODE, using the estimated parameters $\hat{\boldsymbol{\theta}}_b(t)$. We prove a central limit theory asymptotics for $\sqrt{n}(\hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}_*(t))$ (where $\boldsymbol{\theta}_*(t)$ are the 'optimal' parameters) and, as a consequence, for $\sqrt{n}(\hat{q}_b - q)$. Further, we prove that the limiting variance of the error is minimized for CE loss.

Theorem 1 (Informal)**.** *Under regularity assumptions, further assume that, as $n \to \infty$, $n^{1/4}b_n \to 0$ and $\sqrt{n}\,b_n/\log(1/b_n) \to \infty$. Then*

$$\sqrt{n}(\hat{q}_n - q) \Rightarrow \mathsf{N}_k(0, \boldsymbol{V})\,, \tag{9}$$

where the covariance matrix $\boldsymbol{V} \in \mathbb{R}^{k\times k}$ is given explicitly in terms of the loss function ℓ.

Further $\boldsymbol{V}$ is minimized for CE loss.

A formal statement, with full specification of the assumptions, is given in Appendix J. Note that we assume an undersmoothed regime whereby $b \ll 1/\sqrt{n}$. Generalizations are possible but calculations are less transparent, and we defer them to a further publication.

While there is some subtleties in the actual proof, the intuition for this results is quite straightforward. The variance $\boldsymbol{V}$ takes the form of an integral over t of a term depending on the estimation error at t: this is contrast to earlier work (Chewi et al., 2026) where the denoisers at different t are carefully tied. By classical optimality of the maximum likelihood estimator (Van der Vaart, 2000), CE loss is optimal (in asymptotic sense) at any given t, and hence is overall optimal.

5 CONDITIONAL GENERATION

We study conditional generation using DiffLTM with two approaches: Classifier-free guidance (CFG) (Ho & Salimans, 2022), and Sequential Monte Carlo (SMC) (Doucet et al., 2001). We extend CFG to range and linear conditions, and compare them with baselines in Figure 1.

5.1 TRAIN-TIME INTERVENTION: CLASSIFIER-FREE GUIDANCE

Denoiser. In addition to $\boldsymbol{r}_t$, the conditional denoiser takes as input the condition $\mathcal{C}$ and outputs $f_\theta(\boldsymbol{r}_t, \mathcal{C}, t)$ to generate rows satisfying condition $\mathcal{C}$. As anticipated in the introduction, we consider three types of conditions: (i) *equality conditions*: for a categorical column i, these take the form $x_i = a$; (ii) *range conditions*, for a numerical column i, $x_i \in [L_i, U_i]$; and (iii) *linear conditions*, which constrain a linear function of the numerical columns $\sum_i w_i x_i \in [L, U]$. Multiple of these conditions can be imposed simultaneously.

The denoiser takes as input a tensor that specifies, for each column, the noisy values as well as a specification of the various conditions that involve that column (e.g. the weight w_i for a linear constraint). We also input the value of the linear constraint at the current state.

Each condition type is accompanied by a Boolean mask, indicating whether the corresponding condition is present. Whenever an equality condition is specified for column i, we replace the noisy state $r_{t,i}$ with a special token τ_i. For numerical columns, τ_i is a learned scalar, whereas for categorical columns it is a learned embedding. Beyond these modification to the input representation, the remainder of the architecture is identical to that used for unconditional generation.

Training. During CFG training (Ho & Salimans, 2022), we need to sample a condition $\mathcal{C}$. We proceed as follows: we draw a masking probability $p_{\text{mask}} \sim \text{Unif}(0, 1)$ and independently select each column for conditioning with probability p_{mask}. Selected categorical columns are always assigned equality conditions. For each selected numerical column, the equality condition is replaced by a range condition with probability p_{rng}. To construct a linear condition, we sample the number of participating numerical columns, m_{lin}, uniformly from $0, \ldots, m_{\text{num}}$ and then select m_{lin} numerical columns. With probability p_{lin}, we impose a linear condition over these selected columns and remove any equality or range conditions previously assigned to them.

Equality conditions are constructed by setting the conditioned value of each selected column i to its value x_i in the corresponding training row. For numerical range conditions, we must ensure that the optimal conditional denoiser targets the appropriate truncated distribution, $p_{\text{target}}(\boldsymbol{x}) \propto p(\boldsymbol{x})\, \mathbb{1}\{L_i < x_i < U_i\}$. Specifically, we first sample an interval width $\Delta \sim \text{Unif}(\Delta_{\min}, \Delta_{\max})$, then sample the lower bound as $L_i \sim \text{Unif}(x_i - \Delta, x_i)$, and set $U_i = L_i + \Delta$. For a linear condition involving m_{lin} numerical columns, we sample the corresponding weight vector as $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, I_{m_{\text{lin}}})$. We then construct the bounds L_b and U_b using the same interval-width sampling procedure, centered around the value of the linear function $\boldsymbol{w}^\intercal \boldsymbol{x}$. Finally, we normalize the linear condition by $|\boldsymbol{w}|_2$, scaling both the weights w_i and the bounds L_b and U_b accordingly.

5.2 INFERENCE-TIME METHODS

We experiment with two inference-time methods: rejection sampling and sequential Monte Carlo (SMC). Rejection sampling is unbiased but highly inefficient for rare conditions, so we use it as a reference (***Rejection*** in Figure 1).

SMC is a class of methods widely used across statistics and signal processing to perform posterior sampling in sequential systems (Doucet et al., 2001), which was recently used in conjunction with diffusion models (Wu et al., 2023). At a high level, SMC entails initializing K reverse-diffusion trajectories and reweighting them in accordance with how likely each trajectory is to eventually produce a synthetic row satisfying the requested condition (see Algorithm 3 in Appendix I for details)—the main algorithmic question that then arises is how to approximate this probability. To this end, we make three approximations: (i) columns are conditionally independent given the current trajectory state; (ii) categorical columns follow the distribution induced by the predicted denoiser logits; and (iii) clean numerical columns follow a Gaussian distribution centered at the denoiser prediction. We note that choosing logit mode enables assumption (ii), adding to the previously outlined reasons regarding the theoretical and practical optimality of this choice. SMC enables conditioning

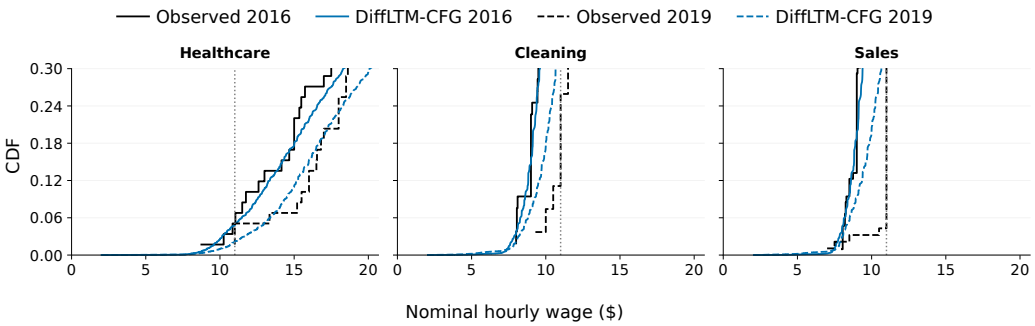


Figure 3: The counterfactual effect of Proposition 206 tracks each occupation's exposure to the minimum wage. The gap between dashed curves (i.e. true 2019 values and DiffLTM predicted counterfactual) is large for cleaning and sales workers, whose observed wages pile up at $11, and negligible for healthcare workers, whose wages already sat above it (see Appendix E.4 for details).

on categorical predicates as well as arbitrary numerical conditions, a capability not offered by prior work such as *Harpoon* (Shankar et al., 2026).

5.3 Reproducing econometric patterns via conditional generation

We use two important econometric ideas and test whether our model can capture the behavioral patterns captured by them: counterfactual analysis and Mincer equation on returns to education. Due to space limitations, we focus on the former, deferring the latter to Appendix G. Results presented in this section use the same denoiser architecture as in previous ones, with additional adaptive layer norm (AdaLN) Peebles & Xie (2023). AdaLN did not have measurable impact on other experiments.

Counterfactual analysis. An entity s_* receives a policy intervention at time period t_*, which is thought to have an effect on a certain outcome $\overline{Y}(s_*, t)$ at times $t \geq t_*$. What would the outcome have been if the intervention had not happened?

As a case study, we will consider the effect of minimum wage raise. Proposition 206, approved by Arizona voters in November 2016, raised the state minimum wage from $8.05 to $10.00 on January 1 2017, followed by increases to $10.50 in 2018, $11.00 in 2019, and $12.00 in 2020 (Ballotpedia). What would have been wages in Arizona had these measures not passed?

Synthetic control (Abadie et al., 2010; 2015; Abadie, 2021) finds non-negative weights w_s summing to one such that $\overline{Y}(s_*, t) \approx \sum_{s \neq s_*} w_s \, \overline{Y}(s, t)$ for $t < t_*$. Assuming the treatment does not affect other units, the same combination serves as the counterfactual for $t \geq t_*$. In our case s_* is Arizona and $\overline{Y}(s, t)$ is the 10th-percentile wage in state s.

We formalize this problem as *conditional generation for an unseen subpopulation*. We view each record as a sample from a joint distribution over unit S (e.g., state), time T, covariates and outcome Y. The training data contain the treated unit s_* only before t_*, and the post-intervention period $t \geq t_*$ only for the other units, so the combination $(S = s_*, T \geq t_*)$ is never observed (Arizona in 2017–2019, shaded in Figure 2). We train DiffLTM on these data and generate the missing subpopulation by conditioning on $S = s_*$ and $T \geq t_*$. This has two advantages: (i) unlike a single summary statistic per time step, it estimates the full counterfactual distribution, including individual-level effects; and (ii) it is robust to reparametrizations of the outcome.

Data and method. In order to investigate the effect of Proposition 206, we utilize wage records from the 2010–2019 Current Population Survey (CPS) data, made available by the NBER (National Bureau of Economic Research). After filters (see Appendix E.1), the dataset comprises 489,225 rows corresponding to employment records from all 50 states and DC. As a baseline, we construct a counterfactual following the methodology of Abadie et al. (2010). The resulting synthetic Arizona places 47.7% of its weight on Virginia, 25.2% on Texas, 14.6% on North Dakota, 11.2% on

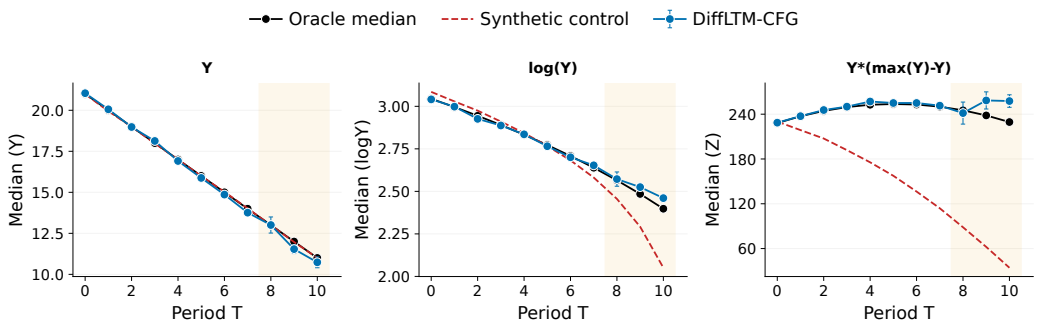


Figure 4: DiffLTM is robust to reparametrizing the outcome whereas synthetic control is not. Median of $f(Y)$ for a synthetic-panel state held out in the shaded periods, for $f(y) = y$, $\log y$, $y(\max y - y)$. Both match the oracle for the additive $f(y) = y$; under nonlinear f, synthetic control's linear fit breaks down while DiffLTM-CFG stays close (Section 5.3; Appendix F).

Wyoming, and 1.3% on South Carolina. Our approach consists in training DiffLTM after removing Arizona data from 2017 and after; see Appendix E for details.

Results. Figure 2 shows that Arizona's observed tenth-percentile nominal hourly wage follows the statutory minimum almost exactly. We compare this with the counterfactual prediction provided by synthetic control and DiffLTM. Obviously there is no way of determining 'ground truth' in this case, but the conditional generation approach is in rough agreement with synthetic control and provides a natural prediction. We repeat the procedure for Pennsylvania, which remained at the federal minimum throughout. Consistent with the design, the observed post-2016 wages match very closely with the one predicted by DiffLTM. We observe that in factor our prediction is somewhat more accurate than the one from synthetic control.

Individual-level effects. Synthetic control produces one counterfactual summary statistic per year. Conditional generation instead samples from the *joint* counterfactual distribution and produces individual-level records. By further conditioning on demographic/occupational we can assess which subpopulations are more affected. As an example, in Fig. 3 we stratify based on occupation (see Appendix E.4 for full analysis and Fig. 6 for stratification based on education.)

Robustness to reparameterization. A potential concern with synthetic control is that the estimate heavily relies on the assumption of a linear relationship between outcomes $Y(s,t)$ for different units. In particular, this assumption is fragile under reparametrizations of the underlying variable. Consider —as in application example— the case in which $\overline{Y}(s,t)$ is the α quantile of a variable of interest Y conditional on s,t, and define $X = f(Y)$ for a nonlinear function f, and the corresponding quantile $\overline{X}(s,t)$. If the linear relationship $\overline{Y}(s_*,t) \approx \sum_{s\neq s_*} w_s^Y \overline{Y}(s,t)$ holds approximately then, generically, $\overline{X}(s_*,t) \approx \sum_{s\neq s_*} w_s^X \overline{X}(s,t)$ will not hold. Figure 4 demonstrates in a synthetic example that conditional generation is robust to reparametrizations, while synthetic control is not.

6 Discussion and Limitations

We presented the design of a generative model for tabular data, based on continuous diffusions. The model has state of the art capabilities in generation and conditional generation. An important methodological element is the use of learned embeddings and CE loss for categoricals, for which we provide a mathematical rationale. Finally, we trained our model on datasets that are one order of magnitude larger than previous published work, and demonstrated the ability to recover econometric patterns, providing complementary information to earlier validations.

An important missing element is cross-datasets modeling. In preliminary work, we trained a cross-datasets version of our architecture, but defer this to a future report. An important limitation to our econometrics application is that we did not develop rigorous statistical guard-rails for our approach. We consider it an important avenue for future work.

AI USE STATEMENT

We used AI assistants to polish writing, support literature review, and assist with code development. All references included in the paper were manually verified by the authors, and all AI-generated code was reviewed and tested for correctness. We did not use AI tools to write code for the core methodological components of the work. We additionally used generative AI tools for drafting the proofs in Appendix J. The proofs were reviewed and edited by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We use either publicly available datasets or provide complete details of the datasets constructed in this work in Appendix B.1. Model architectures and hyperparameter settings are documented in Appendix C and Appendix D. All experiments are repeated across multiple random seeds, and we report means together with standard errors throughout the paper. We commit to releasing the complete codebase upon acceptance of the paper.

REFERENCES

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature*, 59(2):391–425, 2021.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pp. 290–306. PMLR, 2022.

Ballotpedia. Arizona minimum wage and paid time off, proposition 206 (2016). `https://ballotpedia.org/Arizona_Minimum_Wage_and_Paid_Time_Off,_Proposition_206_(2016)`. Accessed: 2026-09-13.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2022.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, and Omar Montasser. Ddpm score matching and distribution learning. In *The Thirty Ninth Annual Conference on Learning Theory*, pp. 1434–1435. PMLR, 2026.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer, 2001.

Floor Eijkelboom, Grigory Bartosh, Christian A Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024.

Andrés Guzmán-Cordero, Floor Eijkelboom, and Jan-Willem Van De Meent. Exponential family variational flow matching for tabular data generation. *arXiv preprint arXiv:2506.05940*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Valter Hudovernik, Minkai Xu, Juntong Shi, Lovro Šubelj, Stefano Ermon, Erik Štrumbelj, and Jure Leskovec. Reldiff: Relational data generative modeling with graph-based diffusion models. *arXiv preprint arXiv:2506.00710*, 2025.

Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. How well does your tabular generator learn the structure of tabular data? *arXiv preprint arXiv:2503.09453*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Mohamed Amine Ketata, David Lüdke, Leo Schwinn, and Stephan Günnemann. Joint relational database generation via graph-conditional diffusion models. *Advances in Neural Information Processing Systems*, 38:163644–163672, 2026.

Thomas Lemieux. The "mincer equation" thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics*, pp. 127–145. Springer, 2006.

Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. Goggle: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2023.

Jacob A. Mincer. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, 1974.

Markus Mueller, Kathrin Gruber, and Dennis Fok. Continuous diffusion for mixed-type tabular data. In *International Conference on Learning Representations*, volume 2025, pp. 39186–39225, 2025.

Markus Mueller, Kathrin Gruber, and Dennis Fok. Cascaded flow matching for heterogeneous tabular data with mixed-type features. *arXiv preprint arXiv:2601.22816*, 2026.

Nazia Nafis, Inaki Esnaola, Alvaro Martinez-Perez, Maria-Cruz Villa-Uriol, and Venet Osmani. Critical challenges and guidelines in evaluating synthetic tabular data: A systematic review. *arXiv preprint arXiv:2504.18544*, 2025.

National Bureau of Economic Research. Cps merged outgoing rotation groups: Annual data. `https://data.nber.org/morg/annual/`. Accessed: 2026-09-13.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Davide Scassola, Sebastiano Saccani, and Luca Bortolussi. Graph-conditional flow matching for relational data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25209–25217, 2026.

SDMetrics. *Synthetic Data Metrics*. DataCebo, Inc., 10 2023. URL `https://docs.sdv.dev/sdmetrics/`. Version 0.12.0.

Aditya Shankar, Yuandou Wang, Rihan Hai, and Lydia Chen. Harpoon: Generalised manifold guidance for conditional tabular diffusion. In *International Conference on Learning Representations*, volume 2026, pp. 143083–143109, 2026.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: a mixed-type diffusion model for tabular data generation. In *International Conference on Learning Representations*, volume 2025, pp. 37353–37375, 2025.

Jacob Si, Zijing Ou, Mike Qu, Zhengrui Xiang, and Yingzhen Li. Tabrep: training tabular diffusion models with a simple and effective continuous representation. *arXiv preprint arXiv:2504.04798*, 2025.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

U.S. Census Bureau. 2014–2018 american community survey public use microdata sample (PUMS) data dictionary, 2018. URL https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2014-2018.csv. Accessed: 2026-09-14.

Boris Van Breugel and Mihaela Van Der Schaar. Position: why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 48976–48993, 2024.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Zhengchao Wan, Qingsong Wang, Gal Mishne, and Yusu Wang. Elucidating flow matching ode dynamics via data geometry and denoisers. In *Forty-second International Conference on Machine Learning*, 2025.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*, volume 2024, pp. 52829–52857, 2024.

APPENDIX

A Further related work **14**

B Datasets and metrics **14**

B.1 Datasets details 14

B.2 Metrics definitions 14

C Unconditional generation: Experiment details **16**

C.1 Architecture parameters 16

C.2 Training 16

C.3 Baselines implementation 16

C.4 Metrics computation and confidence intervals 17

D Conditional generation: Experiment details **17**

D.1 CFG 17

D.2 Inference-time methods 18

D.3 Evaluation methodology 18

D.4 Baseline implementation 18

E Arizona proposition 206 experiment details **18**

E.1 Dataset 18

E.2 Synthetic control counterfactual 19

E.3 DiffLTM counterfactual 19

E.4 Individual-level analyses 20

F Reparametrization in synthetic control details **21**

F.1 Data distribution 21

F.2 Training and generation 21

G Mincer returns to education **21**

H Additional empirical results **22**

I Pseudocode **23**

J Asymptotic analysis of discrete variables generation **25**

J.1 Setting and assumptions 25

J.2 Asymptotics of parameter estimates 32

J.3 Undersmoothing regime 34

J.4 Choice of the time-sampling density 35

J.5 Proofs for Sections J.1 and J.2 36

A Further related work

The literature on generative models for tabular data is rapidly growing. This section provides a few pointers beyond the ones mentioned in context in the main text.

Evaluation is well recognized to be an outstanding challenge because of the heterogeneity of tabular data (Nafis et al., 2025). Metrics based on statistical accuracy (Alaa et al., 2022), downstream utility of synthetic data (Xu et al., 2019), latent causal relationships (Jiang et al., 2025) have been developed.

A variety of techniques and ideas have been brought to bear to develop generative models for tabular data. Among others, generative adversarial networks (Xu et al., 2019), graph neural networks (Liu et al., 2023), based on autoregressive models Borisov et al. (2022), latent space diffusions (Zhang et al., 2024), variational flow matching Guzmán-Cordero et al. (2025).

A small number of recent papers develop models for relational databases beyond the single-table setting Hudovernik et al. (2025); Ketata et al. (2026); Scassola et al. (2026).

B Datasets and metrics

B.1 Datasets details

Construction of ACSLabour. We first provide a detailed description of the contruction of the ACSLabour dataset. As mentioned in Section 2.1, all datasets used herein are derived from the ACS PUMS. The ACS is a survey administered by the U.S. Census Bureau, while PUMS provides de-identified person-level survey records containing demographic, geographic, educational, employment, and income information. Folktables (Ding et al., 2021) constructs standardized machine-learning datasets from ACS PUMS by specifying a population filter, a set of retained columns, and a prediction target and two of our utilized datasets (ACSPubCov and ACSMobility) are derived from this work, concatenated over the 5-year survey period $2014 - 2018$.

In addition to these, we construct *ACSLabour*, a dataset derived from the ACS PUMS which proves particularly useful for econometric experiments considered in this work. The existing folktables dataset ACSIncome discretizes income into a binary prediction target whereas ACSLabour retains continuous log wage, and is designed to build predictive models for this outcome. This choice enables analyses such as regressions on the continuous log wage (see AppendixG). We also retain a broader set of labor and demographic columns, including high-cardinality occupation, industry, geographic, and birthplace variables a detailed list of which is provided in Table 3. ACSLabour therefore is both a regression benchmark and a stress test of whether the generative model can recover rare, semantically meaningful subpopulations within a large table.

From the full $2014 - 2018$ PUMS dataset, for ACSLabour we retain only those rows satisfying:

- `AGEP` ≥ 16;
- `ESR` $\in \{1, 2\}$, corresponding to civilians employed either at work or temporarily absent from work;
- `WAGP` is observed and strictly positive; and
- `WKHP` > 0.

Datasets summary. Table 4 summarizes the size, column types, and designated prediction task for each of the three datasets utilized in our experiments.

Preprocessing. All preprocessing transformations are fit on the training partition and then applied unchanged to the validation and test partitions. Numerical columns are quantile transformed, while categorical columns are integer-encoded.

B.2 Metrics definitions

We provide formal definitions of metrics deferred from Section 2.2 as well as introduce a new metric, *Conditional Shape (rare)*.

