# Total Robustness in Bayesian Nonlinear Regression for Measurement Error Problems under Model Misspecification

Mengqi Chen †

*Department of Statistics, University of Warwick, United Kingdom.*

Charita Dellaporta

*Department of Statistical Science, University College London, United Kingdom.*

Thomas B. Berrett

*Department of Statistics, University of Warwick, United Kingdom.*

Theodoros Damoulas

*Department of Statistics & Department of Computer Science, University of Warwick, United Kingdom.*

**Abstract**. Modern regression analyses are often undermined by covariate measurement error, misspecification of the regression model, and misspecification of the measurement error distribution. We present, to the best of our knowledge, the first Bayesian nonparametric framework targeting *total robustness* that tackles all three challenges in general nonlinear regression. The framework assigns a Dirichlet process prior to the latent covariate–response distribution and updates it with posterior pseudo-samples of the latent covariates, thereby providing the Dirichlet process posterior with observation-informed latent inputs and yielding estimators that minimise the discrepancy between Dirichlet process realisations and the model-induced joint law. This design allows practitioners to (i) encode prior beliefs, (ii) choose between pseudo-sampling latent covariates or working directly with error-prone observations, and (iii) tune the influence of prior and data. We establish generalisation bounds that tighten whenever the prior or pseudo-sample generator aligns with the underlying data generating process, ensuring robustness without sacrificing consistency. A gradient-based algorithm enables efficient computations; simulations and two real-world studies show lower estimation error and reduced estimation sensitivity to misspecification compared to Bayesian and frequentist competitors. The framework, therefore, offers a practical and interpretable paradigm for trustworthy regression when data and models are jointly imperfect.

**Keywords:** Bayesian nonparametric learning, measurement error, misspecification, robustness.

†*Address for correspondence:* Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
Email: Mengqi.Chen.2@warwick.ac.uk

## 1. Introduction

Contemporary robust statistical regression analysis often faces three ubiquitous threats: covariate measurement error (ME), regression model misspecification, and misspecification of the ME distribution. However, existing robust approaches typically target only one of them, failing when also confronted by another. We introduce a robust Bayesian statistical framework that simultaneously protects against all three sources of bias, delivering *total robustness*. To clarify why this is needed, and to understand the limits of current methodology, we begin by examining the distinct challenges posed by ME and the many faces of misspecification.

*ME in covariates* arises in numerous fields, including economic, biomedical, and environmental studies, where recorded values deviate from the true unknown signal [Hausman et al., 1995, Brakenhoff et al., 2018, Haber et al., 2021, Curley, 2022]. This discrepancy can be classical (when the observation is a noisy version of the true covariate) or Berkson (when the true covariate has a random offset from a nominal target), as established by Berkson [1950] and summarised by Carroll et al. [2006]. These two canonical forms of ME, which we focus on in this paper, are part of a broader typology that also includes differential, multiplicative, and systematic ME [see Buonaccorsi, 2010, Chapter 1 for a full taxonomy]. If ignored, ME commonly biases estimates and can degrade inference on quantities of interest [Gustafson, 2003]. A substantial literature addresses ME [Deming, 1943, Stefanski, 1985, Cook and Stefanski, 1994, Wang, 2004, Schennach, 2013, Hu et al., 2020], yet many rely on restrictive assumptions, such as known error distributions or replicate measurements [Delaigle et al., 2006, McIntyre and Stefanski, 2011]. Nonparametric approaches, including deconvolution and Bayesian frameworks, were developed to alleviate such assumptions [Delaigle and Hall, 2016, Dellaporta and Damoulas, 2023].

*Model misspecification* arises whenever the model family cannot explain the true data generating process (DGP). In regression, this may occur because the true regression function lies outside the assumed parametric family, the noise distribution is misspecified, or the data contain an $\varepsilon$-contaminated fraction of outliers generated by a different law. Classical robust

frequentist approaches such as Huber's M-estimators and Hampel's influence-curve framework aim to limit the impact of atypical observations [Huber, 1992, Hampel, 1974]. In the Bayesian paradigm, generalised posteriors replace the likelihood with a loss or divergence to maintain coherent inference under model misspecification [Bissiri et al., 2016, Grünwald and Van Ommen, 2017, Jewson et al., 2018, Knoblauch et al., 2022] and extend to intractable likelihood problems [Matsubara et al., 2024]. Divergence-based updates built on the Maximum Mean Discrepancy (MMD) [Gretton et al., 2012] further reduce sensitivity to contamination without requiring precise knowledge of the misspecification form [Briol et al., 2019, Alquier and Gerber, 2024]. Although these techniques mitigate bias from regression-model misspecification, they assume accurately measured covariates. Extensions to ME settings are nontrivial because the influence function calculations and divergence minimisations rely on unbiased estimating equations in the observed covariates. Replacing the latter by proxies can violates key regularity conditions.

*Misspecification in the ME mechanism* is also damaging: Yi and Yan [2021] show that even advanced corrections become biased when the assumed error law is wrong in parameter estimation and hypothesis tests. Roy and Banerjee [2006] develop an EM algorithm for generalised linear models (GLM) with heavy-tailed ME (Student-t) and potentially multimodal covariate distributions, but they rely on external validation data to identify the ME variance. Later, Cabral et al. [2014] relax this requirement by imposing a Student-t family for the ME distribution and estimating its parameters via EM, though their framework is restricted to linear regression models. More flexible estimators, such as the phase-function technique of Delaigle and Hall [2016] and the Bayesian nonparametric formulations of Dellaporta and Damoulas [2023] avoid the need for a known ME distribution, but neither deal with regression model misspecification.

Existing literature has presented several strands of work that mitigate either ME bias or misspecification, and a smaller but growing body attempt to address them *simultaneously*. Corrected-score methods, which adjust the score equations so that their expectation

remains zero in the presence of classical ME [Nakamura, 1990], were extended by Huang [2014], who documented pathological multiple roots under sizeable ME and proposed remedial constraints, and later by Huang [2016], who analysed the existence ME and model misspecification in GLMs. Both still require accurate knowledge of error moments. In the longitudinal setting, Zhang et al. [2018] propose robust estimating equations that downweight outliers while correcting for ME, yet their approach likewise depends on replicate measurements and linearity. Doubly-robust estimators from causal inference combine outcome and exposure models to guard against single-model failure but still presuppose one correctly specified component [Funk et al., 2011, Ghassami et al., 2022]. Recent Bayesian nonparametric work assigns a Dirichlet process (DP) prior to latent covariates [Dellaporta and Damoulas, 2023], but it achieves consistency only when the scale of ME goes to zero. Thus, despite incremental progress, existing methods hinge on strong auxiliary data or restrictive assumptions. This gap has been recognised across various settings, prompting repeated calls for methods that can jointly and robustly address ME and misspecification [Gustafson, 2002, Hu et al., 2020, Zhou et al., 2023].

To answer these calls, we define what we call a framework targeting *total robustness*: the first approach, to the best of our knowledge, that simultaneously accommodates covariate ME, misspecification