Column	Type	Description
AGEP	Numerical	Age in completed years.
WKHP	Numerical	Usual hours worked per week during the previous 12 months.
log_wage	Numerical	Regression target: natural logarithm of annual wage and salary income in constant 2018 dollars, computed as $\log(\texttt{WAGP} \times \texttt{ADJINC}/10^6)$.
ST	Categorical	State or territory of residence.
PUMA	Categorical	Public Use Microdata Area of residence; codes are unique only within state.
SEX	Categorical	Sex.
RAC1P	Categorical	Recoded detailed race.
HISP	Categorical	Detailed Hispanic-origin recode.
SCHL	Categorical	Highest level of educational attainment.
POBP	Categorical	Place of birth.
CIT	Categorical	Citizenship status.
ENG	Categorical	Self-reported ability to speak English.
OCCP	Categorical	Occupation code.
INDP	Categorical	Industry code.
COW	Categorical	Class of worker.
MAR	Categorical	Marital status.
HICOV	Categorical	Health-insurance coverage status.
WKW	Categorical	Interval-coded number of weeks worked during the previous 12 months.

Table 3: Modeled columns in ACSLabour. Categorical variables retain their original ACS PUMS codes; code definitions and semantic labels are provided in the ACS 2014–2018 PUMS Data Dictionary U.S. Census Bureau (2018).

Dataset	Rows	Numerical	Categorical	Target	Task
ACSLabour	6,892,317	3	15	log_wage	Regression
ACSPubCov (Ding et al., 2021)	5,916,565	2	19	PUBCOV	Classification
ACSMobility (Ding et al., 2021)	1,345,437	4	17	MIG	Classification

Table 4: Summary of the ACS-derived datasets

Shape. Shape assesses fidelity of the univariate marginal distributions. For numerical column j, let $S_j = 1 - \mathrm{KS}(\widehat{F}_j^{\mathrm{real}}, \widehat{F}_j^{\mathrm{syn}})$, where KS denotes the Kolmogorov–Smirnov distance and $\widehat{F}_j$ is the empirical CDF of column j in the real or synthetic data. For categorical column j, let $S_j = 1 - D_{\mathrm{TV}}(\widehat{p}_j^{\mathrm{real}}, \widehat{p}_j^{\mathrm{syn}})$, where D_{TV} is the total variation distance and $\widehat{p}_j$ is the empirical distribution of column j. We report

$$\mathrm{Shape} := \frac{1}{m} \sum_{j=1}^{m} S_j. \tag{10}$$

Trend. Trend is the pairwise analogue of Shape, assessing fidelity of pairwise (bivariate) relationships. For each pair of columns (j, ℓ), if both columns are numerical, let $T_{j\ell} = 1 - \frac{1}{2}|\widehat{\rho}_{j\ell}^{\mathrm{real}} - \widehat{\rho}_{j\ell}^{\mathrm{syn}}|$, where $\widehat{\rho}_{j\ell}$ is the empirical Pearson correlation. Otherwise, let $T_{j\ell} = 1 - D_{\mathrm{TV}}(\widehat{p}_{j\ell}^{\mathrm{real}}, \widehat{p}_{j\ell}^{\mathrm{syn}})$, where $\widehat{p}_{j\ell}$ is the empirical joint distribution of the two columns, with numerical columns discretized when necessary. We report

$$\mathrm{Trend} := \frac{2}{m(m-1)} \sum_{j<\ell} T_{j\ell}. \tag{11}$$

Conditional Shape (Rare). For a fixed $\tau \in (0, 1)$, we denote by $\mathcal{A}_j^{\mathrm{rare}} \subseteq \mathcal{A}_j$ the subset of values that appear in column j with frequency below τ (by default $\tau = 0.05$). We define

$$\mathrm{CShape}_{\mathrm{rare}} := \frac{1}{m} \sum_{j=1}^{m} \frac{1}{|\mathcal{A}_j^{\mathrm{rare}}|} \sum_{a \in \mathcal{A}_j^{\mathrm{rare}}} S(j, a). \tag{12}$$

This metric emphasizes even more accuracy on rare sub-populations. We note that Conditional Trend and Conditional Trend (Rare) can be defined analogously to the above.

MLE Ratio. We train an XGBoost model on synthetic data and evaluate it on held-out real data (*Train Synthetic, Test Real*; TSTR). We divide this score by that of an identically configured XGBoost model trained on real data and evaluated on the same held-out set (*Train Real, Test Real*; TRTR). Then,

$$\text{MLE Ratio} := \frac{\text{Score}_{\text{TSTR}}}{\text{Score}_{\text{TRTR}}}. \tag{13}$$

We use AUC for classification and R^2 for regression; values near one indicate that the synthetic data preserve downstream predictive utility.

C Unconditional generation: Experiment details

In this section, we provide a full description of the experiment setup leading to the results reported in Tables 1 and 2. We recall that the model obtained by the architecture/training choices of (i) learnt embeddings, (ii) logits mode and (iii) CE loss is utilized in Table 1 to compare against baselines.

C.1 Architecture parameters

When used, the learned embedding matrix is implemented using PyTorch's `nn.Embedding`. In all such experiments we set the embedding dimension to $d_{\text{emb}} = 16$. Numerical columns are transformed using a quantile transformation. The concatenated noisy representation $\boldsymbol{r}_t$ is first linearly projected to a hidden width of $d_h = 1024$, which is subsequently processed by a shared MLP comprising $L = 5$ fully connected hidden layers of width $d_h = 1024$ and using SiLU activations. The diffusion time t is encoded using a 1024-dimensional sinusoidal embedding, which is further processed by two fully connected layers of width 1024 with SiLU activations. This time embedding is added to the projected input representation before it is passed through the shared MLP. On the ACSLabour dataset, the unconditional model contains approximately 9.84M parameters, including the learned embedding tables.

C.2 Training

All the DiffLTM variants are trained for 2000 epochs with a batch size 8192 and AdamW ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We use an initial learning rate of 3×10^{-4} with a cosine decay schedule. Following earlier work (Song et al., 2021b), we use exponential moving average(EMA) of parameters with decay of 0.997 for evaluation. The final checkpoint is selected using lowest validation loss.

All experiments were run on a single node with 192 CPU cores and 8 NVIDIA A100 80 GB GPUs, using Ubuntu 22.04 and CUDA 13.0. We train each model using three different random seeds and use bfloat16 (bf16) precision for all training runs. Each individual run, including both training and evaluation, takes less than 12 GPU-hours.

C.3 Baselines implementation

We benchmark DiffLTM against two recent state-of-the-art generative models for tabular data that also use continuous diffusion: TabRep (Si et al., 2025) and TabCascade (Mueller et al., 2026).

To ensure a fair comparison, we use the official implementations released by the authors:

- For TabRep, we use the flow-matching variant, TabRep-Flow, available at `https://github.com/jacobyhsi/TabRep`.
- For TabCascade, we use the implementation available at `https://github.com/muellermarkus/tabcascade`.

Importantly, we use the exact same training, validation, and test splits as for DiffLTM. We train each model end-to-end using its official implementation and generate a synthetic dataset containing n_{train} rows, matching the size of the real training set. Regardless of the generation method, all synthetic datasets are evaluated using the same standardized evaluation setup described in the subsequent Appendix C.4.

We note that the default training budgets used in the official implementations are substantially smaller, potentially due to their evaluation on smaller datasets. To match the training budget, we therefore train all models for 2000 epochs with batch size of 8192 and select the checkpoint with the lowest validation loss.

C.4 Metrics computation and confidence intervals

Upon training the model, we generate n_{train} (i.e. size of the training data) synthetic rows for each method. To compute the fidelity metrics (Shape, Trend, and Conditional Shape), we subsample this to n_{test} synthetic rows and compare them against the held-out real test set.

The MLE metrics involve randomness in fitting the downstream prediction model; we account for this by repeating both TRTR and TSTR evaluations 10 times. More precisely, in each round we sample n_{test} rows without replacement from the real training set for TRTR and from the synthetic pool for TSTR, fit a fresh XGBoost model, and evaluate it on the full real test set. We separately average the 10 TRTR and 10 TSTR scores and compute their ratio to obtain the final MLE score.

For all metrics, we report the mean and standard error across three training seeds.

D Conditional generation: Experiment details

We now provide a description of the experimental setup from Figure 1, as well as all the additional results on conditional generation in Appendix H.

D.1 CFG

The CFG denoiser architecture is described in Figure 5.

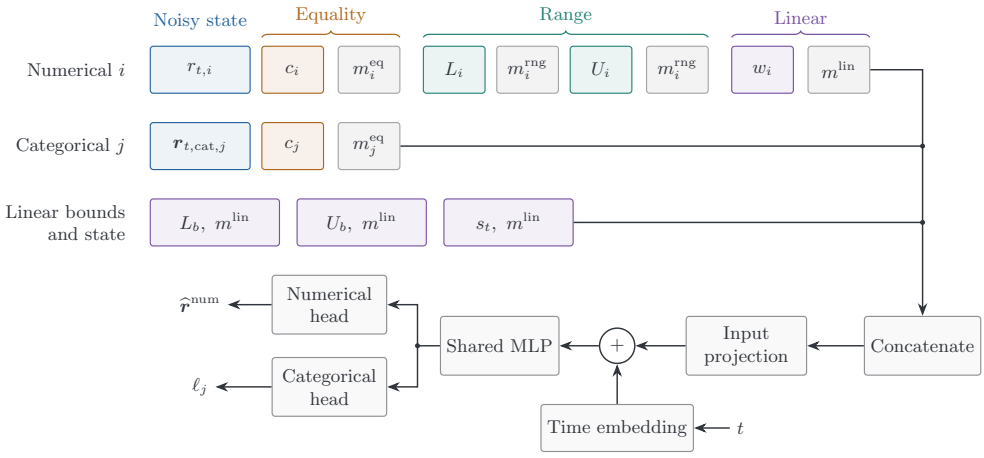


Figure 5: The DiffLTM conditional denoiser differs from the unconditional one only at the input. For each column, the noisy state is concatenated with the parameters of every condition involving that column (equality value, range bounds, or linear weight), each paired with a mask indicating whether the condition is present. A linear condition additionally contributes its bounds and its value s_t at the current noisy state. The concatenated input is then projected and processed by the same time-conditioned MLP and output heads as the unconditional model, and setting all masks to zero recovers the unconditional model (Section 5.1; Appendix D.1).

The training procedure is identical to that reported in Appendix C.2 with the following choices specific to the CFG variant of DiffLTM. We use CFG dropout probability $p_{\text{dropout}} = 0.15$, $p_{\text{rng}} = 0.5$, and $p_{\text{lin}} = 0.5$. For numerical range conditions, we use width limits $(w_{\min}, w_{\max}) = (0.02, 1.0)$, while for linear conditions we use $(\Delta_{\min}, \Delta_{\max}) = (0.02, 1.0)$. These limits are specified as frac-

tions of the corresponding variable or linear-function span, rather than as absolute widths. For generation, the CFG conditioning strength of $\omega = 1$ gave the best performance.

D.2 Inference-time methods

Rejection sampling and SMC can be used in conjunction with any unconditionally trained model—in our experiments, we utilize the same checkpoint used to obtain the results in Table 1, i.e. with CE loss in logit parameterization and using the learnt embeddings. The architecture and training details for this model are provided in Appendix C. For SMC sampling in Algorithm 3, we use an effective sample size (ESS) threshold of $\alpha = 0.95$ and a particle multiplier of $K = 1$. For rejection sampling, for each condition $\mathcal{C}$, we retain the first $\min(100{,}000, n^{\mathcal{C}}_{\text{test}})$ generated rows that satisfy the condition, where $n^{\mathcal{C}}_{\text{test}}$ denotes the number of test-set rows satisfying $\mathcal{C}$.

D.3 Evaluation methodology

We use a fixed set of 50 slice conditions $\mathcal{C}$ for each task (equality, numerical range, and linear conditioning) with each slice containing at least 100 rows in the test set. The same 50 slices are used for evaluating all methods. To construct an equality condition, we first sample $k \in 1, 2, 3, 4$, then select k categorical columns uniformly without replacement and assign their conditioned values using a uniformly sampled row. For numerical range conditions, we sample $k \in 1, 2, 3$ numerical columns uniformly without replacement. For each selected numerical column i, we sample the interval in quantile space by first drawing a quantile width $q_i \sim \text{Uniform}(0.002, 0.8)$ and then a lower quantile bound $a_i \sim \text{Unif}(0, 1 - q_i)$. Set the upper quantile bound to $a_i + q_i$. These quantile bounds are then mapped back to the original numerical space to get (L_i, H_i). The bounds for integer columns are rounded to the nearest integers. For linear conditions, we use the same distribution as during training. For all three tasks, we repeatedly sample candidate conditions until the resulting slice falls within the desired rarity range (like $1 - 3\%$) in the training set and contains at least 100 rows in the test set.

For each slice condition $\mathcal{C}$, we evaluate $\min(100{,}000, n^{\mathcal{C}}_{\text{test}})$ synthetic rows against the same number of real rows from the test set. When computing the fidelity metrics, we retain all generated synthetic rows, including those that do not satisfy $\mathcal{C}$, for every generation method. To compute the empirical ideal score for $\mathcal{C}$, we compare $n^{\mathcal{C}}_{\text{eval}}$ real test rows against an equal number of rows from the real training set. The rarity of $\mathcal{C}$ is defined by its frequency in the training set. For each condition $\mathcal{C}$, we compute the fidelity metrics using the three checkpoints obtained from the three independent training seeds. Each metric for each slice is averaged across these three runs. In the final fidelity plots, conditions are grouped into rarity buckets (e.g., 1–3%), and we report the mean and standard error of the scores across the slice conditions within each bucket.

D.4 Baseline implementation

For a fair comparison, we apply Harpoon (Shankar et al., 2026) guidance to the MSE-loss, representation-mode variant of DiffLTM. This allows controlled comparison of conditional generation capability in which the denoiser architecture and diffusion process are held fixed, with results presented in Figure 1 and Appendix H. We use the constant guidance scale of 0.2 as suggested by the authors. Since the Harpoon implementation does not include linear constraints, we do not evaluate it on the linear-conditioning task.

E Arizona proposition 206 experiment details

We provide details on the experimental setup to study effects of Proposition 206 (cf. Figures 2 and 6), as well as present additional results.

E.1 Dataset

We utilize wage records from the 2010–2019 Current Population Survey (CPS) data, made available by the NBER (National Bureau of Economic Research). We retain employed private-sector workers

aged 16–64 with self reported hourly wages between $\$2.00$ and $\$99.98$. The resulting dataset contains 489,225 rows corresponding to observations from all 50 states and DC. The retained columns are state, year, age, sex, education, race, broad occupation, and $\log$(hourly wage), with wages normalized to constant 2019 units using CPI-U-RS, a provided inflation measure.

E.2 Synthetic control counterfactual

Our statistic of interest is the tenth-percentile hourly wage each year; this is the region where the effect of a minimum wage policy shock should be most visible. The donor pool $\mathcal{D}$ consists of the 21 states that remained at the federal $\$7.25$ minimum wage throughout 2010–2019: AL, GA, ID, IN, IA, KS, KY, LA, MS, NH, NC, ND, OK, PA, SC, TN, TX, UT, VA, WI, and WY. To pick the synthetic control weights, we construct the following 14 features (all measured pre-2017) for each donor state and AZ:

- Fraction of workers aged 16–24
- Fraction of workers with at most a high school education
- Fraction of workers with at least a bachelor degree
- Fraction of workers in the following occupations:
  - Sales
  - Cleaning or personal care
  - Food preparation and serving
- Mean $\log$(hourly wage)
- 10th percentile wage in each pre-intervention year (2010–2016).

The population features chosen in addition to the pre-intervention 10th percentile value are those that make the individual more likely to be affected by a minimum wage increase.

Let $\boldsymbol{x}_{\mathrm{AZ}} \in \mathbb{R}^{14}$ denote Arizona's obtained feature vector and $\boldsymbol{x}_j$ the corresponding feature vector for donor state j (all standardized). We choose the weights by minimizing the mean-squared distance:

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\| \boldsymbol{x}_{\mathrm{AZ}} - \sum_{j \in \mathcal{D}} w_j \boldsymbol{x}_j \right\|_2^2 \quad \text{subject to} \quad w_j \geq 0, \qquad \sum_{j \in \mathcal{D}} w_j = 1. \tag{14}$$

The post-intervention synthetic-control outcome Y is then

$$\widehat{Y}_{\mathrm{AZ},t} = \sum_{j \in \mathcal{D}} \widehat{w}_j Y_{j,t}.$$

The exact analogous procedure is used to construct the synthetic-control counterfactual for PA.

E.3 DiffLTM counterfactual

From the dataset, we remove all rows ($2,539$ in total) satisfying

$$\textsc{State} = \mathrm{AZ} \quad \text{and} \quad \textsc{Year} \in \{2017, 2018, 2019\}$$

and train a CFG-enabled DiffLTM model with the architecture/training choices described in Appendix D.1. Following prior work (Peebles & Xie, 2023), we additionally use MLP blocks with adaptive layer normalization (AdaLN) (Perez et al., 2018). Specifically, the time embedding is used to predict per-layer scale and shift parameters, which are applied to the layer-normalized features before each linear layer.

At inference, we provide condition $\textsc{State} = \mathrm{AZ}$ and the requested post-treatment year and sample all remaining worker attributes and wages jointly. We generate $20,000$ synthetic rows for the slice corresponding to each year and calculate the 10th percentile value of the hourly wage (converted into nominal \$). We repeat generation with four random seeds; the mean obtained value is reported. The exact analogous procedure is used to construct the DiffLTM counterfactual for PA.

E.4 INDIVIDUAL-LEVEL ANALYSES

Synthetic control produces one counterfactual summary statistic per year (in the above case, the 10th percentile hourly wage value). Conditional generation instead samples from the *joint* counterfactual distribution and produces individual-level records. By further conditioning on other columns corresponding to demographic/occupational profiles and producing counterfactual distributions *per subpopulation*, we can assess which subpopulations are more affected by the policy.

As an example, we stratify based on education level (shown in Figure 6) and broad occupational description, since we expect certain education and occupation profiles to be more affected by a minimum-wage increase. Using the model trained with all post-intervention Arizona rows excised, we conditionally generate $20{,}000$ rows for each of the following slices:

- STATE $=$ AZ, YEAR $\in \{2016, 2019\}$, and EDU $\in \{\text{high school}, \text{bachelor's}\}$;
- STATE $=$ AZ, YEAR $\in \{2016, 2019\}$, and OCC $\in \{\text{healthcare}, \text{cleaning}, \text{sales}\}$.

The 2016 slices provide a calibration value because data available during training, whereas the 2019 slices correspond to the held-out, post-intervention period. We consider two metrics of policy exposure. First, we calculate the Kolmogorov–Smirnov (KS) distance between the real and synthetic wage distributions in each year and report its change from 2016 to 2019 for group g:

$$\Delta_{\mathrm{KS}}(g) = \mathrm{KS}\big(F^{\mathrm{real}}_{g,2019}, F^{\mathrm{syn}}_{g,2019}\big) - \mathrm{KS}\big(F^{\mathrm{real}}_{g,2016}, F^{\mathrm{syn}}_{g,2016}\big)\,.$$

If the 2019 distance is similar to its 2016 calibration value, this suggests that the policy shock for that subpopulation is small relative to the model's baseline approximation error. Second, because Arizona's minimum wage was \$11.00 on January 1, 2019, we calculate the *sub*-\$11 *reduction*:

$$\Delta_{<11}(g) = \mathsf{P}_{\mathrm{syn}}(\text{hourly wage} < \$11 \mid g, 2019) - \mathsf{P}_{\mathrm{real}}(\text{hourly wage} < \$11 \mid g, 2019).$$

This measures the fraction of the counterfactual population predicted to earn below \$11 that is absent from the observed post-policy distribution.

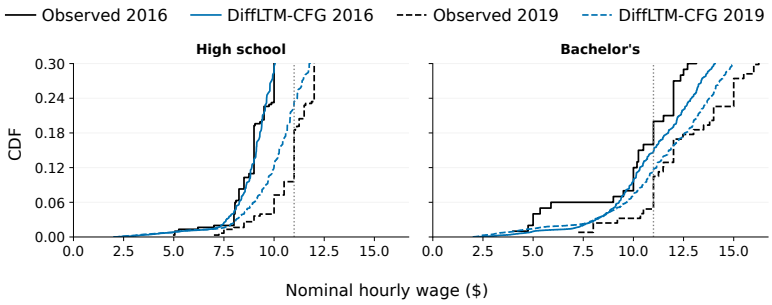


Figure 6: DiffLTM generates the full counterfactual wage distribution for any subgroup. Arizona wage CDFs (lowest 30%); 2016 seen in training, 2019 held out. Dotted: \$11 minimum set by Proposition 206. The policy effect, the gap between dashed curves, is large for high-school workers and small for bachelor's-degree workers relative to their 2016 gap (Section 5.3; Appendix E.4).

The KS changes behave as expected (see Figures 6 and 3). The distance increases by 0.073 for workers with a high-school education, compared with only 0.012 for workers with a bachelor's degree. Across occupations, it decreases by 0.001 for healthcare workers, but increases by 0.137 for cleaning workers and 0.224 for sales workers. The corresponding sub-\$11 reductions are 13.7 percentage points for high-school workers and 6.8 points for workers with a bachelor's degree. Across occupations, they are 2.2 points for healthcare, 25.8 points for cleaning, and 32.7 points for sales. We note that the observed data also exhibit substantial mass at exactly \$11, particularly among the more exposed groups.

F REPARAMETRIZATION IN SYNTHETIC CONTROL DETAILS

As alluded to in Section 5.3 and illustrated in Figure 4, synthetic control can be very brittle, in particular showing sharp performance degradation when the target variable is non-linearly transformed. We now precisely describe the experimental setup underlying this illustration.

F.1 DATA DISTRIBUTION

To study this question in a controlled setting, we construct a synthetic dataset with state $S \in \{1, \ldots, 11\}$, time period $T \in \{0, \ldots, 10\}$, and exogenous noise $U \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$. The outcome is generated as

$$Y = \mu_S + \mu_T + \sigma U,$$

where

$$\boldsymbol{\mu}_T = [11 \quad 10 \quad \cdots \quad 1]$$
$$\boldsymbol{\mu}_S = [0 \quad 0.02 \quad \cdots \quad 0.08 \quad 10 \quad 19.92 \quad 19.94 \quad \cdots \quad 20.0]$$
$$\boldsymbol{\sigma} = 0.5.$$

We draw 1,200 independent rows for each (S, T) pair (by sampling $1,200$ corresponding U i.i.d.), giving a 4-column table with columns (Y, S, T, U). State 6 is subject to an intervention during periods $T \in \{8, 9, 10\}$. The construction is chosen so that state 6, for which $\mu_6 = 10$, lies between two clustered groups of donor states. The task is to learn a DiffLTM-CFG model on this table of four columns and then conditionally generate Y for the unseen slices which is $S = 6$ and $T \in \{8, 9, 10\}$.

To study robustness to reparameterization, we transform the outcome Y using a function $f(Y)$. The training data are then given by the table $(f(Y), S, T, U)$. The three experiments in Figure 4 correspond to three choices of f: (i) $f(y) = y$, (ii) $f(y) = \log y$, and (iii) $f(y) = y(\max(y) - y)$. The third setting is included to evaluate robustness under a non-monotone reparameterization, where $\max(y)$ denotes the maximum value of Y in the training set. Figure 4 compares the median of the distribution generated by our model with the median of the observed data and the median of the distribution generated by synthetic control.

F.2 TRAINING AND GENERATION

Along with learned embeddings and logit parameterization, we use cross-entropy (CE) loss for the categorical features (S and T) and mean-squared error (MSE) loss for the numerical features (Y and U). Given the smaller dataset sizes, we use a smaller model for these experiments with embedding dimension $d_{\mathrm{emb}} = 4$, hidden width $d_h = 256$, and depth $L = 1$, resulting in approximately $350{,}000$ trainable parameters. Unlike the other experiments, we standardize the numerical columns rather than applying a quantile transformation. The model is trained for $1{,}000$ epochs with a batch size of 4096, and the final checkpoint is selected based on the minimum EMA validation loss. Similar to Appendix E.3, we use AdaLN for these experiments. All other training and architectural details follow Appendix C.

For each choice of $f(Y)$, we repeat the experiment across three training seeds. For each trained model, we generate $10{,}000$ samples for every (S, T) pair using four different generation seeds. Each median reported in Figure 4 is first averaged across the four generation seeds, and the standard error bars are then computed across the three training seeds.

G MINCER RETURNS TO EDUCATION

An important property of synthetic data is that analyses performed on the synthetic data lead to the same downstream conclusions as analyses performed on the real data. The MLE ratio in (13) measures this property in one way. Another test with a more direct semantic interpretation is whether econometric conclusions drawn from the ACSLabour dataset continue to hold when the analysis is repeated on synthetic data. A simple first test is to fit a linear regression and ask whether the coefficients fit on real and synthetic data match.

For the ACSLabour dataset in particular, since the derived target is log wage, the most natural testbed is the Mincer earnings equation introduced by Mincer (1974) (see also Lemieux (2006)). We use the quadratic specification

$$\log(W) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 S + \epsilon,$$

where W is wage, X is potential years of experience, S is years of schooling, and ϵ is exogenous noise. We obtain S by mapping the ACS education categories to completed years of schooling and define potential experience as $X = \texttt{AGEP} - S - 6$, assuming that schooling begins at age six and that work begins immediately after schooling. Rows for which $X < 0$ are excluded. The coefficient β_3 captures the association between schooling and log wages, while β_1 and β_2 describe the wage-experience profile.

For each subpopulation, we first compute the least squares fit on the real slice to obtain β_{real}. We then generate synthetic rows using each of CFG and SMC and fit OLS to obtain β_{CFG} and β_{SMC}. We repeat generation with three random seeds. Finally, we compute $\|\beta_{\text{real}} - \beta_{\text{syn}}\|_2^2$ for each method and plot it as a function of subpopulation rarity in Figure 7a, defined as the percentage of real training rows belonging to that subpopulation. The considered slices are the same as those used in Figure 1, with the important exclusion of those defined using the wage, schooling, or age columns.

As a sanity check, we first perform unconditional generation from the full dataset. In this setting

$$\beta_{\text{real}} = [7.60583 \quad 0.08540 \quad -0.001346 \quad 0.13039]\,,$$
$$\beta_{\text{syn}} = [7.70212 \quad 0.09132 \quad -0.001499 \quad 0.12002]\,.$$

The synthetic data closely recover all four coefficients. Note in particular the positive schooling coefficient as well as the positive linear and negative quadratic experience coefficients: these reproduce the known results on (i) higher education translating to higher earnings at each each level, and (ii) the concave wage-experience profile.

More importantly, conditional generation should preserve relationships that differ substantially from those of the full distribution. One such example is the subpopulation reporting no English proficiency ($\approx 0.95\%$ of the dataset), where the fitted coefficients on real and (conditionally generated) synthetic data are given by

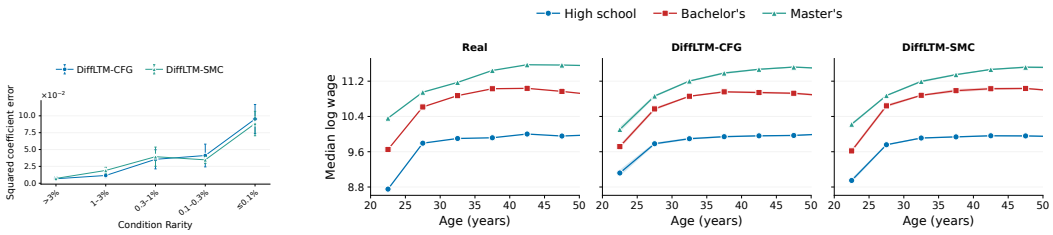
$$\beta_{\text{real}} = [9.21015 \quad 0.02995 \quad -0.000395 \quad 0.01247]\,,$$
$$\overline{\beta}_{\text{CFG}} = [9.13192 \quad 0.03331 \quad -0.000440 \quad 0.01450]\,,$$
$$\overline{\beta}_{\text{SMC}} = [9.14925 \quad 0.03519 \quad -0.000458 \quad 0.01018]\,.$$

The schooling coefficient falls from 0.130 in the full dataset to 0.0125 in this subpopulation, while the initial experience coefficient falls from 0.0854 to 0.03; suggesting that the return on education and experience for this group is substantially lower than that for the full population. Both CFG and SMC successfully recover this rare subgroup-specific conclusion.

Finally, we examine the implied returns-to-education curves for China-born female private-sector workers. We condition on this subpopulation, in turn, on high-school, bachelor's, and master's education, while jointly generating age and log wage. We then plot median log wage across five-year age bins to test whether the synthetic data preserve both the ordering of earnings by education and the concave age-earnings profiles observed in the real data; as illustrated in Figure 7b, this curve is successfully reproduced by both CFG and SMC.

H Additional empirical results

We present results on fidelity vs rarity across fidelity metrics and datasets—see Table 5. The protocol used for selecting slices, training the models, conditional generation, and metrics evaluation is exactly that described in Appendix D to obtain Figure 1. We remark that our conclusion of our methods outperforming the Shankar et al. (2026) baseline is very robust across metrics and datasets.



(a) Coeff. error versus rarity. (b) Returns to education (China-born, Female, Private Industry).

Figure 7: Conditionally generated data preserve subgroup-specific wage structure in ACSLabour. (a) Squared error between Mincer coefficients fitted on generated and real slices grows smoothly with rarity, similarly for CFG and SMC. (b) For China-born female private-sector workers, both methods reproduce the education ordering and the concave age–earnings profile of the real data (Appendix G).

Fidelity Metric	Dataset	Figure
Shape	ACSLabour	Figure 8
Trend	ACSLabour	Figure 9
Conditional Shape (Rare)	ACSLabour	Figure 10
Shape	ACSPubCov	Figure 11
Trend	ACSPubCov	Figure 12
Conditional Shape	ACSPubCov	Figure 13
Conditional Shape (Rare)	ACSPubCov	Figure 14
Shape	ACSMobility	Figure 15
Trend	ACSMobility	Figure 16
Conditional Shape	ACSMobility	Figure 17
Conditional Shape (Rare)	ACSMobility	Figure 18

Table 5: Additional fidelity vs. rarity results across fidelity metrics and datasets.

I PSEUDOCODE

Algorithm 2: Generation with DiffLTM

Input: trained denoiser f_{θ^*}, embeddings $\{\phi_j\}$, noise schedule σ_t, grid $1 = t_0 > \cdots > t_T = 0$

Initialize $\boldsymbol{r}^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{t_0}^2 I_D)$

for $i = 0, \ldots, T-1$ **do**

  Evaluate the denoiser at $(\boldsymbol{r}^{(i)}, t_i)$

  if *categorical output is logits* **then**

    Compute $\widehat{p}_{j,c}^{(i)} \leftarrow \mathrm{softmax}(\boldsymbol{\ell}_j^{(i)})_c$ and $\widehat{\boldsymbol{r}}_{\mathrm{cat},j}^{(i)} \leftarrow \sum_{c=1}^{C_j} \widehat{p}_{j,c}^{(i)} \phi_j(c)$

  Form the full denoised estimate $\widehat{\boldsymbol{r}}^{(i)}$

  Update $\boldsymbol{r}^{(i+1)} \leftarrow \boldsymbol{r}^{(i)} + (\sigma_{t_{i+1}} - \sigma_{t_i}) \frac{\boldsymbol{r}^{(i)} - \widehat{\boldsymbol{r}}^{(i)}}{\sigma_{t_i}}$

Return the final numerical coordinates directly

if *categorical output is representation* **then**

  Decode each categorical column as $\widehat{x}_j \leftarrow \arg\min_c \|\widehat{\boldsymbol{r}}_{\mathrm{cat},j} - \phi_j(c)\|_2$

else

  Decode each categorical column as $\widehat{x}_j \leftarrow \arg\max_c \widehat{p}_{j,c}$

Output: generated mixed-type row $\boldsymbol{x}_{\mathrm{syn}}$

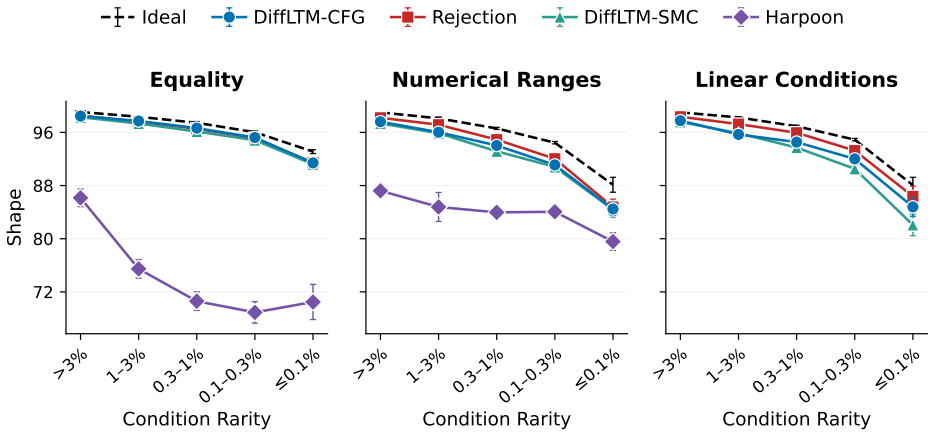


Figure 8: Shape vs rarity on ACSLabour Dataset.

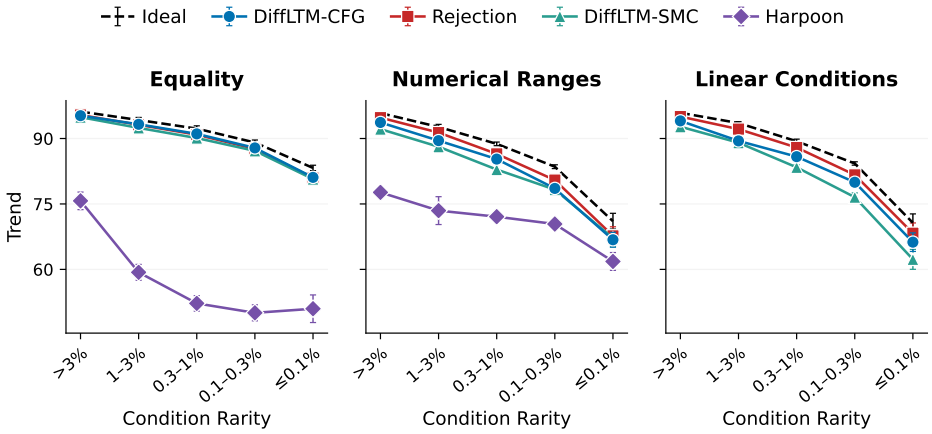


Figure 9: Trend vs rarity on ACSLabour Dataset.

Algorithm 3: SMC for sampling $p(\boldsymbol{x} \mid \mathcal{C})$

Input: condition $\mathcal{C}$, K particles, Noise schedule, n_{syn} output rows, effective sample size (ESS) threshold αK

for $k \in [K]$ **do**

  $\boldsymbol{r}_k^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_{t_0}^2 I)$; $\log w_k \leftarrow 0$; $\log p_{\text{prev},k} \leftarrow 0$

  `// Initialize`

for $i = 0, \ldots, T-1$ **do**

  Estimate $p_k^{(i)}(\mathcal{C})$

  $\log w_k \leftarrow \log w_k + \log p_k^{(i)}(\mathcal{C}) - \log p_{\text{prev},k}$ for all k `// reweight`

  if $\text{ESS}(\log \boldsymbol{w}) < \alpha K$ *and* $i < T-1$ **then**

    Systematically resample the pairs $\{(\boldsymbol{r}_k^{(i)}, p_k^{(i)}(\mathcal{C}))\}_{k=1}^K$ according to $\boldsymbol{w}$; set $\log w_k \leftarrow 0$

    `// resample`

  $\log p_{\text{prev},k} \leftarrow \log p_k^{(i)}(\mathcal{C})$ for all k

  $\boldsymbol{r}_k^{(i+1)} \leftarrow \textsc{SdeStep}(\boldsymbol{r}_k^{(i)}, t_i \to t_{i+1})$ for all k `// move`

Let $\widehat{\boldsymbol{x}}_k$ be the candidate synthetic row produced by trajectory $\boldsymbol{r}_k^{(T)}$

$\log w_k \leftarrow \log w_k + \log \mathbf{1}\{\widehat{\boldsymbol{x}}_k \text{ satisfies } \mathcal{C}\} - \log p_{\text{prev},k}$ for all k

Output: n_{syn} rows systematically resampled from $\{\widehat{\boldsymbol{x}}_k\}_{k=1}^K$ according to $\boldsymbol{w}$

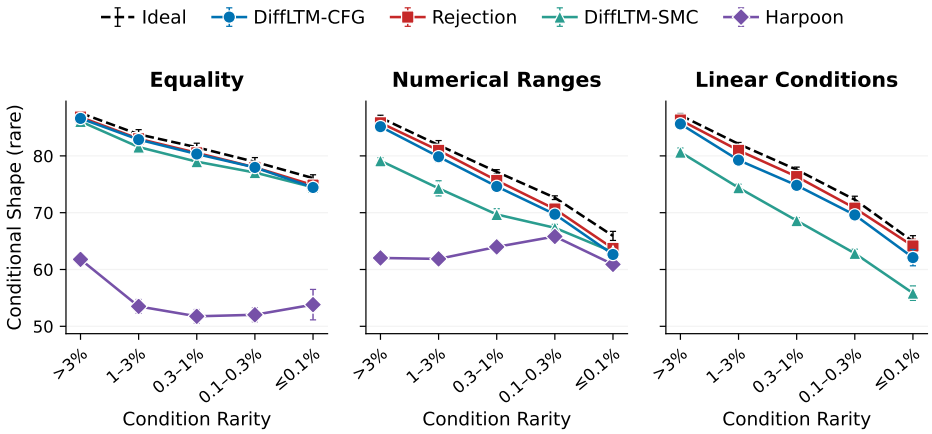


Figure 10: Conditional Shape (Rare) vs rarity on ACSLabour Dataset.

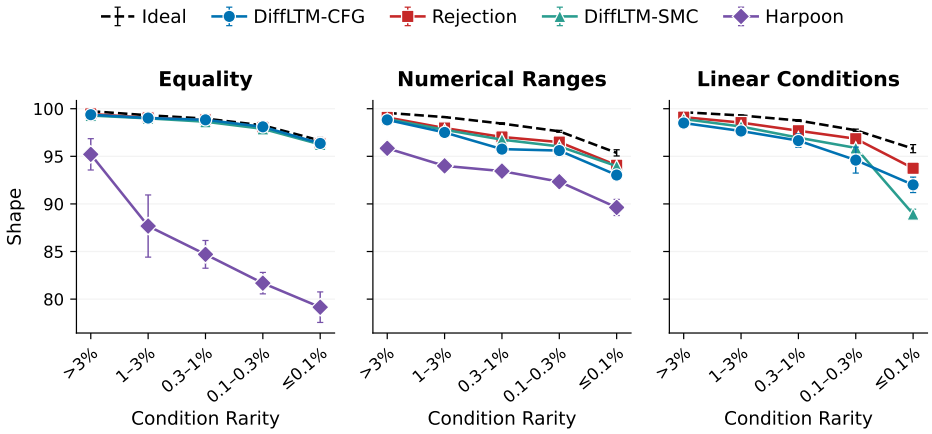


Figure 11: Shape vs rarity on ACSPubCov Dataset.

J Asymptotic analysis of discrete variables generation

In this appendix we give a formal presentation of the theoretical results of Section 4 establishing optimality of CE loss in the undersmoothed regime. Section J.1 introduces the setting. Section J.2 describes the asymptotics of parameters estimates. Section J.3 focuses on the undersmoothing regime $b \ll n^{-1/2}$ and Section J.4 characterizes the optimal sampling density λ in this setting. (This is not needed for our main result, but is nevertheless a useful byproduct of the analysis.) Finally, proofs of the various technical steps are given in Section J.5.

J.1 Setting and assumptions

We have a categorical distribution $(q(\sigma) : \sigma \in \{0, \dots, k-1\})$, and embeddings $\boldsymbol{x}(\sigma) \in \mathbb{R}^d$, $\sigma \in \{0, \dots, k-1\}$, which we assume to be normalized so that $\|\boldsymbol{x}(\sigma)\| = 1$ for all $\sigma \in \{0, \dots, k-1\}$. Throughout, we denote by $\boldsymbol{X}$ a random variable in $\mathbb{R}^d$ with $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}(\sigma)) = q(\sigma)$, for $\sigma \in \{0, \dots, k-1\}$, and denote the law of $\boldsymbol{X}$ by $\mathbb{P}_q$ whenever we want to emphasize its dependence on q. We assume the embeddings are distinct and all class probabilities are positive. We parametrize

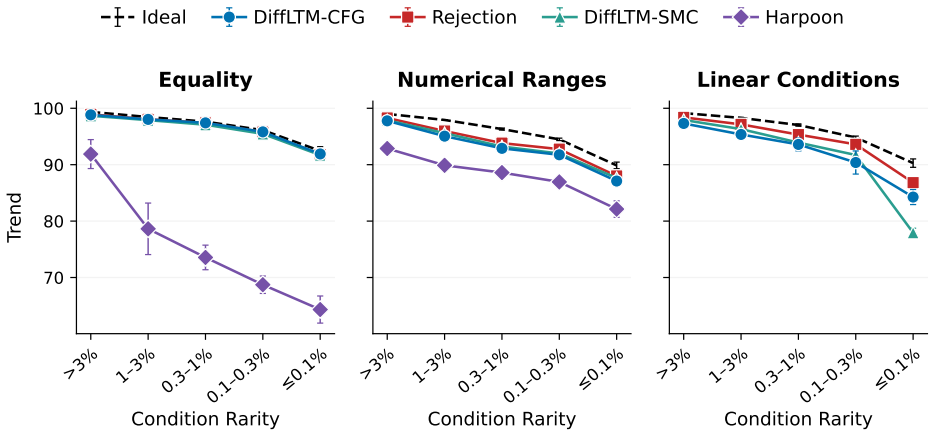


Figure 12: Trend vs rarity on ACSPubCov Dataset.

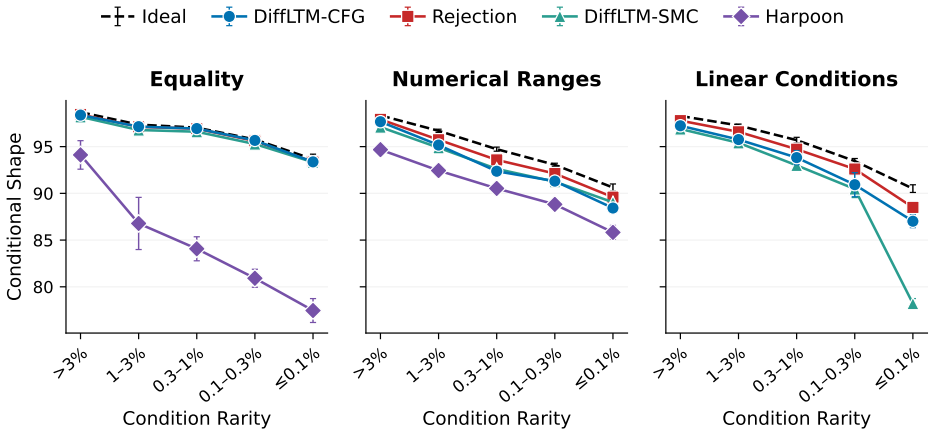


Figure 13: Conditional Shape vs rarity on ACSPubCov Dataset.

the probabilities by log-weights $\{r(\sigma)\}_{\sigma=1}^{k-1}$ as follows

$$q(\sigma) = \frac{1}{Z_0} e^{-r(\sigma)}, \qquad \sigma \in \{1, \ldots, k-1\}, \tag{15}$$

$$q(0) = \frac{1}{Z_0}, \qquad Z_0 := 1 + \sum_{\sigma=1}^{k-1} e^{-r(\sigma)}. \tag{16}$$

We consider the following stochastic process over $t \in [0, T]$:

$$\boldsymbol{X}_t = a(t)\boldsymbol{X} + \tau_0(t)\boldsymbol{B}_{s(t)}, \tag{17}$$

where $(\boldsymbol{B}_s)_{s\geq 0}$ is a standard Brownian motion, independent of $\boldsymbol{X} \sim \mathbb{P}_q$, $a : [0, T] \to \mathbb{R}$ is non-decreasing with $a(0) = 0$ and $a(T) = 1$, and $s : [0, T] \to \mathbb{R}$ is non-decreasing $s(0) \geq 0$, $\tau_0(t) \geq 0$ for all $t \in [0, T]$, and $\tau_0(T)^2 s(T) = 0$. Further, all of these functions are continuously differentiable. The variance at time t is denoted by $\tau(t)^2 := \tau_0(t)^2 s(t)$.

The stochastic process $\{\boldsymbol{X}_t\}_{t\in[0,T]}$ is also characterized as the strong solutions of the following stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)\,\mathrm{d}t + \sqrt{g(t)}\mathrm{d}\boldsymbol{W}_t, \tag{18}$$

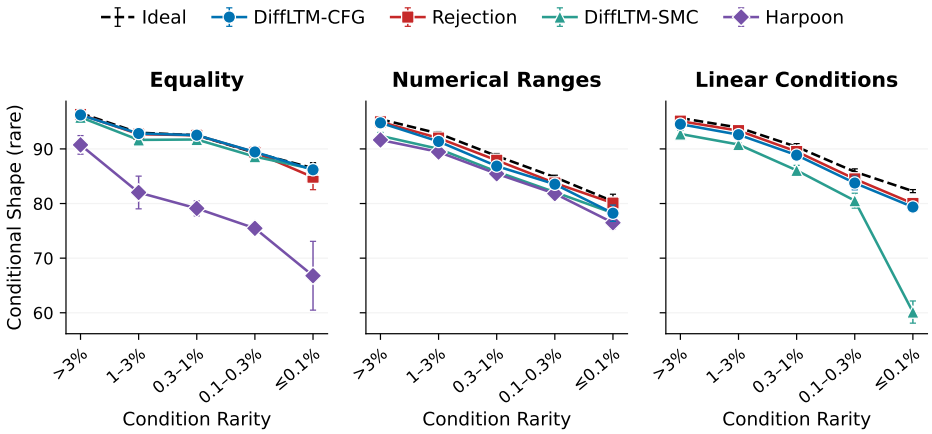


Figure 14: Conditional Shape (Rare) vs rarity on ACSPubCov Dataset.

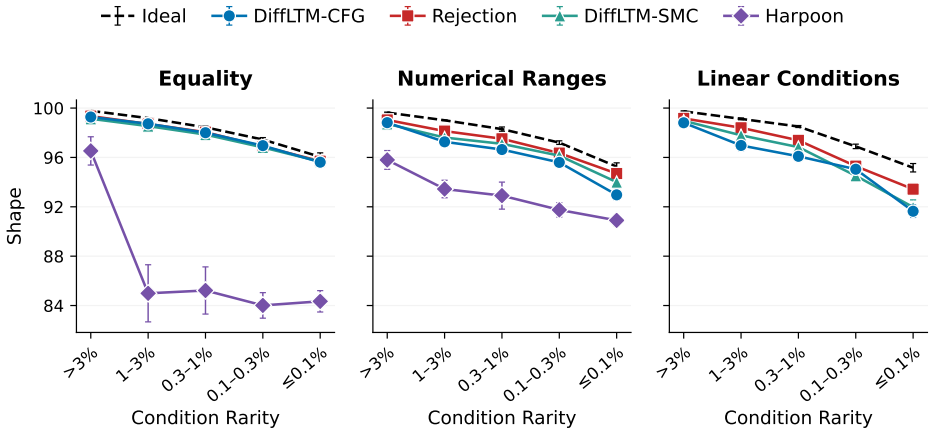


Figure 15: Shape vs rarity on ACSMobility Dataset.

where $\boldsymbol{W}_t$ is a standard Brownian motion and

$$\boldsymbol{f}(\boldsymbol{x},t) = \tau_0(t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{a(t)}{\tau_0(t)}\right)\boldsymbol{m}(\boldsymbol{x},t) + \frac{\dot{\tau}_0(t)}{\tau_0(t)}\boldsymbol{x}\,, \qquad g(t) = \tau_0(t)^2\dot{s}(t)\,, \tag{19}$$

$$\boldsymbol{m}(\boldsymbol{x},t) := \mathbb{E}[\boldsymbol{X}|\boldsymbol{X}_t = \boldsymbol{x}]\,. \tag{20}$$

The evolution of the distribution of $\boldsymbol{X}_t$ is given by the following Fokker-Planck PDE

$$\partial_t p(\boldsymbol{x},t) = -\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{f}(\boldsymbol{x},t)p(\boldsymbol{x},t)) + \frac{1}{2}g(t)\Delta_{\boldsymbol{x}}p(\boldsymbol{x},t)\,, \tag{21}$$

$$p(\boldsymbol{x},0) = \frac{1}{(2\pi\tau(0)^2)^{d/2}}\,e^{-\|\boldsymbol{x}\|^2/(2\tau(0)^2)}\,. \tag{22}$$

The one-time marginals of $\boldsymbol{X}_t$ are the same as the ones of the probability flow ODE:

$$\mathrm{d}\boldsymbol{Z}_t = \boldsymbol{h}(\boldsymbol{Z}_t,t)\,\mathrm{d}t\,, \tag{23}$$

$$\boldsymbol{h}(\boldsymbol{z},t) = \boldsymbol{f}(\boldsymbol{z},t) - \frac{1}{2}g(t)\nabla_{\boldsymbol{z}}\log p(\boldsymbol{z},t)\,. \tag{24}$$

Tweedie's formula reads

$$\boldsymbol{m}(\boldsymbol{x},t) = \frac{1}{a(t)}\left\{\boldsymbol{x} + \tau_0(t)^2 s(t)\,\nabla\log p(\boldsymbol{x},t)\right\}\,. \tag{25}$$

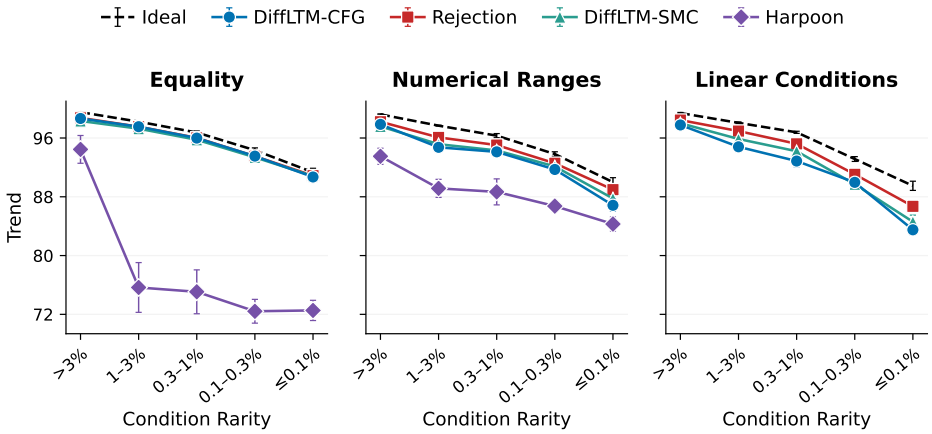


Figure 16: Trend vs rarity on ACSMobility Dataset.

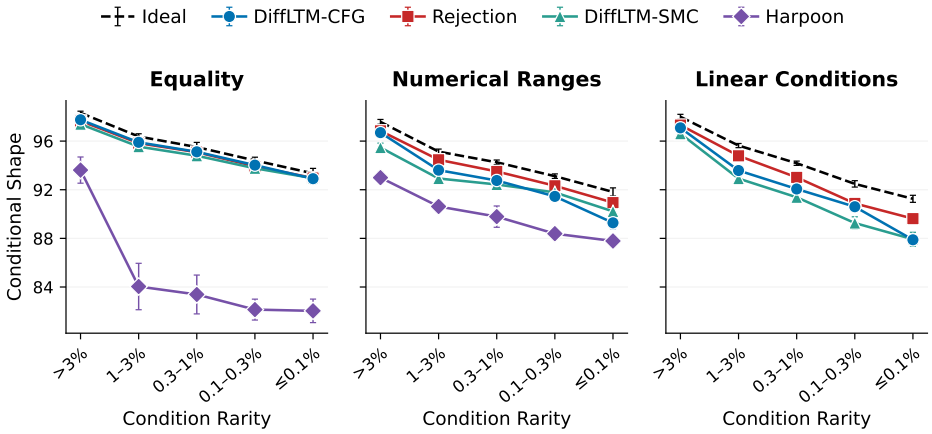


Figure 17: Conditional Shape vs rarity on ACSMobility Dataset.

This yields

$$\boldsymbol{h}(\boldsymbol{z},t) = \left\{\dot{a}(t) - \frac{a(t)\dot{\tau}_0(t)}{\tau_0(t)} - \frac{a(t)\dot{s}(t)}{2s(t)}\right\}\boldsymbol{m}(\boldsymbol{z},t) + \left\{\frac{\dot{\tau}_0(t)}{\tau_0(t)} + \frac{\dot{s}(t)}{2s(t)}\right\}\boldsymbol{z} \tag{26}$$

$$= \frac{\dot{a}(t)}{a(t)}\boldsymbol{z} + \tau_0(t)^2 s(t)\left\{\frac{\dot{a}(t)}{a(t)} - \frac{\dot{\tau}_0(t)}{\tau_0(t)} - \frac{\dot{s}(t)}{2s(t)}\right\}\nabla \log p(\boldsymbol{z},t)\,. \tag{27}$$

We are given i.i.d. data triples $\{(t_i, \sigma_i, \boldsymbol{Y}_i) : i \le n\}$. Their common law is specified by $\sigma_i \in \{0, 1, \dots, k-1\}$, $\mathbb{P}(\sigma_i = \sigma) = q(\sigma)$, $t_i \sim \lambda(t)\cdot \mathrm{d}t$, independently of σ_i, where λ is a probability density on $[0, T]$, and

$$\boldsymbol{Y}_i|_{(t_i,\sigma_i)} \sim \mathsf{N}\big(\boldsymbol{x}(\sigma_i)\, a(t_i), \tau(t_i)^2 \boldsymbol{I}_d\big), \qquad i \le n\,. \tag{28}$$

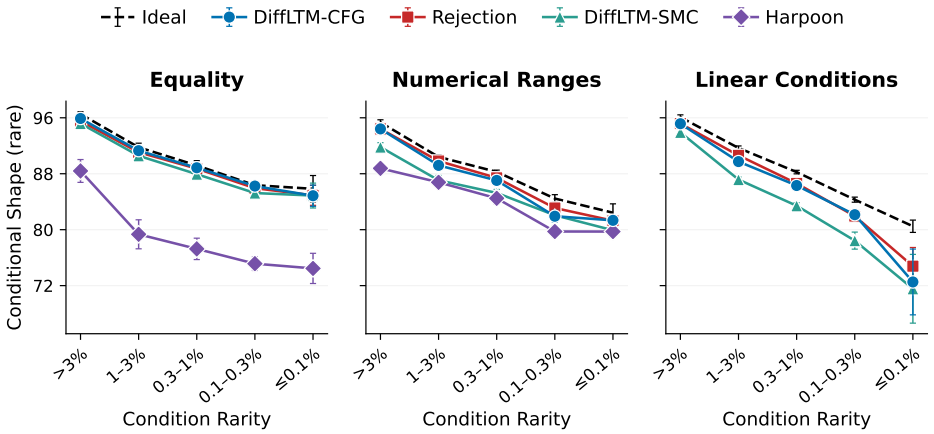


Figure 18: Conditional Shape (Rare) vs rarity on ACSMobility Dataset.

Algorithm 1: Training DiffLTM

Input: training data, denoiser f_θ, embeddings $\{\phi_j\}$, noise schedule σ_t, categorical parameterization and loss type

repeat

- Sample a minibatch of rows $\boldsymbol{x}$ and construct their continuous representations $\boldsymbol{r}$ using (3)
- Sample $t \sim \mathrm{Unif}[0,1]$ and $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I_D)$
- Construct the noisy input $\boldsymbol{r}_t \leftarrow \boldsymbol{r} + \sigma_t \boldsymbol{z}$
- Evaluate $f_\theta(\boldsymbol{r}_t, t)$
- Compute $\mathcal{L}_{\text{num}} \leftarrow \frac{1}{m_{\text{num}}} \|\widehat{\boldsymbol{r}}_{\text{num}} - \boldsymbol{r}_{\text{num}}\|_2^2$
- **if** *categorical output is logits* **then**
  - Compute $\widehat{p}_{j,c} \leftarrow \mathrm{softmax}(\boldsymbol{\ell}_j)_c$ for each categorical column j
- **if** *categorical loss is MSE* **then**
  - If logits mode, construct $\widehat{\boldsymbol{r}}_{\text{cat},j} \leftarrow \sum_{c=1}^{C_j} \widehat{p}_{j,c} \phi_j(c)$ using (4)
  - Compute $\mathcal{L}_{\text{cat}} \leftarrow \mathcal{L}_{\text{cat}}^{\text{MSE}}$ using (5)
- **else**
  - Compute $\mathcal{L}_{\text{cat}} \leftarrow \mathcal{L}_{\text{cat}}^{\text{CE}}$ using (5)
- Update θ and, for learned embeddings, $\{E_j\}$ using gradients of $\mathcal{L}_{\text{num}} + \mathcal{L}_{\text{cat}}$

until *convergence*

We learn a generative model of the form equation 23 where $\boldsymbol{m}$ is replaced by

$$\hat{\boldsymbol{m}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(t)) = \sum_{\sigma=0}^{k-1} \hat{q}_\sigma(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(t)), \boldsymbol{x}(\sigma)\,, \tag{29}$$

$$\hat{q}_0(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{x}, t, \boldsymbol{\theta})}\,, \quad \hat{q}_\sigma(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{x}, \boldsymbol{\theta})} \exp\left\{\theta_{\sigma,0} + \langle \boldsymbol{\theta}_\sigma^+, \boldsymbol{x}\rangle\right\}, \quad \sigma \in \{1, \ldots, k-1\}\,, \tag{30}$$

$$z(\boldsymbol{x}, \boldsymbol{\theta}) = 1 + \sum_{\sigma=1}^{k-1} \exp\{\theta_{\sigma,0} + \langle \boldsymbol{\theta}_\sigma^+, \boldsymbol{x}\rangle\}\,. \tag{31}$$

We write $\boldsymbol{\theta}_\sigma = (\theta_{\sigma,0}, \boldsymbol{\theta}_\sigma^+) \in \mathbb{R}^{d+1}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}) \in E$, where $E := \mathbb{R}^{(d+1)\times(k-1)}$ is the parameter space.

The function $\boldsymbol{\theta}(t)$ is a parameter learned via the following weighted ERM:

$$\hat{\boldsymbol{\theta}}_b(t) := \arg\min \left\{ \hat{R}_{b,t}(\boldsymbol{\theta}) : \ \boldsymbol{\theta} \in K(t) \right\}, \tag{32}$$

$$\hat{R}_{b,t}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} w_b(t_i - t)\, \ell_{\sigma_i}(\boldsymbol{Y}_i; \boldsymbol{\theta}). \tag{33}$$

where $w_b(t) = w_1(t/b)$ is an unnormalized weight function and b is a bandwidth parameter and $t \mapsto K(t) \subseteq E$ specifies the time-dependent optimization domain. All empirical and population minimizations at time t use the same set $K(t)$, subject to Assumption 5 below.

We define the population risk at time t R_t and its non-zero bandwidth version $R_{b,t}$ as:

$$R_t(\boldsymbol{\theta}) := \mathbb{E}\left[\ell_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G}; \boldsymbol{\theta}\big) \right], \tag{34}$$

$$R_{b,t}(\boldsymbol{\theta}) := \mathbb{E}\left[w_b(U - t) \ell_\Sigma\big(a(U)\boldsymbol{x}(\Sigma) + \tau(U)\boldsymbol{G}; \boldsymbol{\theta}\big) \right], \tag{35}$$

where U has density λ, $\Sigma \sim q$, $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, and U, Σ, and $\boldsymbol{G}$ are independent. Finally, we define the corresponding population estimators as

$$\boldsymbol{\theta}^{b=0}(t) := \underset{\boldsymbol{\theta} \in K(t)}{\arg\min}\, R_t(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}^{b}(t) := \underset{\boldsymbol{\theta} \in K(t)}{\arg\min}\, R_{b,t}(\boldsymbol{\theta}), \tag{36}$$

and the 'target' parameters as $\boldsymbol{\theta}^*(t)$, where, for $\sigma \in \{1, \ldots, k-1\}$,

$$\theta^*_{\sigma,0}(t) := -r(\sigma), \qquad \boldsymbol{\theta}^{*,+}_{\sigma}(t) := \frac{a(t)}{\tau(t)^2}\{\boldsymbol{x}(\sigma) - \boldsymbol{x}(0)\}. \tag{37}$$

(The model equation 30 matches the true conditional distribution for $\hat{\boldsymbol{\theta}}(t) = \boldsymbol{\theta}^*(t)$.)

Our first assumption is that the loss function identifies the true parameters at population level on the compact time neighborhood under consideration.

Assumption 1 (Fisher consistency on J)**.** *For every $t \in J$, the population risk R_t has a unique minimizer over $K(t)$, and*

$$\boldsymbol{\theta}^*(t) = \boldsymbol{\theta}^{b=0}(t) \in \operatorname{int}\big(K(t)\big). \tag{38}$$

We emphasize that we allow the sets $K(t)$ to expand with t.

Our second assumption is that the generative process is given by the probability flow ODE equation 23, with learned parameters $\hat{\boldsymbol{\theta}}_b$.

Assumption 2 (Generative process)**.** *Assume a, τ_0, and s are smooth and $a(0) = 0$, $a(T) = 1$, $\tau_0(T)^2 s(T) = 0$. We consider the generation process given by the ODE:*

$$\mathrm{d}\boldsymbol{Z}_t = \hat{\boldsymbol{h}}(\boldsymbol{Z}_t, t)\,\mathrm{d}t, \quad \boldsymbol{Z}_0 \sim \mathsf{N}(0, \tau(0)^2 \boldsymbol{I}_d). \tag{39}$$

where

$$\hat{\boldsymbol{h}}(\boldsymbol{z}, t) := c_1(t)\hat{\boldsymbol{m}}(\boldsymbol{z}, t; \hat{\boldsymbol{\theta}}_b(t)) + c_2(t)\boldsymbol{z}, \tag{40}$$

$$c_1(t) := \dot{a}(t) - \frac{a(t)\dot{\tau}_0(t)}{\tau_0(t)} - \frac{a(t)\dot{s}(t)}{2s(t)}, \tag{41}$$

$$c_2(t) := \frac{\dot{\tau}_0(t)}{\tau_0(t)} + \frac{\dot{s}(t)}{2s(t)}. \tag{42}$$

The next assumption ensures that the $b \to 0$ limit of the population minimizer $\boldsymbol{\theta}^b(t)$ is well behaved.

Assumption 3 (Undersmoothing regime)**.** *We assume $b = b_n$ and impose throughout the undersmoothing conditions*

$$n^{1/4} b_n \to 0, \qquad \frac{\sqrt{n}\, b_n}{\log(1/b_n)} \longrightarrow \infty. \tag{43}$$

All parameter derivatives are taken in the Euclidean space E with Frobenius inner product and norm. Hessians and covariance matrices are understood as linear operators on E; operator composition is denoted by $\circ$. All statistical limits keep k, d fixed.

We next state some standard conditions for the kernel.

Assumption 4 (Kernel)**.** *The kernel w_1 is nonnegative, symmetric, supported on $[-1,1]$, Hölder continuous on $\mathbb{R}$ with exponent $\alpha \in (0,1]$, and $w_1(0) > 0$. Define*

$$\mu_0 := \int_{-1}^{1} w_1(v)\,\mathrm{d}v > 0, \qquad \mu_2 := \int_{-1}^{1} v^2 w_1(v)\,\mathrm{d}v. \tag{44}$$

Assumption 5 (Zero-bandwidth regularity on $I \subset [0,T)$)**.** *Fix a compact interval $I \subset [0,T)$ and a compact relative neighborhood $J \subset [0,T)$ of I.*

1. *There is a compact convex localization set $K_0 \subset E$ with nonempty interior (depending on I) such that*

$$\inf_{t\in J} \operatorname{dist}\big(\boldsymbol{\theta}^*(t), E \setminus K_0\big) > 0, \tag{45}$$

and one of the following holds:

   (a) *The sets $K(t)$ are nonempty, compact, and convex for $t \in J$; their graph*

$$\mathcal{K}_J := \{(t,\boldsymbol{\theta}) : t \in J,\ \boldsymbol{\theta} \in K(t)\} \tag{46}$$

   is a compact subset of $J \times K_0$; and the target has a uniform interior margin,

$$\inf_{t\in J} \operatorname{dist}\big(\boldsymbol{\theta}^*(t), E \setminus K(t)\big) > 0. \tag{47}$$

   (b) *$K(t) = E$ for every $t \in J$, and $\ell_\sigma(\boldsymbol{y};\cdot) : E \to \mathbb{R}$ is finite and convex for every $(\sigma,\boldsymbol{y})$.*

2. *The losses and their first three parameter derivatives are jointly continuous in $(\boldsymbol{y},\boldsymbol{\theta})$ on an open neighborhood of $\mathbb{R}^d \times K_0$. For finite C,κ,*

$$\max_{0\le j\le 3} \sup_{\boldsymbol{\theta}\in K_0} \|D^j_{\boldsymbol{\theta}} \ell_\sigma(\boldsymbol{y};\boldsymbol{\theta})\| \le C(1+\|\boldsymbol{y}\|)^\kappa, \qquad 0 \le \sigma \le k-1. \tag{48}$$

   Here $D^0\ell = \ell$ and the norms for $j \ge 2$ are multilinear operator norms.

3. *The functions a,τ,λ are continuous on J, $\inf_J \lambda > 0$, and*

$$\inf_{t\in J} \lambda_{\min}\big(\nabla^2_{\boldsymbol{\theta}} R_t(\boldsymbol{\theta}^{b=0}(t))\big) > 0. \tag{49}$$

Assumption 6 (Positive noise)**.** *The schedule satisfies $\tau(t) > 0$ for $0 \le t < T$.*

Since $\tau(t) > 0$, for any $t < T$ the true density is given by

$$p_t(\boldsymbol{z}) = \sum_{\sigma=0}^{k-1} q(\sigma) \frac{\exp\{-\|\boldsymbol{z} - a(t)\boldsymbol{x}(\sigma)\|^2/(2\tau(t)^2)\}}{(2\pi\tau(t)^2)^{d/2}}. \tag{50}$$

For $0 \le s < t < T$, let $\boldsymbol{\Phi}^*_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ denote the flow associated to the ODE equation 23. Write $\boldsymbol{\Phi}^*_{t,s} := (\boldsymbol{\Phi}^*_{s,t})^{-1}$. For $0 < t < T$, define the terminal basins

$$\mathcal{A}_\sigma(t) := \left\{ \boldsymbol{z} \in \mathbb{R}^d : \lim_{u\uparrow T} \boldsymbol{\Phi}^*_{t,u}(\boldsymbol{z}) = \boldsymbol{x}(\sigma) \right\}, \qquad 0 \le \sigma \le k-1, \tag{51}$$

The next lemma gives some simple regularity properties of the basins and is essentially extracted from Wan et al. (2025).

Lemma J.1 (Terminal convergence and measurable basins Wan et al. (2025))**.** *Under Assumptions 2 and 6, the $0 \le s < t < T$, $\boldsymbol{\Phi}^*_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth diffeomorphism between. For each $0 \le s < T$, $\lim_{t\to T} \boldsymbol{\Phi}^*_{s,t}(\boldsymbol{z})$ exists and belongs to $\{\boldsymbol{x}(0),\dots,\boldsymbol{x}(k-1)\}$ for Lebesgue-almost every starting point $\boldsymbol{z}$ at time s. The sets $\mathcal{A}_\sigma(s)$, with the same definition also at $s = 0$, are Borel measurable and partition $\mathbb{R}^d$ up to a null set. For $0 \le s < t < T$,*

$$\boldsymbol{\Phi}^*_{s,t}(\mathcal{A}_\sigma(s)) = \mathcal{A}_\sigma(t), \qquad \int_{\mathcal{A}_\sigma(t)} p_t(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = q(\sigma). \tag{52}$$

For self-containedness, we give a prove in Appendix J.5.1.

Score matching and cross-entropy are two popular losses

$$\ell_\sigma^{\mathrm{SM}}(\boldsymbol{y};\boldsymbol{\theta}) = \|\boldsymbol{x}(\sigma) - \hat{\boldsymbol{m}}(\boldsymbol{y};\boldsymbol{\theta})\|^2, \tag{53}$$

$$\ell_\sigma^{\mathrm{CE}}(\boldsymbol{y};\boldsymbol{\theta}) = -\log \hat{q}_\sigma(\boldsymbol{y};\boldsymbol{\theta}). \tag{54}$$

Proposition J.2 (Verification for score matching and cross-entropy)**.** *Suppose a, τ, λ are continuous on $J \subset [0,T)$ and $\inf_J \tau > 0$, $\inf_J \lambda > 0$. Choose a compact convex parameter box K_0 containing $\{\boldsymbol{\theta}^*(t) : t \in J\}$ in its interior.*

- (a) *For cross-entropy, Assumptions 1 and 5 hold on J with $K(t) = E$.*
- (b) *For score matching loss Assumptions 1 and 5 hold for any compact family $K(t)$ $K(t)$ whose graph is compact in $J \times K_0$ and which satisfies equation 47, provided that the embeddings are affinely independent:*

$$\operatorname{rank}\big[\boldsymbol{x}(1) - \boldsymbol{x}(0), \ldots, \boldsymbol{x}(k-1) - \boldsymbol{x}(0)\big] = k - 1. \tag{55}$$

If a, τ are C^3 on a neighborhood of J, then the risks are jointly C^3 in $(t, \boldsymbol{\theta})$ there.

The proof is in Appendix J.5.2.

J.2 Asymptotics of parameter estimates

We next study the asymptotics of the empirical risk minimization problem under Assumption 3. Throughout $I \subset (0,T)$ is compact.

Define the loss gradient

$$\boldsymbol{\psi}_\sigma(\boldsymbol{y};\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \ell_\sigma(\boldsymbol{y};\boldsymbol{\theta}) \in E, \tag{56}$$

Note that by Assumptions 1 and 5,

$$\mathbb{E}\big\{\boldsymbol{\psi}_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G};\boldsymbol{\theta}^*(t)\big)\big\} = 0\,, \tag{57}$$

where $\Sigma \sim q$, $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, and Σ and $\boldsymbol{G}$ are independent. We also define the Hessian and covariance

$$\boldsymbol{H}^*(t) := \nabla^2_{\boldsymbol{\theta}} R_t(\boldsymbol{\theta}^*(t)), \tag{58}$$

$$\boldsymbol{J}^*(t) := \mathbb{E}\left[\boldsymbol{\psi}_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G};\boldsymbol{\theta}^*(t)\big) \otimes \boldsymbol{\psi}_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G};\boldsymbol{\theta}^*(t)\big)\right]. \tag{59}$$

Here $\boldsymbol{H}^*(t)$ is the self-adjoint linear operator associated to the Hessian quadratic form, and for matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{(d+1)\times(k-1)}$,

$$\nabla^2_{\boldsymbol{\theta}} R_t(\boldsymbol{\theta}^*(t))[\boldsymbol{A}, \boldsymbol{B}] = \langle \boldsymbol{A}, \boldsymbol{H}^*(t)[\boldsymbol{B}] \rangle\,, \tag{60}$$

$$(\boldsymbol{A} \otimes \boldsymbol{B})[\boldsymbol{C}] := \langle \boldsymbol{B}, \boldsymbol{C} \rangle \boldsymbol{A}. \tag{61}$$

Similarly $\boldsymbol{J}^*(t)$ is an operator on $\mathbb{R}^{(d+1)\times(k-1)}$.

For the finite bandwidth empirical risk minimization problem, define

$$\boldsymbol{H}_b(t) := \nabla^2_{\boldsymbol{\theta}} R_{b,t}(\boldsymbol{\theta}^b(t)), \tag{62}$$

$$\boldsymbol{Z}_{b,t} := w_b(U - t)\boldsymbol{\psi}_\Sigma(a(U)\boldsymbol{x}(\Sigma) + \tau(U)\boldsymbol{G};\boldsymbol{\theta}^b(t)), \tag{63}$$

$$\boldsymbol{\Omega}_b(t,s) := \mathbb{E}[\boldsymbol{Z}_{b,t} \otimes \boldsymbol{Z}_{b,s}], \qquad \mathbb{E}\boldsymbol{Z}_{b,t} = 0, \tag{64}$$

$$\widehat{\boldsymbol{S}}_n(t,\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^n w_b(t_i - t)\boldsymbol{\psi}_{\sigma_i}(\boldsymbol{Y}_i;\boldsymbol{\theta}). \tag{65}$$

Proposition J.3 (Localized population bias)**.** *Under Assumptions 1, 4, and 5, let $I \subset (0,T)$ be compact. Suppose λ is C^2 on an open neighborhood of I, and $(u,\boldsymbol{\theta}) \mapsto R_u(\boldsymbol{\theta})$ is C^3 on an open neighborhood of $\{(t,\boldsymbol{\theta}^*(t)) : t \in I\}$ in $(0,T) \times E$. Then, uniformly on I,*

$$\boldsymbol{\theta}^b(t) - \boldsymbol{\theta}^*(t) = b^2\boldsymbol{\beta}(t) + o(b^2), \tag{66}$$

$$\boldsymbol{\beta}(t) = -\frac{\mu_2}{2\mu_0\lambda(t)}\boldsymbol{H}^*(t)^{-1}\left[\partial_u^2\{\lambda(u)\boldsymbol{S}(u,\boldsymbol{\theta})\}\big|_{u=t,\,\boldsymbol{\theta}=\boldsymbol{\theta}^*(t)}\right], \qquad \boldsymbol{S}(u,\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} R_u(\boldsymbol{\theta}). \tag{67}$$

The derivatives ∂_u in equation 67 hold the parameter $\boldsymbol{\theta}^*$ fixed; they are not total derivatives along $\boldsymbol{\theta}^*(u)$.

The proof is in Appendix J.5.4.

The next proposition characterize the effect of an error $\boldsymbol{\Delta}(t)$ on the parameters estimates affects the sampling probabilities, using probability flow, to the linear order.

Proposition J.4 (Volume linear response)**.** *Under Assumptions 2 and 6, fix a compact interval $I \subset (0, T)$ and let $\boldsymbol{h}_*$ be the probability-flow velocity obtained by setting $\hat{\boldsymbol{\theta}}_b = \boldsymbol{\theta}^*$.*

Let

$$\hat{\boldsymbol{\theta}}_b(t) = \boldsymbol{\theta}^*(t) + \boldsymbol{\Delta}(t), \qquad \boldsymbol{\Delta} \in L^\infty(I; E), \qquad \boldsymbol{\Delta}(t) = 0 \quad (t \notin I), \tag{68}$$

and denote by $\mathrm{P}_{\boldsymbol{\Delta}}$ the law of $\boldsymbol{Z}_T$ under the corresponding generation process. For a column $\boldsymbol{\Delta}_\rho(t) \in \mathbb{R}^{d+1}$ of $\boldsymbol{\Delta}(t)$, set

$$\overline{\boldsymbol{z}} := (1, \boldsymbol{z}^\intercal)^\intercal, \qquad \pi^*_\rho(\boldsymbol{z}, t) := \hat{q}_\rho(\boldsymbol{z}; \boldsymbol{\theta}^*(t)), \tag{69}$$

$$\mathcal{M}_t(\boldsymbol{z})[\boldsymbol{\Delta}(t)] := \sum_{\rho=1}^{k-1} \pi^*_\rho(\boldsymbol{z}, t)\{\boldsymbol{x}(\rho) - \boldsymbol{m}(\boldsymbol{z}, t)\} \left\langle \boldsymbol{\Delta}_\rho(t), \overline{\boldsymbol{z}} \right\rangle. \tag{70}$$

Then, as $\|\boldsymbol{\Delta}\|_\infty \to 0$ over bounded measurable perturbations supported on I,

$$\mathrm{P}_{\boldsymbol{\Delta}} = \sum_{\sigma=0}^{k-1} \{q(\sigma) + \mathcal{L}_\sigma(\boldsymbol{\Delta})\}\delta_{\boldsymbol{x}(\sigma)} + O(\|\boldsymbol{\Delta}\|_\infty^2), \tag{71}$$

where

$$\mathcal{L}_\sigma(\boldsymbol{\Delta}) = \int_I \langle \boldsymbol{L}_\sigma(t), \boldsymbol{\Delta}(t) \rangle \, \mathrm{d}t, \tag{72}$$

$$\langle \boldsymbol{L}_\sigma(t), \boldsymbol{D} \rangle = -\int_{\mathcal{A}_\sigma(t)} \nabla_{\boldsymbol{z}} \cdot \{c_1(t) p_t(\boldsymbol{z}) \mathcal{M}_t(\boldsymbol{z})[\boldsymbol{D}]\} \, \mathrm{d}\boldsymbol{z}. \tag{73}$$

or equivalently

$$[\boldsymbol{L}_\sigma(t)]_{j,\rho} = -c_1(t) \int_{\mathcal{A}_\sigma(t)} \nabla_{\boldsymbol{z}} \cdot \left\{ p_t(\boldsymbol{z}) \pi^*_\rho(\boldsymbol{z}, t)\{\boldsymbol{x}(\rho) - \boldsymbol{m}(\boldsymbol{z}, t)\}\overline{z}_j \right\} \mathrm{d}\boldsymbol{z}, \tag{74}$$

for $0 \le j \le d$ and $1 \le \rho \le k-1$, where $\overline{z}_0 = 1$ and $\overline{z}_j = z_j$ for $j \ge 1$. The integrals are absolutely convergent, and $\boldsymbol{L}_\sigma \in C(I; E)$ is bounded. Further

$$\sum_{\sigma=0}^{k-1} \mathcal{L}_\sigma(\boldsymbol{\Delta}) = 0, \qquad \sum_{\sigma=0}^{k-1} \boldsymbol{L}_\sigma(t) = 0, \qquad t \in I. \tag{75}$$

The proof is in Appendix J.5.5.

Example 1 (Binary variables)**.** *For $d = 1$, $k = 2$, $\boldsymbol{x}(0) = -1$, $\boldsymbol{x}(1) = 1$, and $q(1) = p$, the denoiser is*

$$\hat{\boldsymbol{m}}(z; \boldsymbol{\theta}) = \tanh\left(\frac{\theta_{1,0} + \theta_1^+ z}{2} \right). \tag{76}$$

If $q(+1) = p = 1 - q(-1)$, and F_t is the scalar mixture distribution function:

$$F_t(x) = (1-p)\Phi\left(\frac{x + a(t)}{\sigma(t)} \right) + p\Phi\left(\frac{x - a(t)}{\tau(t)} \right), \tag{77}$$

the two basins are $\mathcal{A}_-(t) = (-\infty, \gamma(t))$ and $\mathcal{A}_+(t) = (\gamma(t), \infty)$ with $\gamma(t) = F_t^{-1}(1-p)$.

Writing $\eta_*(z, t) = (\theta^*_{1,0}(t) + \theta^{*,+}_1(t) z)/2$, integrating the spatial derivative in equation 74 over this half-line gives

$$\boldsymbol{L}_1(t) = \frac{1}{2} c_1(t) p_t(\gamma(t)) \operatorname{sech}^2\left((\theta^*_{1,0}(t) + \theta^{*,+}_1(t) z)/2 \right) \begin{pmatrix} 1 \\ \gamma(t) \end{pmatrix}. \tag{78}$$

J.3 UNDERSMOOTHING REGIME

Define the intrinsic fixed-time sandwich operator

$$\boldsymbol{Q}^*(t) := \boldsymbol{H}^*(t)^{-1} \circ \boldsymbol{J}^*(t) \circ \boldsymbol{H}^*(t)^{-1}, \tag{79}$$

where $\boldsymbol{H}^*(t)$, $\boldsymbol{J}^*(t)$ are the Hessian and covariance defined in Eqs. equation 58 and equation 59.

Proposition J.5 (Shrinking-bandwidth empirical minimizers)**.** *Under Assumptions 1, 4, 5, and 3, let $I \subset (0,T)$ be compact. Set $b = b_n$ and $e_n = \sqrt{\log(1/b)/(nb)}$. With probability tending to one, the empirical minimizers are unique and belong to $\mathrm{int}(K(t))$ simultaneously on I, and, defining $\widehat{\boldsymbol{S}}_n(t,\boldsymbol{\theta})$ as in Eq. equation 65,*

$$\|\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}^b\|_\infty = O_{\mathrm{P}}(e_n), \tag{80}$$

$$\hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}^b(t) = -\boldsymbol{H}_b(t)^{-1}[\widehat{\boldsymbol{S}}_n(t,\boldsymbol{\theta}^b(t))] + \boldsymbol{r}_n(t), \tag{81}$$

$$\|\boldsymbol{r}_n\|_\infty = O_{\mathrm{P}}(e_n^2) = o_{\mathrm{P}}(n^{-1/2}). \tag{82}$$

The proof is in Appendix J.5.6.

Proposition J.6 (Integrated parameter central limit theorem)**.** *Under Assumptions 1, 3, 4, and 5, let $I \subset (0,T)$ be compact. Suppose λ is C^2 on an open neighborhood of I, and $(u,\boldsymbol{\theta}) \mapsto R_u(\boldsymbol{\theta})$ is C^3 on an open neighborhood of $\{(t,\boldsymbol{\theta}^*(t)) : t \in I\}$ in $(0,T)\times E$. For every $\boldsymbol{\phi} \in C(I;\mathbb{R}^{(d+1)\times(k-1)})$,*

$$\sqrt{n}\int_I \Big\langle \hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}^*(t), \boldsymbol{\phi}(t)\Big\rangle \,\mathrm{d}t \Rightarrow \mathsf{N}\big(0, Q^*(\boldsymbol{\phi})\big), \tag{83}$$

$$Q^*(\boldsymbol{\phi}) := \int_I \frac{1}{\lambda(t)} \,\langle \boldsymbol{\phi}(t), \boldsymbol{Q}^*(t)[\boldsymbol{\phi}(t)]\rangle \,\mathrm{d}t, \tag{84}$$

The conclusion holds jointly for every finite collection of test functions; the limiting covariance associated with $\boldsymbol{\phi}$ and $\widetilde{\boldsymbol{\phi}}$ is

$$Q^*(\boldsymbol{\phi},\widetilde{\boldsymbol{\phi}}) := \int_I \frac{1}{\lambda(t)} \Big\langle \boldsymbol{\phi}(t), \boldsymbol{Q}^*(t)[\widetilde{\boldsymbol{\phi}}(t)]\Big\rangle \,\mathrm{d}t. \tag{85}$$

The proof is in Appendix J.5.7.

For convenience, recall that, with $\Sigma \sim q$ and $\boldsymbol{G} \sim \mathsf{N}(0,\boldsymbol{I}_d)$ independent,

$$\boldsymbol{H}^*(t) := \nabla^2_{\boldsymbol{\theta}} R_t(\boldsymbol{\theta}^*(t)), \tag{86}$$

$$\boldsymbol{J}^*(t) := \mathbb{E}_{\Sigma,\boldsymbol{G}}\left[\boldsymbol{\psi}_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G};\boldsymbol{\theta}^*(t)\big) \otimes \boldsymbol{\psi}_\Sigma\big(a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G};\boldsymbol{\theta}^*(t)\big)\right], \tag{87}$$

$$\boldsymbol{Q}^*(t) := \boldsymbol{H}^*(t)^{-1} \circ \boldsymbol{J}^*(t) \circ \boldsymbol{H}^*(t)^{-1}. \tag{88}$$

We are now in position to state a formal version of Theorem 1, given by the next theorem and corollary.

Theorem 2 (Joint limit for generated class probabilities)**.** *Under Assumptions 2, 3, 4, and 6, suppose Assumptions 1 and 5 hold on every compact subinterval of $(0,T)$ for the same domain map $t \mapsto K(t)$. Suppose λ is C^2 on $(0,T)$, and $(u,\boldsymbol{\theta}) \mapsto R_u(\boldsymbol{\theta})$ is C^3 on an open neighborhood of $\{(t,\boldsymbol{\theta}^*(t)) : t \in I\}$ for every compact $I \subset (0,T)$. Let $\boldsymbol{L}_\sigma(t)$ be the continuous kernels in Eq. equation 74.*

Use $\hat{\boldsymbol{\theta}}_b(t)$ on all of $(0,T)$ in the probability flow ODE equation 23. Let P_n be its terminal law conditional on the training data, and set $\hat{q}_n(\sigma) := \mathrm{P}_n(\{\boldsymbol{x}(\sigma)\})$. For $0 < \varepsilon < T/2$, let $I_\varepsilon = [\varepsilon, T-\varepsilon]$ and let $\mathrm{P}_{n,\varepsilon}$ be the terminal law obtained by using $\hat{\boldsymbol{\theta}}_b(t)$ on I_ε and $\boldsymbol{\theta}^*(t)$ outside I_ε. Assume that the endpoint contributions are negligible at the central-limit scale: for every $\eta > 0$,

$$\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}\,\|\mathrm{P}_n - \mathrm{P}_{n,\varepsilon}\|_{\mathrm{TV}} > \eta\right) = 0. \tag{89}$$

Assume also that, for every $0 \le \sigma \le k-1$,

$$\int_0^T \frac{1}{\lambda(t)} \langle \boldsymbol{L}_\sigma(t), \boldsymbol{Q}^*(t)[\boldsymbol{L}_\sigma(t)]\rangle \,\mathrm{d}t < \infty. \tag{90}$$

For $0 \le \sigma, \alpha \le k-1$, define

$$V_{\sigma\alpha}(\lambda) := \int_0^T \frac{1}{\lambda(t)} \langle \boldsymbol{L}_\sigma(t), \boldsymbol{Q}^*(t)[\boldsymbol{L}_\alpha(t)] \rangle \, \mathrm{d}t. \tag{91}$$

Then

$$\sqrt{n}(\hat{q}_n - q) \Rightarrow \mathsf{N}_k(0, \boldsymbol{V}(\lambda)). \tag{92}$$

Moreover, $\boldsymbol{V}(\lambda)\mathbf{1} = 0$.

The proof is in Appendix J.5.8. The theorem concerns conditional model probabilities, not empirical frequencies from an additional finite batch of generated samples.

Corollary J.7 (Efficiency of cross-entropy)**.** *For $t \in (0,T)$ and $\boldsymbol{\vartheta}$ in a neighborhood of $\boldsymbol{\theta}^*(t)$, let $P_{t,\boldsymbol{\vartheta}}$ be the conditional softmax model under which $\boldsymbol{Y}$ has its true marginal density p_t and*

$$P_{t,\boldsymbol{\vartheta}}(\Sigma = \sigma \mid \boldsymbol{Y} = \boldsymbol{y}) = \hat{q}_\sigma(\boldsymbol{y}; \boldsymbol{\vartheta}). \tag{93}$$

Consider any loss ℓ satisfying the assumptions of Theorem 2 and the local Fisher-consistency condition

$$\mathbb{E}_{t,\boldsymbol{\vartheta}} \left[\nabla_{\boldsymbol{\theta}} \ell_\Sigma(\boldsymbol{Y}; \boldsymbol{\vartheta}) \right] = 0 \tag{94}$$

for every such $(t, \boldsymbol{\vartheta})$, with differentiation under the expectation permitted. For fixed λ, let $\boldsymbol{V}^\ell(\lambda)$ be the covariance in the theorem for this loss and let $\boldsymbol{V}^{\mathrm{CE}}(\lambda)$ be the one for cross-entropy. Then

$$\boldsymbol{V}^{\mathrm{CE}}(\lambda) \preceq \boldsymbol{V}^\ell(\lambda), \tag{95}$$

where $\preceq$ is the Loewner order.

The proof is in Appendix J.5.9.

J.4 Choice of the time-sampling density

Proposition J.8 (Optimal time-sampling density)**.** *Fix a positive semidefinite $k \times k$ matrix $\boldsymbol{W}$. For an admissible time-sampling density λ, let $\boldsymbol{\xi}_\lambda \sim \mathsf{N}_k(0, \boldsymbol{V}(\lambda))$ and define*

$$C_{\sigma\alpha}(t) := \langle \boldsymbol{L}_\sigma(t), \boldsymbol{Q}^*(t)[\boldsymbol{L}_\alpha(t)] \rangle, \tag{96}$$

$$h_W(t) := \mathrm{tr}\{\boldsymbol{W}\boldsymbol{C}(t)\}, \tag{97}$$

$$\mathcal{R}_W(\lambda) := \mathbb{E}\left[\langle \boldsymbol{\xi}_\lambda, \boldsymbol{W}\boldsymbol{\xi}_\lambda \rangle\right] = \mathrm{tr}\{\boldsymbol{W}\boldsymbol{V}(\lambda)\} = \int_0^T \frac{h_W(t)}{\lambda(t)} \, \mathrm{d}t. \tag{98}$$

Under the assumptions of Theorem 2,

$$n \, \langle \hat{q}_n - q, \boldsymbol{W}(\hat{q}_n - q) \rangle \Rightarrow \langle \boldsymbol{\xi}_\lambda, \boldsymbol{W}\boldsymbol{\xi}_\lambda \rangle. \tag{99}$$

If $H_W := \int_0^T \sqrt{h_W(t)} \, \mathrm{d}t < \infty$, then

$$\inf_{\lambda \ge 0:\, \int_0^T \lambda = 1} \mathcal{R}_W(\lambda) = H_W^2. \tag{100}$$

If $0 < H_W < \infty$, the minimizing density is

$$\lambda_{\mathrm{opt}}(t) = \frac{\sqrt{h_W(t)}}{H_W}. \tag{101}$$

The minimum is attained whenever λ_{opt} is admissible for Theorem 2; otherwise the right-hand side of equation 100 is the infimum over admissible positive approximations to λ_{opt}.

Proposition J.9 (SNR-reparametrization invariance)**.** *Suppose the loss is cross-entropy or squared denoising loss and set $\rho(t) := a(t)/\tau(t)$. If ρ is C^1 and strictly monotone on $(0,T)$, with $\dot{\rho}(t) \neq 0$, then there is a positive semidefinite matrix function $\widetilde{\boldsymbol{C}}$ depending only on the normalized signal-to-noise coordinate $r = \rho(t)$ such that*

$$\boldsymbol{C}(t) = \dot{\rho}(t)^2 \widetilde{\boldsymbol{C}}(\rho(t)). \tag{102}$$

Consequently the optimal Gaussian-limit risk is

$$\inf_\lambda \mathcal{R}_W(\lambda) = \left\{ \int_{\rho((0,T))} \sqrt{\mathrm{tr}\{\boldsymbol{W}\widetilde{\boldsymbol{C}}(r)\}} \, \mathrm{d}r \right\}^2, \tag{103}$$

whenever the improper integral is finite. It therefore depends on the schedule only through its SNR range and is invariant under monotone time reparametrization.

The proofs are in Appendices J.5.10 and J.5.11.

J.5 Proofs for Sections J.1 and J.2

J.5.1 Proof of Lemma J.1

Note that by construction $\boldsymbol{m}(\boldsymbol{z}, t) = \hat{\boldsymbol{m}}(\boldsymbol{z}; \boldsymbol{\theta}^*(t))$. On every compact time interval below T, the velocity $\boldsymbol{h}_*$ has linear growth and bounded spatial derivatives of every positive order. Indeed, $\boldsymbol{h}_* = c_1\boldsymbol{m} + c_2\boldsymbol{z}$, the embeddings are bounded, the posterior is a softmax with bounded slope parameters, and $\inf \tau > 0$ on such an interval. Thus the flow and its inverse are smooth spatial diffeomorphisms. The mixture density equation 50 solves the continuity equation for $\boldsymbol{h}_*$, so a trajectory initialized with density p_s has density p_t at every $s \le t < T$.

For each fixed t, represent the marginal as $\boldsymbol{Y}_t = a(t)\boldsymbol{X} + \tau(t)\boldsymbol{G}_t$, where $\boldsymbol{G}_t \sim \mathsf{N}(0, \boldsymbol{I}_d)$ is independent of $\boldsymbol{X}$. Since $c_2 = \dot{\tau}/\tau$ and $c_1 = \dot{a} - a\dot{\tau}/\tau$, we have

$$\boldsymbol{h}_*(\boldsymbol{z}, t) = \dot{a}(t)\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}_t = \boldsymbol{z}] + \dot{\tau}(t)\mathbb{E}[\boldsymbol{G}_t \mid \boldsymbol{Y}_t = \boldsymbol{z}]. \tag{104}$$

Positive noise at zero implies $s(0) > 0$. Since s is non-decreasing, $\tau = \tau_0\sqrt{s}$ is smooth on $[0, T]$, and both $\dot{a}$ and $\dot{\tau}$ are integrable. For a true trajectory initialized with density p_s, conditional Jensen and Fubini-Tonelli give

$$\mathbb{E}\int_s^T \|\boldsymbol{h}_*(\boldsymbol{Z}_t, t)\| \,\mathrm{d}t \le \mathbb{E}\|\boldsymbol{X}\| \int_s^T |\dot{a}(t)| \,\mathrm{d}t + \mathbb{E}\|\boldsymbol{G}_t\| \int_s^T |\dot{\tau}(t)| \,\mathrm{d}t < \infty. \tag{105}$$

Therefore, almost every trajectory $t \mapsto \boldsymbol{Z}_t$ has finite length and hence has a terminal limit $\boldsymbol{Z}_T = \lim_{t\uparrow T} \boldsymbol{Z}_t$. Its law is $\boldsymbol{Z}_T \sim \mathbb{P}_q$, because $p_t(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ converges weakly to $\mathbb{P}_q$ as $t \uparrow T$. Since p_s is everywhere positive, the exceptional set of starting points is also Lebesgue-null.

The limit conditions defining each basin are Borel because the trajectories are continuous before T. The distinct embeddings make the basins disjoint. Flow composition and invertibility imply the set identity in equation 52. Applying the terminal-law conclusion to a trajectory initialized with p_t gives the mass identity in that equation.

J.5.2 Proof of Proposition J.2

We begin with Assumption 1. Equal embedding norms imply the posterior log-odds

$$\log \frac{\mathbb{P}(\Sigma = \rho \mid \boldsymbol{X}_t = \boldsymbol{y})}{\mathbb{P}(\Sigma = 0 \mid \boldsymbol{X}_t = \boldsymbol{y})} = -r(\rho) + \frac{a(t)}{\tau(t)^2}\langle \boldsymbol{x}(\rho) - \boldsymbol{x}(0), \boldsymbol{y}\rangle. \tag{106}$$

Thus $\pi^*(\boldsymbol{y}, t) := (\hat{q}_\sigma(\boldsymbol{y}; \boldsymbol{\theta}^*(t)))_{\sigma=0}^{k-1}$ is the true posterior. Conditional expectation gives the excess risks

$$R_t^{\mathrm{CE}}(\boldsymbol{\theta}) - R_t^{\mathrm{CE}}(\boldsymbol{\theta}^*) = \mathbb{E}\left[\sum_{\sigma=0}^{k-1} \pi_\sigma^*(\boldsymbol{Y}, t) \log \frac{\pi_\sigma^*(\boldsymbol{Y}, t)}{\hat{q}_\sigma(\boldsymbol{Y}; \boldsymbol{\theta})}\right], \tag{107}$$

$$R_t^{\mathrm{SM}}(\boldsymbol{\theta}) - R_t^{\mathrm{SM}}(\boldsymbol{\theta}^*) = \mathbb{E}\|\hat{\boldsymbol{m}}(\boldsymbol{Y}; \boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{Y}, t)\|^2. \tag{108}$$

Hence, $R_t^{\mathrm{L}}(\boldsymbol{\theta}) - R_t^{\mathrm{L}}(\boldsymbol{\theta}^*) \ge 0$ for $\mathrm{L} \in \{\mathrm{CE}, \mathrm{SM}\}$.

For cross-entropy $R_t^{\mathrm{CE}}(\boldsymbol{\theta}) - R_t^{\mathrm{CE}}(\boldsymbol{\theta}^*) = 0$ $\hat{q}_\sigma(\boldsymbol{Y}; \boldsymbol{\theta}) = hq_\sigma(\boldsymbol{Y}; \boldsymbol{\theta}_*)$ for almost every $\boldsymbol{Y}, \sigma$, which in turns implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. For squared loss, $R_t^{\mathrm{SM}}(\boldsymbol{\theta}) - R_t^{\mathrm{SM}}(\boldsymbol{\theta}^*) = 0$ implies $\hat{\boldsymbol{m}}(\boldsymbol{Y}; \boldsymbol{\theta}) = \boldsymbol{m}(\boldsymbol{Y}, t)$ for almost every $\boldsymbol{Y}$, which in turns implies $\hat{q}_\sigma(\boldsymbol{Y}; \boldsymbol{\theta}) = hq_\sigma(\boldsymbol{Y}; \boldsymbol{\theta}_*)$ because of Eq. equation 55, whence $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

We now pass to Assumption 5. For $\boldsymbol{A} \in E$, let $v_0 = 0$ and $v_\rho = \langle \boldsymbol{A}_\rho, \overline{\boldsymbol{y}}\rangle$, $\overline{\boldsymbol{y}} = (1, \boldsymbol{y}^\intercal)^\intercal$. Let $\mathcal{M}_t(\boldsymbol{y}) = D_{\boldsymbol{\theta}}\hat{\boldsymbol{m}}(\boldsymbol{y}; \boldsymbol{\theta}^*(t))$. The Hessian quadratic forms at the optimum are

$$\langle \boldsymbol{A}, \boldsymbol{H}^{*,\mathrm{CE}}(t)[\boldsymbol{A}]\rangle = \mathbb{E}\left[\sum_{\rho=1}^{k-1} \pi_\rho^* v_\rho^2 - \left(\sum_{\rho=1}^{k-1} \pi_\rho^* v_\rho\right)^2\right], \tag{109}$$

$$\langle \boldsymbol{A}, \boldsymbol{H}^{*,\mathrm{SM}}(t)[\boldsymbol{A}]\rangle = 2\mathbb{E}\|\mathcal{M}_t(\boldsymbol{Y})[\boldsymbol{A}]\|^2. \tag{110}$$

All posterior coordinates are positive, whence the conditional variance in the first line vanishes only if all v_ρ equal $v_0 = 0$.

Under the affine independence assumption Eq. equation 55 the differential of the mean with respect to probability perturbations whose coordinates sum to zero has full rank $k-1$. Hence the second line also vanishes only if all $v_\rho = 0$ almost everywhere. Full support of $\boldsymbol{Y}$ then gives $\boldsymbol{A} = 0$. Both Hessians are positive definite; continuity and compactness of J make this uniform. This proves point 3 of Assumption 5.

The unique population minimizer over E is therefore also the unique minimizer over every $K(t)$ containing $\boldsymbol{\theta}^*(t)$. The compact-graph and uniform-margin hypotheses give precisely the constrained-domain alternative in Assumption 5; cross-entropy also satisfies its unconstrained convex alternative. This proves point 1 of Assumption 5.

To prove point 2, note that derivatives of softmax with respect to logits are bounded at every fixed order. Each parameter differentiation supplies at most a factor $\|\overline{\boldsymbol{y}}\|$. Consequently both losses and their first three parameter derivatives have polynomial envelopes, with third derivative bounded by $C(1+\|\boldsymbol{y}\|)^3$. The cross-entropy loss itself grows at most linearly on K_0 and is convex on E; squared loss is bounded but may be nonconvex, which explains the compact time-dependent domains. Gaussian moments justify differentiation and continuity under expectation. Differentiating $\boldsymbol{y} = a(t)\boldsymbol{x}(\Sigma) + \tau(t)\boldsymbol{G}$ also gives polynomial Gaussian envelopes for three joint derivatives if a, τ are C^3. This proves all the claims.

For completeness, the score operators used in the covariance are particularly explicit for these losses:

$$[\boldsymbol{\psi}_\sigma^{\mathrm{CE}}(\boldsymbol{y};\boldsymbol{\theta})]_\rho = (\hat{q}_\rho(\boldsymbol{y};\boldsymbol{\theta}) - \mathbf{1}_{\{\sigma=\rho\}})\overline{\boldsymbol{y}}, \tag{111}$$

$$\boldsymbol{\psi}_\sigma^{\mathrm{SM}}(\boldsymbol{y};\boldsymbol{\theta}) = 2(D_{\boldsymbol{\theta}}\hat{\boldsymbol{m}}(\boldsymbol{y};\boldsymbol{\theta}))^*[\hat{\boldsymbol{m}}(\boldsymbol{y};\boldsymbol{\theta}) - \boldsymbol{x}(\sigma)]. \tag{112}$$

Here the superscript $*$ on the differential denotes its adjoint. For cross-entropy the conditional score covariance equals its Hessian, so $\boldsymbol{J}^{*,\mathrm{CE}} = \boldsymbol{H}^{*,\mathrm{CE}}$ and $\boldsymbol{Q}^{*,\mathrm{CE}} = (\boldsymbol{H}^{*,\mathrm{CE}})^{-1}$.

J.5.3 Zero-bandwidth limit of the population risk

Lemma J.10. *Under Assumptions 1, 4, and 5, define*

$$d_b(t) := \int_0^T w_b(u-t)\lambda(u)\,\mathrm{d}u, \qquad \overline{R}_{b,t} := R_{b,t}/d_b(t). \tag{113}$$

Then $c_- b \le d_b(t) \le c_+ b$ uniformly on I for all sufficiently small b, and

$$\lim_{b\to 0} \max_{0\le j\le 2} \sup_{t\in I,\,\boldsymbol{\theta}\in K_0} \|D^j\overline{R}_{b,t}(\boldsymbol{\theta}) - D^j R_t(\boldsymbol{\theta})\| = 0. \tag{114}$$

The minimizers $\boldsymbol{\theta}^b(t)$ are unique and belong to $\mathrm{int}(K(t))$ for all sufficiently small b, uniformly on I, and

$$\lim_{b\to 0} \|\boldsymbol{\theta}^b - \boldsymbol{\theta}^*\|_\infty = 0, \tag{115}$$

$$\lim_{b\to 0} \sup_{t\in I} \left\| \frac{\boldsymbol{H}_b(t)}{d_b(t)} - \boldsymbol{H}^*(t) \right\|_{\mathrm{op}} = 0. \tag{116}$$

For every fixed neighborhood of the target curve, the population risk has a uniform positive gap over feasible points outside that neighborhood; in the unconstrained case the corresponding gap holds on K_0. If $I \Subset (0,T)$, then also, uniformly over I,

$$\boldsymbol{H}_b(t)^{-1} = \frac{1}{b\mu_0\lambda(t)}\boldsymbol{H}^*(t)^{-1} + o(b^{-1})\,. \tag{117}$$

Proof. The Gaussian representation and the polynomial upper bounds in Assumption 5 imply joint continuity of $D^j R_t(\boldsymbol{\theta})$ on $J \times K_0$ for $j \le 2$. The bounds on d_b follow from local positivity of λ, compactness of I, and $w_1(0) > 0$. Since $\overline{R}_{b,t}$ is a normalized weighted average of R_u over $|u-t| \le b$, uniform continuity gives equation 114, including when $0 \in I$.

In the constrained case, compactness of $\mathcal{K}_J$ (as defined in Assumption 5), continuity, and uniqueness of the minimizer give, for every $\delta > 0$, a positive gap for $R_t(\boldsymbol{\theta}) - R_t(\boldsymbol{\theta}^*(t))$ uniformly over

$$\{(t,\boldsymbol{\theta}) : t \in I,\ \boldsymbol{\theta} \in K(t),\ \|\boldsymbol{\theta} - \boldsymbol{\theta}^*(t)\| \ge \delta\}. \tag{118}$$

In the unconstrained case, the same compactness argument gives the gap on $I \times K_0$. The $j = 0$ part of equation 114 therefore places every constrained minimizer, and every minimizer localized to K_0 in the unconstrained case, in an arbitrarily small neighborhood of the target curve.

The $j = 2$ convergence, equation 49, and the uniform interior margin give strict convexity on fixed-radius balls around the target curve. Thus the constrained minimizers are unique and interior. In the unconstrained case, convexity makes the localized interior stationary point global, while local strict convexity makes it the unique global minimizer. This proves uniform convergence of $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^*$; evaluating the uniform Hessian convergence along these minimizers proves equation 116.

Finally, if $I \Subset (0, T)$, the kernel is not truncated by the temporal boundary and $d_b(t)/b \to \mu_0 \lambda(t)$ uniformly. Combining this with equation 116 and continuity of matrix inversion on uniformly positive definite operators proves equation 117. □

J.5.4 Proof of Proposition J.3

Lemma J.10 gives unique localized minimizers converging uniformly to $\boldsymbol{\theta}^*$. Since $I \Subset (0, T)$, the localized score equation can be written

$$0 = b^{-1} \nabla R_{b,t}(\boldsymbol{\theta}^b(t)) = \mathbf{F}(b, t, \boldsymbol{\theta}^b(t)), \tag{119}$$

$$\mathbf{F}(b, t, \boldsymbol{\theta}) := \int_{-1}^{1} w_1(v) \lambda(t + bv) \boldsymbol{S}(t + bv, \boldsymbol{\theta}) \, \mathrm{d}v \,, \tag{120}$$

where $\boldsymbol{S}(t, \boldsymbol{\theta})$ is defined in Eq. equation 67. This is C^2 jointly near the zero-bandwidth curve, and $D_{\boldsymbol{\theta}} \mathbf{F}(0, t, \boldsymbol{\theta}^*(t)) = \mu_0 \lambda(t) \boldsymbol{H}^*(t)$ is uniformly invertible. The implicit function theorem gives a C^2 solution in b, uniformly over I, identified with $\boldsymbol{\theta}^b$ for positive b by uniqueness. Kernel symmetry makes $\mathbf{F}$ and its local solution even in b, so $\partial_b \boldsymbol{\theta}^b|_{b=0} = 0$. Differentiating twice gives

$$\partial_b^2 \boldsymbol{\theta}^b(t)\big|_{b=0} = -\frac{\mu_2}{\mu_0 \lambda(t)} \boldsymbol{H}^*(t)^{-1} \left[\partial_u^2 (\lambda(u) \boldsymbol{S}(u, \boldsymbol{\theta}))\big|_{u=t,\, \boldsymbol{\theta}=\boldsymbol{\theta}^*(t)} \right]. \tag{121}$$

Taylor's theorem with a uniform remainder proves equation 66–equation 67.

J.5.5 Proof of Proposition J.4

As pointed out in previous proofs already, the posterior log-odds satisfy

$$\log \frac{\mathbb{P}(\Sigma = \rho \mid \boldsymbol{X}_t = \boldsymbol{z})}{\mathbb{P}(\Sigma = 0 \mid \boldsymbol{X}_t = \boldsymbol{z})} = -r(\rho) + \frac{a(t)}{\tau(t)^2} \langle \boldsymbol{x}(\rho) - \boldsymbol{x}(0), \boldsymbol{z} \rangle \,. \tag{122}$$

Thus $\pi_\rho^*(\boldsymbol{z}, t) = \mathbb{P}(\Sigma = \rho \mid \boldsymbol{X}_t = \boldsymbol{z})$, $\hat{\boldsymbol{m}}(\boldsymbol{z}, t; \boldsymbol{\theta}^*(t)) = \boldsymbol{m}(\boldsymbol{z}, t)$, and $\boldsymbol{h}_*$ is the true probability-flow velocity with density p_t.

We next compute the differential of the denoiser. For $\rho \geq 1$, the perturbation of the ρ-th logit at $(\boldsymbol{z}, t)$ is

$$\mathrm{d}\eta_\rho = \langle \boldsymbol{\Delta}_\rho(t), \overline{\boldsymbol{z}} \rangle \,, \qquad \mathrm{d}\eta_0 = 0. \tag{123}$$

The softmax differential is

$$D\pi_\alpha^*[\boldsymbol{\Delta}] = \pi_\alpha^* \left\{ \mathrm{d}\eta_\alpha - \sum_{\rho=1}^{k-1} \pi_\rho^* \mathrm{d}\eta_\rho \right\}. \tag{124}$$

Since $\boldsymbol{m} = \sum_{\alpha=0}^{k-1} \pi_\alpha^* \boldsymbol{x}(\alpha)$, summing the last display against $\boldsymbol{x}(\alpha)$ gives

$$\begin{aligned} D_{\boldsymbol{\theta}} \hat{\boldsymbol{m}}(\boldsymbol{z}, t; \boldsymbol{\theta}^*(t))[\boldsymbol{\Delta}(t)] &= \sum_{\rho=1}^{k-1} \pi_\rho^*(\boldsymbol{z}, t) \{ \boldsymbol{x}(\rho) - \boldsymbol{m}(\boldsymbol{z}, t) \} \langle \boldsymbol{\Delta}_\rho(t), \overline{\boldsymbol{z}} \rangle \\ &= \mathcal{M}_t(\boldsymbol{z})[\boldsymbol{\Delta}(t)]. \end{aligned} \tag{125}$$

Write $I = [t_-, t_+]$, $\varepsilon = \|\boldsymbol{\Delta}\|_\infty$, and $\boldsymbol{h}_{\eta\boldsymbol{\Delta}}$ for the velocity with parameters $\boldsymbol{\theta}^* + \eta\boldsymbol{\Delta}$, $0 \leq \eta \leq 1$. Outside I all these velocities equal $\boldsymbol{h}_*$. The first velocity variation is

$$\boldsymbol{v}_t[\boldsymbol{\Delta}](\boldsymbol{z}) := \partial_\eta \boldsymbol{h}_{\eta\boldsymbol{\Delta}}(\boldsymbol{z}, t)\big|_{\eta=0} = c_1(t) \mathcal{M}_t(\boldsymbol{z})[\boldsymbol{\Delta}(t)], \tag{126}$$

$$\boldsymbol{h}_{\boldsymbol{\Delta}}(\boldsymbol{z}, t) - \boldsymbol{h}_*(\boldsymbol{z}, t) = \boldsymbol{v}_t[\boldsymbol{\Delta}](\boldsymbol{z}) + O\big(\varepsilon^2 (1 + \|\boldsymbol{z}\|)^2\big). \tag{127}$$

For sufficiently small ε, all parameters $\boldsymbol{\theta}^*(t) + \eta\boldsymbol{\Delta}(t)$ lie in a fixed compact neighborhood of the target curve. Write

$$\boldsymbol{h}_{\eta\boldsymbol{\Delta}}(\boldsymbol{z}, t) = c_2(t)\boldsymbol{z} + \boldsymbol{b}_{\eta,t}(\boldsymbol{z}), \qquad \boldsymbol{b}_{\eta,t}(\boldsymbol{z}) := c_1(t)\hat{\boldsymbol{m}}(\boldsymbol{z}, t; \boldsymbol{\theta}^*(t) + \eta\boldsymbol{\Delta}(t)). \tag{128}$$

The softmax probabilities lie in $[0, 1]$, the embeddings are fixed, and the slope parameters range over a compact set. Consequently, for every fixed spatial order $r \geq 0$,

$$\sup_{t\in I,\, \eta\in[0,1],\, \boldsymbol{z}\in\mathbb{R}^d} \|D_{\boldsymbol{z}}^r \boldsymbol{b}_{\eta,t}(\boldsymbol{z})\| \leq C_r, \tag{129}$$

$$\|\partial_\eta^j D_{\boldsymbol{z}}^r \boldsymbol{b}_{\eta,t}(\boldsymbol{z})\| \leq C_r \varepsilon^j (1 + \|\boldsymbol{z}\|)^j, \qquad j = 1, 2. \tag{130}$$

Thus the velocities are measurable in time, smooth in space, and globally Lipschitz in space, with constants uniform in t and η.

Let $\boldsymbol{\Phi}^\eta_{t_-,t}$ be the forward flow of $\boldsymbol{h}_{\eta\boldsymbol{\Delta}}$ and let $\Psi^\eta_t := (\boldsymbol{\Phi}^\eta_{t_-,t})^{-1}$. If $\boldsymbol{X}^\eta_t = \boldsymbol{\Phi}^\eta_{t_-,t}(\boldsymbol{x})$, variation of constants gives

$$\boldsymbol{X}^\eta_t = \alpha_t \boldsymbol{x} + \alpha_t \int_{t_-}^t \alpha_u^{-1} \boldsymbol{b}_{\eta,u}(\boldsymbol{X}^\eta_u)\, \mathrm{d}u, \qquad \alpha_t := \exp\left\{\int_{t_-}^t c_2(u)\, \mathrm{d}u\right\} = \frac{\tau(t)}{\tau(t_-)}. \tag{131}$$

The factors α_t are bounded above and away from zero on I, and equation 129 therefore yields

$$\sup_{t\in I,\, \eta\in[0,1],\, \boldsymbol{z}\in\mathbb{R}^d} \|\Psi^\eta_t(\boldsymbol{z}) - \alpha_t^{-1}\boldsymbol{z}\| \leq C, \qquad 1 + \|\Psi^\eta_t(\boldsymbol{z})\| \asymp 1 + \|\boldsymbol{z}\|. \tag{132}$$

For completeness, the spatial Jacobian $\boldsymbol{A}^\eta_t := D_{\boldsymbol{x}}\boldsymbol{X}^\eta_t$ and the first two parameter variations $\boldsymbol{Y}^\eta_t := \partial_\eta \boldsymbol{X}^\eta_t$ and $\boldsymbol{W}^\eta_t := \partial_\eta^2 \boldsymbol{X}^\eta_t$ satisfy

$$\dot{\boldsymbol{A}}^\eta_t = D_{\boldsymbol{z}}\boldsymbol{h}_{\eta\boldsymbol{\Delta}}(\boldsymbol{X}^\eta_t, t)\boldsymbol{A}^\eta_t, \qquad \boldsymbol{A}^\eta_{t_-} = \boldsymbol{I}_d, \tag{133}$$

$$\dot{\boldsymbol{Y}}^\eta_t = D_{\boldsymbol{z}}\boldsymbol{h}_{\eta\boldsymbol{\Delta}}\boldsymbol{Y}^\eta_t + \partial_\eta \boldsymbol{h}_{\eta\boldsymbol{\Delta}}, \qquad \boldsymbol{Y}^\eta_{t_-} = 0, \tag{134}$$

$$\dot{\boldsymbol{W}}^\eta_t = D_{\boldsymbol{z}}\boldsymbol{h}_{\eta\boldsymbol{\Delta}}\boldsymbol{W}^\eta_t + D^2_{\boldsymbol{z}}\boldsymbol{h}_{\eta\boldsymbol{\Delta}}[\boldsymbol{Y}^\eta_t, \boldsymbol{Y}^\eta_t] + 2D_{\boldsymbol{z}}\partial_\eta \boldsymbol{h}_{\eta\boldsymbol{\Delta}}[\boldsymbol{Y}^\eta_t] + \partial^2_\eta \boldsymbol{h}_{\eta\boldsymbol{\Delta}}, \qquad \boldsymbol{W}^\eta_{t_-} = 0, \tag{135}$$

where the coefficients without displayed arguments are evaluated at $(\boldsymbol{X}^\eta_t, t)$. Equations equation 129– equation 130 and Gronwall's inequality bound $\boldsymbol{A}^\eta_t$ and its inverse uniformly and give, for $j = 1, 2$ and $0 \leq r \leq 2$,

$$\|\partial_\eta^j D_{\boldsymbol{z}}^r \Psi^\eta_t(\boldsymbol{z})\| \leq C\varepsilon^j (1 + \|\boldsymbol{z}\|)^N. \tag{136}$$

Here the inverse-flow bounds follow by differentiating $\boldsymbol{\Phi}^\eta_{t_-,t}(\Psi^\eta_t(\boldsymbol{z})) = \boldsymbol{z}$ in $\boldsymbol{z}$ and η.

Denote by $p^{\eta\boldsymbol{\Delta}}_t(\boldsymbol{z})$ the density at time t for the velocity $\boldsymbol{h}_{\eta\boldsymbol{\Delta}}$. Since the perturbation vanishes before t_-, its initial density is $p^{\eta\boldsymbol{\Delta}}_{t_-} = p_{t_-}$. The change-of-variables formula therefore gives

$$p^{\eta\boldsymbol{\Delta}}_t(\boldsymbol{z}) = p_{t_-}(\Psi^\eta_t(\boldsymbol{z}))\, |\det D_{\boldsymbol{z}}\Psi^\eta_t(\boldsymbol{z})|, \tag{137}$$

which implies the bounds

$$\left|\partial^j_\eta p^{\eta\boldsymbol{\Delta}}_t(\boldsymbol{z})\right| \leq C\varepsilon^j (1 + \|\boldsymbol{z}\|)^N e^{-c\|\boldsymbol{z}\|^2}, \qquad j = 0, 1, 2. \tag{138}$$

Here and below constants are uniform over $\varepsilon \in (0, \varepsilon_0)$ (for $\epsilon_0 > 0$ sufficiently small), $t \in I$, and $\eta \in [0, 1]$. For $j = 0$, the convention $\varepsilon^0 = 1$ means only that the density itself is uniformly dominated; no smallness in ε is claimed. The same argument applied after one spatial differentiation gives

$$\left\|\partial^j_\eta \nabla p^{\eta\boldsymbol{\Delta}}_t(\boldsymbol{z})\right\| \leq C\varepsilon^j (1 + \|\boldsymbol{z}\|)^N e^{-c\|\boldsymbol{z}\|^2}, \qquad j = 0, 1, 2. \tag{139}$$

The Gaussian factor follows from equation 132 with the Gaussian decay of p_{t_-} and its derivatives; all remaining factors in the differentiated density formula equation 137 have polynomial growth.

Define the signed density variation $\dot{p}_t[\boldsymbol{\Delta}] := \partial_\eta p^{\eta\boldsymbol{\Delta}}_t\Big|_{\eta=0}$, where the dot here denotes variation in η, not time differentiation. Taylor's formula and equation 138 give

$$|p^{\boldsymbol{\Delta}}_t(\boldsymbol{z}) - p_t(\boldsymbol{z})| \leq C\varepsilon (1 + \|\boldsymbol{z}\|)^N e^{-c\|\boldsymbol{z}\|^2}, \tag{140}$$

$$\sup_{t\in I} \|p^{\boldsymbol{\Delta}}_t - p_t - \dot{p}_t[\boldsymbol{\Delta}]\|_{L^1(\mathbb{R}^d)} \leq C\varepsilon^2. \tag{141}$$

Transport over measurable basins. Let $\mathcal{U}_{s,t}$ be the transport operator on signed densities for the true flow from s to t. Explicitly, for $f \in L^1(\mathbb{R}^d)$,

$$\mathcal{U}_{s,t}[f](\boldsymbol{z}) := f\big(\boldsymbol{\Phi}^*_{t,s}(\boldsymbol{z})\big)\,\big|\det D_{\boldsymbol{z}}\boldsymbol{\Phi}^*_{t,s}(\boldsymbol{z})\big|, \qquad t_- \le s \le t \le t_+, \tag{142}$$

where $\boldsymbol{\Phi}^*_{t,s} = (\boldsymbol{\Phi}^*_{s,t})^{-1}$. Thus $\mathcal{U}_{s,t}[f]$ is the density of the pushforward of the signed measure $f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ by $\boldsymbol{\Phi}^*_{s,t}$; in particular it preserves total signed mass and is an L^1 contraction. Differentiating the continuity equation in η gives, in its integrated or distributional sense,

$$\partial_t \dot{p}_t[\boldsymbol{\Delta}] + \nabla_{\boldsymbol{z}} \cdot (\boldsymbol{h}_* \dot{p}_t[\boldsymbol{\Delta}]) = -\nabla_{\boldsymbol{z}} \cdot (p_t \boldsymbol{v}_t[\boldsymbol{\Delta}]), \qquad \dot{p}_{t_-}[\boldsymbol{\Delta}] = 0. \tag{143}$$

Hence

$$\dot{p}_v[\boldsymbol{\Delta}] = -\int_{t_-}^{v} \mathcal{U}_{t,v}[\nabla_{\boldsymbol{z}} \cdot (p_t \boldsymbol{v}_t[\boldsymbol{\Delta}])]\,\mathrm{d}t, \qquad v \in I. \tag{144}$$

The integrals are well-defined in L^1 by the Gaussian bounds equation 138, equation 139. The basin transport identity equation 52 implies, for every integrable signed density f, and any $t < v < T$,

$$\int_{\mathcal{A}_\sigma(v)} \mathcal{U}_{t,v}[f](\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \int_{\mathcal{A}_\sigma(t)} f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{145}$$

This is a change of variables.

We also record an identity for later use. Put $\boldsymbol{V}^{\boldsymbol{\Delta}}_t := \boldsymbol{h}_{\boldsymbol{\Delta}}(\cdot,t) - \boldsymbol{h}_*(\cdot,t)$. Subtracting the two continuity equations gives the source $-\nabla \cdot (p^{\boldsymbol{\Delta}}_t \boldsymbol{V}^{\boldsymbol{\Delta}}_t)$ under the true transport. Consequently,

$$\int_{\mathcal{A}_\sigma(v)} p^{\boldsymbol{\Delta}}_v(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} - q(\sigma) = -\int_{t_-}^{v}\int_{\mathcal{A}_\sigma(t)} \nabla_{\boldsymbol{z}} \cdot \big\{p^{\boldsymbol{\Delta}}_t(\boldsymbol{z})\boldsymbol{V}^{\boldsymbol{\Delta}}_t(\boldsymbol{z})\big\}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}t, \qquad v \in I. \tag{146}$$

Further $\nabla_{\boldsymbol{z}} \cdot \big\{p^{\boldsymbol{\Delta}}_t(\boldsymbol{z})\boldsymbol{V}^{\boldsymbol{\Delta}}_t(\boldsymbol{z})\big\}$ is integrable uniformly in time, by the bounds equation 138, equation 139 above.

After t_+ the perturbed flow equals the true flow. Lemma J.1 and absolute continuity of $p^{\boldsymbol{\Delta}}_{t_+}$ therefore imply

$$\mathrm{P}_{\boldsymbol{\Delta}}(\{\boldsymbol{x}(\sigma)\}) = \int_{\mathcal{A}_\sigma(t_+)} p^{\boldsymbol{\Delta}}_{t_+}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{147}$$

In particular, the perturbed terminal law has no mass outside the embeddings. Equations equation 144 and equation 145, with $v = t_+$, identify the derivative of equation 147 with the linear functional defined by equation 72–equation 74. Integration over the basin partition is an L^1 contraction into the class-mass vector:

$$\sum_{\sigma=0}^{k-1}\left|\int_{\mathcal{A}_\sigma(t_+)} f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\right| \le \|f\|_{L^1(\mathbb{R}^d)}. \tag{148}$$

Thus equation 141 gives the claimed $O(\|\boldsymbol{\Delta}\|^2_\infty)$ remainder in total variation, uniformly over bounded measurable perturbations.

Substituting the velocity variation equation 127 into equation 144 and using equation 145, the coefficient of $\Delta_{j,\rho}(t)$ in the terminal mass variation is precisely equation 74. This gives the pairing equation 72. The kernel integrands have uniform polynomial-times-Gaussian bounds, so they are absolutely integrable and the kernels are bounded.

To prove the continuity claim $\boldsymbol{L}_\sigma \in C(I;E)$, fix $\boldsymbol{D} \in E$ and define

$$F_{t,\boldsymbol{D}}(\boldsymbol{z}) := \nabla_{\boldsymbol{z}} \cdot \{c_1(t)p_t(\boldsymbol{z})\mathcal{M}_t(\boldsymbol{z})[\boldsymbol{D}]\}, \qquad \tilde{\boldsymbol{\Psi}}_t := \boldsymbol{\Phi}^*_{t_+,t}. \tag{149}$$

The definition of the response kernel first gives

$$\langle \boldsymbol{L}_\sigma(t), \boldsymbol{D}\rangle = -\int_{\mathcal{A}_\sigma(t)} F_{t,\boldsymbol{D}}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{150}$$

By equation 52, $\mathbf{\Phi}^*_{t,t_+}(\mathcal{A}_\sigma(t)) = \mathcal{A}_\sigma(t_+)$. Hence the substitution $\boldsymbol{y} = \mathbf{\Phi}^*_{t,t_+}(\boldsymbol{z})$, or equivalently $\boldsymbol{z} = \tilde{\mathbf{\Psi}}_t(\boldsymbol{y})$, changes the moving basin to the fixed measurable set $\mathcal{A}_\sigma(t_+)$ and yields

$$\langle \boldsymbol{L}_\sigma(t), \boldsymbol{D}\rangle = -\int_{\mathcal{A}_\sigma(t_+)} F_{t,\boldsymbol{D}}(\tilde{\mathbf{\Psi}}_t(\boldsymbol{y}))\,|\det D\tilde{\mathbf{\Psi}}_t(\boldsymbol{y})|\,\mathrm{d}\boldsymbol{y}. \tag{151}$$

The integrand is continuous in t and bounded in absolute value by $C\|\boldsymbol{D}\|_{\mathrm{F}}(1+\|\boldsymbol{y}\|)^N e^{-c\|\boldsymbol{y}\|^2}$ uniformly on I. Dominated convergence, applied to a basis of the finite-dimensional space E, proves $\boldsymbol{L}_\sigma \in C(I;E)$.

Finally, to prove equation 75, fix $t \in I$, $0 \le j \le d$, and $1 \le \rho \le k-1$, and set

$$\boldsymbol{G}_{t,j,\rho}(\boldsymbol{z}) := p_t(\boldsymbol{z})\pi^*_\rho(\boldsymbol{z},t)\{\boldsymbol{x}(\rho) - \boldsymbol{m}(\boldsymbol{z},t)\}\overline{z}_j. \tag{152}$$

Because the measurable basins partition $\mathbb{R}^d$ up to a null set and the divergence is absolutely integrable,

$$\sum_{\sigma=0}^{k-1}[\boldsymbol{L}_\sigma(t)]_{j,\rho} = -c_1(t)\int_{\mathbb{R}^d} \nabla_{\boldsymbol{z}}\cdot \boldsymbol{G}_{t,j,\rho}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{153}$$

To evaluate the last integral, choose smooth cutoffs χ_R that equal one on $\{\|\boldsymbol{z}\| \le R\}$, vanish outside $\{\|\boldsymbol{z}\| \le 2R\}$, and satisfy $\|\nabla\chi_R\| \le C/R$. The polynomial-times-Gaussian bounds on $\boldsymbol{G}_{t,j,\rho}$ and its divergence allow dominated convergence and integration by parts, giving

$$\begin{aligned}\int_{\mathbb{R}^d} \nabla_{\boldsymbol{z}}\cdot \boldsymbol{G}_{t,j,\rho}\,\mathrm{d}\boldsymbol{z} &= \lim_{R\to\infty}\int_{\mathbb{R}^d}\chi_R\nabla_{\boldsymbol{z}}\cdot\boldsymbol{G}_{t,j,\rho}\,\mathrm{d}\boldsymbol{z}\\ &= -\lim_{R\to\infty}\int_{\mathbb{R}^d}\langle\nabla\chi_R, \boldsymbol{G}_{t,j,\rho}\rangle\,\mathrm{d}\boldsymbol{z} = 0.\end{aligned} \tag{154}$$

Equation equation 153 therefore proves $\sum_\sigma \boldsymbol{L}_\sigma(t) = 0$. Substitution into equation 72 gives $\sum_\sigma \mathcal{L}_\sigma(\mathbf{\Delta}) = 0$, completing equation 75.

J.5.6 Proof of Proposition J.5

Write $b = b_n$ and $e_n = \sqrt{\log(1/b)/(nb)}$. The bandwidth conditions imply that, eventually, $n^{-1/2} < b < n^{-1/4}$. Hence $\log(1/b) \asymp \log n$, $nb > \sqrt{n}$, and $e_n \to 0$. All constants below may depend on I, J, K_0, but not on n.

Uniform empirical bounds. We first establish

$$\max_{0\le j\le 2}\sup_{t\in I,\,\boldsymbol{\theta}\in K_0}\left\|b^{-1}D^j_{\boldsymbol{\theta}}(\hat{R}_{b,t} - R_{b,t})(\boldsymbol{\theta})\right\| = O_{\mathrm{P}}(e_n). \tag{155}$$

Let P be the law of $Z = (U, \Sigma, \boldsymbol{G})$, let P_n be its empirical measure, and write $\boldsymbol{Y} = a(U)\boldsymbol{x}(\Sigma) + \tau(U)\boldsymbol{G}$. For $j = 0, 1, 2$, define the tensor-valued functions

$$f^{(j)}_{b,t,\boldsymbol{\theta}}(Z) := w_b(U-t)D^j_{\boldsymbol{\theta}}\ell_\Sigma(\boldsymbol{Y};\boldsymbol{\theta}). \tag{156}$$

Differentiation under the expectation is justified by equation 48, so $(\widehat{\mathbb{E}}_n - \mathbb{E})f^{(j)}_{b,t,\boldsymbol{\theta}} = D^j_{\boldsymbol{\theta}}(\hat{R}_{b,t} - R_{b,t})(\boldsymbol{\theta})$. Since E has fixed finite dimension, it suffices to bound the finitely many coordinates of these tensors.

For small b, a nonzero kernel weight with $t \in I$ implies $U \in J$. On J, the functions a, τ, λ are bounded, and the Gaussian representation gives uniformly bounded conditional moments of every fixed polynomial in $\|\boldsymbol{Y}\|$. Consequently,

$$\mathbb{E}\|f^{(j)}_{b,t,\boldsymbol{\theta}}\|^2 \le C\int_{|u-t|\le b}\mathbb{E}[(1+\|\boldsymbol{Y}\|)^{2\kappa} \mid U = u]\lambda(u)\,\mathrm{d}u \le Cb. \tag{157}$$

Set $\widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}} := f^{(j)}_{b,t,\boldsymbol{\theta}}\mathbf{1}_{\{\|\boldsymbol{G}\|\le D\sqrt{\log n}\}}$. Fix $m > 1$. Choosing D sufficiently large gives

$$\mathbb{P}\left(\max_{1\le i\le n}\|\boldsymbol{G}_i\| > D\sqrt{\log n}\right) = O(n^{-m}), \tag{158}$$

$$\sup_{t,\boldsymbol{\theta},j} b^{-1}\mathbb{E}\|f^{(j)}_{b,t,\boldsymbol{\theta}} - \widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}}\| = O(n^{-m}). \tag{159}$$

Here the factor b in the tail expectation again comes from kernel support, and $\boldsymbol{G}$ is independent of U. The truncated functions are upper bounded by $B_n = C(\log n)^{\kappa_0}$ for some fixed κ_0 and variance at most Cb. Hence, by Bernstein inequality

$$\mathbb{P}\left(\max_{0\le j\le 2}\|(\widehat{\mathbb{E}}_n-\mathbb{E})\widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}}\| > C\left\{\sqrt{\frac{bx}{n}}+\frac{B_n x}{n}\right\}\right)\le Ce^{-x}. \tag{160}$$

Take a net in $I\times K_0$ with time mesh at most $bn^{-3/\alpha}$ and parameter mesh at most n^{-3}. Its cardinality is at most Cn^q for a fixed q, since $b^{-1}<n^{1/2}$ and K_0 is finite-dimensional and compact. Hence, a union bound over this net gives, for $x\ge 1$,

$$\mathbb{P}\left(\max_{\text{net},\,0\le j\le 2}\|(\widehat{\mathbb{E}}_n-\mathbb{E})\widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}}\| > C\left\{\sqrt{\frac{bx}{n}}+\frac{B_n x}{n}\right\}\right)\le Cn^q e^{-x}. \tag{161}$$

Choosing $x = A\log n$ with $A > q+1$ and dividing by b bounds the maximum on the net by

$$\max_{\text{net},\,0\le j\le 2}\frac{1}{b}\|(\widehat{\mathbb{E}}_n-\mathbb{E})\widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}}\| = O_{\mathrm{P}}\left(\sqrt{\frac{\log n}{nb}}+\frac{B_n\log n}{nb}\right)=O_{\mathrm{P}}(e_n), \tag{162}$$

because $B_n\sqrt{\log n}/\sqrt{nb}\to 0$ by Assumption 3.

To pass from the net to all of $I\times K_0$, kernel Hölder continuity and the parameter derivative bounds give

$$\|\widetilde{f}^{(j)}_{b,t,\boldsymbol{\theta}}-\widetilde{f}^{(j)}_{b,s,\boldsymbol{\vartheta}}\|\le CB_n\left\{(|t-s|/b)^{\alpha}+\|\boldsymbol{\theta}-\boldsymbol{\vartheta}\|\right\},\qquad j=0,1,2. \tag{163}$$

Increase B_n if necessary to bound derivatives through order three. Convexity of K_0 allows the mean-value bound along the parameter segment. For a point and its nearest net point, the right-hand side is CB_nn^{-3}, so the normalized interpolation error for $\widehat{\mathbb{E}}_n-\mathbb{E}$ is at most $CB_nn^{-3}/b = o(e_n)$. Indeed, $b^{-1}<n^{1/2}$ and $e_n\ge n^{-1/2}$ for large n. Together with equation 159, this proves equation 155.

We also need a uniform bound on empirical third derivatives. Choose $M(\boldsymbol{y}) = C(1+\|\boldsymbol{y}\|)^{\kappa}$ dominating their norms on K_0. Apply the same truncation and time-net argument to $w_b(U-t)M(\boldsymbol{Y})$; its expectation is $O(b)$ and its centered empirical average is $O_{\mathrm{P}}(be_n)$ uniformly in t. Thus

$$\sup_{t\in I}\frac{1}{nb}\sum_{i=1}^{n}w_b(t_i-t)M(\boldsymbol{Y}_i)=O_{\mathrm{P}}(1). \tag{164}$$

This step uses the envelope itself, so no fourth parameter derivative is required.

Localization, existence, and uniqueness. Recall $d_b(t)\asymp b$ from Lemma J.10. The $j=0$ empirical bound equation 155 and equation 114 imply

$$\eta_n := \sup_{t\in I,\,\boldsymbol{\theta}\in K_0}\left|\frac{\hat{R}_{b,t}(\boldsymbol{\theta})}{d_b(t)}-R_t(\boldsymbol{\theta})\right| = o_{\mathrm{P}}(1). \tag{165}$$

In the constrained case put $\mathcal{D}(t)=K(t)$, and in the unconstrained case first put $\mathcal{D}(t)=K_0$. Each $\mathcal{D}(t)$ is compact and contains the unique local minimizer $\boldsymbol{\theta}^*(t)$. The uniform bound equation 165 implies that all empirical minimizers over $\mathcal{D}(t)$ converge uniformly to $\boldsymbol{\theta}^*(t)$ in probability.

By uniform positive definiteness of $\boldsymbol{H}^*(t)$, continuity of the population Hessian, and the uniform interior margins, we can choose $r,c>0$ so that the balls $\mathsf{B}(\boldsymbol{\theta}^*(t),r)$ lie in $\mathrm{int}(\mathcal{D}(t))$ and $\nabla^2R_t(\boldsymbol{\theta})\succeq 2c\,\mathrm{Id}_E$ on these balls. The $j=2$ empirical bound and population Hessian convergence then give, with probability tending to one,

$$\nabla^2\hat{R}_{b,t}(\boldsymbol{\theta})\succeq c\,d_b(t)\,\mathrm{Id}_E\succeq c'b\,\mathrm{Id}_E,\qquad t\in I,\ \boldsymbol{\theta}\in B(\boldsymbol{\theta}^*(t),r). \tag{166}$$

Hence the minimizer $\hat{\boldsymbol{\theta}}_b(t)\in\mathsf{B}(\boldsymbol{\theta}^*(t),r/2)$ is unique in the interior of $\mathcal{D}(t)$ and satisfies the first-order condition $\nabla\hat{R}_{b,t}(\hat{\boldsymbol{\theta}}_b(t))=0$. In the convex uncostrained case, it is globally unique because of strong convexity at $\hat{\boldsymbol{\theta}}_b(t)$.

Uniform rate and linearization. Write $\boldsymbol{\delta}_n(t) := \hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}^b(t)$ and define the averaged empirical Hessian by

$$\widehat{\boldsymbol{H}}_n(t) := \int_0^1 \nabla^2 \hat{R}_{b,t}(\boldsymbol{\theta}^b(t) + u\boldsymbol{\delta}_n(t))\,\mathrm{d}u. \tag{167}$$

The population minimizers also converge uniformly to $\boldsymbol{\theta}^*$ by Lemma J.10, so the entire segment in this integral lies in the balls of equation 166 with probability tending to one. In particular,

$$\sup_{t\in I} \|\widehat{\boldsymbol{H}}_n(t)^{-1}\|_{\mathrm{op}} = O_{\mathrm{P}}(b^{-1}). \tag{168}$$

The empirical first-order condition and the fundamental theorem of calculus give

$$0 = \widehat{\boldsymbol{S}}_n(t, \hat{\boldsymbol{\theta}}_b(t)) = \widehat{\boldsymbol{S}}_n(t, \boldsymbol{\theta}^b(t)) + \widehat{\boldsymbol{H}}_n(t)[\boldsymbol{\delta}_n(t)], \tag{169}$$

$$\boldsymbol{\delta}_n(t) = -\widehat{\boldsymbol{H}}_n(t)^{-1}[\widehat{\boldsymbol{S}}_n(t, \boldsymbol{\theta}^b(t))]. \tag{170}$$

Since $\nabla R_{b,t}(\boldsymbol{\theta}^b(t)) = 0$, the $j = 1$ empirical bound in Eq. equation 155 implies

$$\sup_{t\in I} \|\widehat{\boldsymbol{S}}_n(t, \boldsymbol{\theta}^b(t))\| = O_{\mathrm{P}}(be_n). \tag{171}$$

Combining this with equation 170 proves $\|\boldsymbol{\delta}_n\|_\infty = O_{\mathrm{P}}(e_n)$, as claimed in equation 80.

To replace the averaged empirical Hessian by $\boldsymbol{H}_b(t)$, split

$$\begin{aligned}\widehat{\boldsymbol{H}}_n(t) - \boldsymbol{H}_b(t) = {} & \{\nabla^2 \hat{R}_{b,t}(\boldsymbol{\theta}^b(t)) - \nabla^2 R_{b,t}(\boldsymbol{\theta}^b(t))\} \\ & + \int_0^1 \{\nabla^2 \hat{R}_{b,t}(\boldsymbol{\theta}^b(t) + u\boldsymbol{\delta}_n(t)) - \nabla^2 \hat{R}_{b,t}(\boldsymbol{\theta}^b(t))\}\,\mathrm{d}u.\end{aligned} \tag{172}$$

The first term is uniformly $O_{\mathrm{P}}(be_n)$ by the $j = 2$ empirical bound in Eq. equation 155. By the mean-value theorem and equation 164, the norm of the second term is at most

$$\frac{\|\boldsymbol{\delta}_n(t)\|}{2n} \sum_{i=1}^n w_b(t_i - t) M(\boldsymbol{Y}_i), \tag{173}$$

which is uniformly $O_{\mathrm{P}}(b\|\boldsymbol{\delta}_n\|_\infty) = O_{\mathrm{P}}(be_n)$. Consequently,

$$\|\widehat{\boldsymbol{H}}_n - \boldsymbol{H}_b\|_\infty = O_{\mathrm{P}}(be_n). \tag{174}$$

Both inverse Hessians have uniform norm $O_{\mathrm{P}}(b^{-1})$, with the population bound deterministic. The identity

$$\widehat{\boldsymbol{H}}_n^{-1} - \boldsymbol{H}_b^{-1} = \widehat{\boldsymbol{H}}_n^{-1} \circ (\boldsymbol{H}_b - \widehat{\boldsymbol{H}}_n) \circ \boldsymbol{H}_b^{-1} \tag{175}$$

therefore gives $\|\widehat{\boldsymbol{H}}_n^{-1} - \boldsymbol{H}_b^{-1}\|_\infty = O_{\mathrm{P}}(e_n/b)$. Define

$$\boldsymbol{r}_n(t) := \{\boldsymbol{H}_b(t)^{-1} - \widehat{\boldsymbol{H}}_n(t)^{-1}\}[\widehat{\boldsymbol{S}}_n(t, \boldsymbol{\theta}^b(t))]. \tag{176}$$

Equations equation 170 and equation 171 now yield

$$\boldsymbol{\delta}_n(t) = -\boldsymbol{H}_b(t)^{-1}[\widehat{\boldsymbol{S}}_n(t, \boldsymbol{\theta}^b(t))] + \boldsymbol{r}_n(t), \tag{177}$$

$$\|\boldsymbol{r}_n\|_\infty = O_{\mathrm{P}}(e_n/b)\, O_{\mathrm{P}}(be_n) = O_{\mathrm{P}}(e_n^2). \tag{178}$$

Finally,

$$\sqrt{n}\, e_n^2 = \frac{\log(1/b)}{\sqrt{n}\, b} \longrightarrow 0 \tag{179}$$

by Assumption 3, proving the last assertion of equation 82.

J.5.7 Proof of Proposition J.6

Propositions J.3 and J.5 imply

$$\sqrt{n}\int_I \langle \hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}^*, \boldsymbol{\phi}\rangle \,\mathrm{d}t = -\sqrt{n}\int_I \langle \boldsymbol{H}_b(t)^{-1}[\widehat{\boldsymbol{S}}_n(t,\boldsymbol{\theta}^b(t))], \boldsymbol{\phi}(t)\rangle \,\mathrm{d}t + o_{\mathrm{P}}(1). \tag{180}$$

Indeed the smoothing bias term is $O(\sqrt{n}b^2) = o(1)$ and the stochastic remainder is $O_{\mathrm{P}}(\log(1/b)/(\sqrt{n}b)) = o_{\mathrm{P}}(1)$. For $u \in [0,T]$, $\sigma \in \{0,\dots,k-1\}$, and $\boldsymbol{g} \in \mathbb{R}^d$, set

$$F_{b,\boldsymbol{\phi}}(u,\sigma,\boldsymbol{g}) := \int_I \left\langle \boldsymbol{\phi}(t), \boldsymbol{H}_b(t)^{-1}\left[w_b(u-t)\boldsymbol{\psi}_\sigma\big(a(u)\boldsymbol{x}(\sigma)+\tau(u)\boldsymbol{g};\boldsymbol{\theta}^b(t)\big)\right]\right\rangle \,\mathrm{d}t. \tag{181}$$

Applying Fubini's theorem to the leading term in equation 180 gives

$$\sqrt{n}\int_I \left\langle \hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}^b(t), \boldsymbol{\phi}(t)\right\rangle \,\mathrm{d}t = -\frac{1}{\sqrt{n}}\sum_{i=1}^n F_{b,\boldsymbol{\phi}}(t_i,\sigma_i,\boldsymbol{G}_i) + o_{\mathrm{P}}(1), \tag{182}$$

where $\boldsymbol{Y}_i = a(t_i)\boldsymbol{x}(\sigma_i) + \tau(t_i)\boldsymbol{G}_i$. The summands in equation 182 are i.i.d. within each row and have mean zero, since

$$\mathbb{E}_{\Sigma,U,\boldsymbol{G}}\left[w_b(U-t)\boldsymbol{\psi}_\Sigma\big(a(U)\boldsymbol{x}(\Sigma)+\tau(U)\boldsymbol{G};\boldsymbol{\theta}^b(t)\big)\right] = 0 \tag{183}$$

by the population first-order condition $\nabla R_{b,t}(\boldsymbol{\theta}^b(t)) = 0$.

The Hessian limit equation 117, the convergence $\boldsymbol{\theta}^b \to \boldsymbol{\theta}^*$, and the approximate-identity property of $(b\mu_0)^{-1}w_1(\cdot/b)$ imply, as $b \to 0$,

$$\begin{aligned} F_{b,\boldsymbol{\phi}}(U,\Sigma,\boldsymbol{G}) &\longrightarrow F_{0,\boldsymbol{\phi}}(U,\Sigma,\boldsymbol{G}) \\ &:= \frac{1}{\lambda(U)}\left\langle \boldsymbol{\phi}(U), \boldsymbol{H}^*(U)^{-1}\left[\boldsymbol{\psi}_\Sigma\big(a(U)\boldsymbol{x}(\Sigma)+\tau(U)\boldsymbol{G};\boldsymbol{\theta}^*(U)\big)\right]\right\rangle \end{aligned} \tag{184}$$

in L^2, defining the limit to be zero for $U \notin I$. Extend $\boldsymbol{\phi}$ by zero outside I; if it is nonzero at an endpoint, the boundary layer has width $O(b)$ and a uniformly square-integrable envelope, so it does not affect the limit in Eq. equation 184. Indeed, $|F_{b,\boldsymbol{\phi}}(U,\Sigma,\boldsymbol{G})| \le C(1+\|\boldsymbol{G}\|)^\kappa$ uniformly in small b, since $\|\boldsymbol{H}_b^{-1}\| = O(b^{-1})$ and $\int_I w_b(U-t)\,\mathrm{d}t = O(b)$. Thus all these variables have a common $L^{2+\eta}$ upper bound. Consequently the triangular array in equation 182 satisfies the Lindeberg condition, and its variance converges to

$$\mathbb{E}F_{0,\boldsymbol{\phi}}(U,\Sigma,\boldsymbol{G})^2 = \int_I \frac{1}{\lambda(t)}\left\langle \boldsymbol{\phi}(t), \boldsymbol{H}^*(t)^{-1}\left[\boldsymbol{J}^*(t)\left[\boldsymbol{H}^*(t)^{-1}[\boldsymbol{\phi}(t)]\right]\right]\right\rangle \,\mathrm{d}t \tag{185}$$

$$= Q^*(\boldsymbol{\phi}). \tag{186}$$

The Lindeberg–Feller theorem applied to equation 182, followed by equation 180 and Slutsky's theorem, proves equation 83. Applying the same argument to linear combinations of finitely many test functions proves the final joint convergence claim.

J.5.8 Proof of Theorem 2

Fix $0 < \varepsilon < T/2$ and write $I_\varepsilon = [\varepsilon, T-\varepsilon]$. Let $\boldsymbol{\delta}_{n,\varepsilon}(t) = \hat{\boldsymbol{\theta}}_b(t) - \boldsymbol{\theta}^*(t)$ on I_ε and zero elsewhere. The bias and empirical rate imply

$$\|\boldsymbol{\delta}_{n,\varepsilon}\|_\infty = O_{\mathrm{P}}(e_n + b^2), \tag{187}$$

$$\sqrt{n}\|\boldsymbol{\delta}_{n,\varepsilon}\|_\infty^2 = O_{\mathrm{P}}\left(\frac{\log(1/b)}{\sqrt{n}b} + \sqrt{n}b^4\right) = o_{\mathrm{P}}(1). \tag{188}$$

The volume-response kernels are continuous and bounded on I_ε by Proposition J.4. Define

$$\mathcal{Z}_{n,\varepsilon,\sigma} := \sqrt{n}\,\mathcal{L}_\sigma(\boldsymbol{\delta}_{n,\varepsilon}) = \sqrt{n}\int_{I_\varepsilon}\langle \boldsymbol{L}_\sigma(t), \boldsymbol{\delta}_{n,\varepsilon}(t)\rangle \,\mathrm{d}t. \tag{189}$$

Apply Proposition J.6 jointly to the k test functions $\boldsymbol{L}_\sigma$ to get $\mathcal{Z}_{n,\varepsilon} \Rightarrow \mathsf{N}_k(0, \boldsymbol{V}_\varepsilon(\lambda))$, where $\boldsymbol{V}_\varepsilon$ is defined as in equation 91 with the integral over I_ε. On events of probability tending to one $\boldsymbol{\delta}_{n,\varepsilon}$

is a small bounded measurable perturbation, so Proposition J.4 applies pathwise, with a uniform remainder constant. The density remainder equation 141 and the contraction equation 148 give $O_{\mathrm{P}}(\|\boldsymbol{\delta}_{n,\varepsilon}\|_\infty^2) = o_{\mathrm{P}}(n^{-1/2})$ in total variation. Thus, writing $\hat{q}_{n,\varepsilon}(\sigma) := \mathrm{P}_{n,\varepsilon}(\{\boldsymbol{x}(\sigma)\})$,

$$\sqrt{n}(\hat{q}_{n,\varepsilon} - q) \Rightarrow \mathsf{N}_k(0, \boldsymbol{V}_\varepsilon(\lambda)). \tag{190}$$

Lemma J.1 and absolute continuity at t_+ ensure that the truncated terminal law has precisely the stated discrete support, as in equation 147.

The diagonal integrability condition equation 90 and Cauchy–Schwarz for the positive semidefinite operator $\boldsymbol{Q}^*(t)$ imply that every entry of the covariance integrand is absolutely integrable. Hence $\boldsymbol{V}_\varepsilon(\lambda) \to \boldsymbol{V}(\lambda)$ as $\varepsilon \downarrow 0$. Moreover,

$$\|\hat{q}_n - \hat{q}_{n,\varepsilon}\|_\infty \le \|\mathrm{P}_n - \mathrm{P}_{n,\varepsilon}\|_{\mathrm{TV}}. \tag{191}$$

Equation equation 89, equation 190, and the converging-together theorem therefore prove equation 92 for the process using $\hat{\boldsymbol{\theta}}_b$ on all of $(0, T)$.

Finally, summing the volume kernels over the basin partition gives equation 75; consequently $\boldsymbol{V}_\varepsilon(\lambda)\mathbf{1} = 0$ for every ε, and passing to the limit gives $\boldsymbol{V}(\lambda)\mathbf{1} = 0$.

J.5.9 Proof of Corollary J.7

Fix a loss ℓ in the class stated in the corollary and fix $t \in (0, T)$. Because the embeddings have equal norm, Bayes' formula gives $P_{t,\boldsymbol{\theta}^*(t)} = P_t$, the true joint law of $(\Sigma, \boldsymbol{Y})$ at time t. Write $\boldsymbol{\psi}^\ell := \nabla_{\boldsymbol{\theta}}\ell$ and $\boldsymbol{\psi}^{\mathrm{CE}} := \nabla_{\boldsymbol{\theta}}\ell^{\mathrm{CE}}$. The likelihood score of the conditional model equation 93 is $-\boldsymbol{\psi}^{\mathrm{CE}}$.

Set $\boldsymbol{H}_\ell(t) := \mathbb{E}_t[D_{\boldsymbol{\theta}}\boldsymbol{\psi}^\ell]$ and $\boldsymbol{J}_\ell(t) := \mathbb{E}_t[\boldsymbol{\psi}^\ell \otimes \boldsymbol{\psi}^\ell]$, with all quantities on the right evaluated at $\boldsymbol{\theta}^*(t)$, and use analogous notation for cross-entropy. Differentiating equation 94 with respect to $\boldsymbol{\vartheta}$ at $\boldsymbol{\theta}^*(t)$ gives

$$\mathbb{E}_t[\boldsymbol{\psi}^\ell \otimes \boldsymbol{\psi}^{\mathrm{CE}}] = \boldsymbol{H}_\ell(t). \tag{192}$$

The same identity with $\ell = \ell^{\mathrm{CE}}$ is the information identity $\boldsymbol{J}_{\mathrm{CE}}(t) = \boldsymbol{H}_{\mathrm{CE}}(t)$. Now define the two influence functions

$$\boldsymbol{\zeta}_\ell := \boldsymbol{H}_\ell(t)^{-1}[\boldsymbol{\psi}^\ell], \qquad \boldsymbol{\zeta}_{\mathrm{CE}} := \boldsymbol{H}_{\mathrm{CE}}(t)^{-1}[\boldsymbol{\psi}^{\mathrm{CE}}]. \tag{193}$$

Using equation 192, its adjoint, and the self-adjointness of the Hessians gives

$$\mathbb{E}_t\big[(\boldsymbol{\zeta}_\ell - \boldsymbol{\zeta}_{\mathrm{CE}}) \otimes (\boldsymbol{\zeta}_\ell - \boldsymbol{\zeta}_{\mathrm{CE}})\big] = \boldsymbol{Q}^*_\ell(t) - \boldsymbol{Q}^*_{\mathrm{CE}}(t) \succeq 0. \tag{194}$$

Thus the cross-entropy sandwich operator is no larger than the sandwich operator for ℓ at every t.

Finally, the volume-response kernels $\boldsymbol{L}_\sigma(t)$ in equation 74 depend only on the true distribution and the common parameterization of the denoiser, not on the loss. Hence, for every $v \in \mathbb{R}^k$, equation 91 and equation 194 imply

$$v^\intercal\{\boldsymbol{V}^\ell(\lambda) - \boldsymbol{V}^{\mathrm{CE}}(\lambda)\}v = \int_I \frac{1}{\lambda(t)} \left\langle \sum_{\sigma=0}^{k-1} v_\sigma \boldsymbol{L}_\sigma(t), \{\boldsymbol{Q}^*_\ell(t) - \boldsymbol{Q}^*_{\mathrm{CE}}(t)\} \left[\sum_{\alpha=0}^{k-1} v_\alpha \boldsymbol{L}_\alpha(t)\right] \right\rangle \mathrm{d}t \ge 0. \tag{195}$$

This is precisely equation 95. In particular, squared denoising loss belongs to the comparison class because, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$,

$$\mathbb{E}_{t,\boldsymbol{\vartheta}}\big[\nabla_{\boldsymbol{\theta}}\ell^{\mathrm{SM}}_\Sigma(\boldsymbol{Y};\boldsymbol{\vartheta}) \mid \boldsymbol{Y} = \boldsymbol{y}\big] = -2D_{\boldsymbol{\theta}}\hat{\boldsymbol{m}}(\boldsymbol{y};\boldsymbol{\vartheta})^* \{\mathbb{E}_{t,\boldsymbol{\vartheta}}[\boldsymbol{x}(\Sigma) \mid \boldsymbol{Y} = \boldsymbol{y}] - \hat{\boldsymbol{m}}(\boldsymbol{y};\boldsymbol{\vartheta})\} = 0. \tag{196}$$

J.5.10 Proof of Proposition J.8

For $v \in \mathbb{R}^k$, set $\boldsymbol{L}_v(t) := \sum_{\sigma=0}^{k-1} v_\sigma \boldsymbol{L}_\sigma(t)$. Since $\boldsymbol{Q}^*(t)$ is positive semidefinite,

$$\boldsymbol{v}^\intercal \boldsymbol{C}(t)\boldsymbol{v} = \langle \boldsymbol{L}_v(t), \boldsymbol{Q}^*(t)[\boldsymbol{L}_v(t)]\rangle \ge 0. \tag{197}$$

Thus $\boldsymbol{C}(t) \succeq 0$. Because $\boldsymbol{W} \succeq 0$ as well, $h_W(t) = \mathrm{Tr}(\boldsymbol{W}^{1/2}\boldsymbol{C}(t)\boldsymbol{W}^{1/2}) \ge 0$. The covariance formula equation 91 and Fubini's theorem give

$$\mathrm{Tr}\left(\boldsymbol{W}\boldsymbol{V}(\lambda)\right) = \int_0^T \frac{h_W(t)}{\lambda(t)}\,\mathrm{d}t, \tag{198}$$

which proves equation 98. The continuous mapping theorem applied to equation 92 proves equation 99.

For every probability density λ with finite risk, Cauchy–Schwarz gives

$$H_W^2 = \left\{ \int_0^T \sqrt{\frac{h_W(t)}{\lambda(t)}} \sqrt{\lambda(t)}\, \mathrm{d}t \right\}^2 \le \left\{ \int_0^T \frac{h_W(t)}{\lambda(t)}\, \mathrm{d}t \right\} \left\{ \int_0^T \lambda(t)\, \mathrm{d}t \right\} = \mathcal{R}_W(\lambda). \tag{199}$$

When $0 < H_W < \infty$, equality holds precisely when $\lambda(t)$ is proportional to $\sqrt{h_W(t)}$ on its support. Normalization gives equation 101 and proves equation 100. If this density vanishes where strict positivity is required, positive regularizations converge to it and their risks converge to H_W^2, so the same value remains the infimum. The case $H_W = 0$ is immediate because $h_W = 0$ almost everywhere.

J.5.11 Proof of Proposition J.9

Proof. Write $\rho = a/\tau$ and rescale space by $\boldsymbol{u} = \boldsymbol{z}/\tau(t)$. From the definitions of c_1, c_2,

$$c_2(t) = \frac{\dot\tau(t)}{\tau(t)}, \qquad\qquad c_1(t) = \tau(t)\dot\rho(t). \tag{200}$$

Consequently the true probability-flow equation becomes

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \dot\rho(t)\widetilde{\boldsymbol{m}}(\boldsymbol{u}, \rho(t)), \qquad\qquad \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}\rho} = \widetilde{\boldsymbol{m}}(\boldsymbol{u}, \rho), \tag{201}$$

where $\widetilde{\boldsymbol{m}}(\boldsymbol{u}, r)$ is the posterior mean under the normalized mixture $\boldsymbol{u} = r\boldsymbol{x}(\Sigma) + \boldsymbol{G}$. Thus the normalized density, posterior, and terminal basins depend on time only through $r = \rho(t)$.

Let $A_t : E \to E$ multiply every slope column by $\tau(t)$ and leave its intercept unchanged. For either loss in the proposition,

$$\ell_\sigma(\tau(t)\boldsymbol{u}; \boldsymbol{\theta}) = \widetilde{\ell}_\sigma(\boldsymbol{u}; A_t\boldsymbol{\theta}). \tag{202}$$

Writing $\widetilde{\boldsymbol{H}}^*(r)$, $\widetilde{\boldsymbol{J}}^*(r)$, and $\widetilde{\boldsymbol{Q}}^*(r)$ for the normalized-coordinate operators, the chain rule gives

$$\boldsymbol{H}^*(t) = A_t^* \widetilde{\boldsymbol{H}}^*(\rho(t)) A_t, \tag{203}$$

$$\boldsymbol{J}^*(t) = A_t^* \widetilde{\boldsymbol{J}}^*(\rho(t)) A_t, \tag{204}$$

$$\boldsymbol{Q}^*(t) = A_t^{-1} \widetilde{\boldsymbol{Q}}^*(\rho(t)) (A_t^*)^{-1}. \tag{205}$$

In equation 74, make the change of variables $\boldsymbol{z} = \tau(t)\boldsymbol{u}$. Then

$$p_t(\tau(t)\boldsymbol{u}) = \tau(t)^{-d} \widetilde{p}_{\rho(t)}(\boldsymbol{u}), \qquad\qquad \mathrm{d}\boldsymbol{z} = \tau(t)^d \mathrm{d}\boldsymbol{u}, \tag{206}$$

$$\nabla_{\boldsymbol{z}} \cdot = \tau(t)^{-1} \nabla_{\boldsymbol{u}} \cdot, \qquad\qquad \mathcal{M}_t(\tau(t)\boldsymbol{u})[\boldsymbol{D}] = \widetilde{\mathcal{M}}_{\rho(t)}(\boldsymbol{u})[A_t \boldsymbol{D}]. \tag{207}$$

Together with $c_1 = \tau\dot\rho$ and the normalized-basin identity from equation 201, this yields

$$\boldsymbol{L}_\sigma(t) = \dot\rho(t) A_t^* \widetilde{\boldsymbol{L}}_\sigma(\rho(t)). \tag{208}$$

Combining equation 205 and equation 208 cancels the A_t factors:

$$C_{\sigma\alpha}(t) = \dot\rho(t)^2 \left\langle \widetilde{\boldsymbol{L}}_\sigma(\rho(t)), \widetilde{\boldsymbol{Q}}^*(\rho(t))[\widetilde{\boldsymbol{L}}_\alpha(\rho(t))] \right\rangle. \tag{209}$$

This proves equation 102. In particular, with $\widetilde{h}_W(r) := \mathrm{tr}\{\boldsymbol{W}\widetilde{\boldsymbol{C}}(r)\}$,

$$\int_0^T \sqrt{h_W(t)}\, \mathrm{d}t = \int_0^T |\dot\rho(t)| \sqrt{\widetilde{h}_W(\rho(t))}\, \mathrm{d}t = \int_{\rho((0,T))} \sqrt{\widetilde{h}_W(r)}\, \mathrm{d}r. \tag{210}$$

Equation equation 103 now follows from Proposition J.8. □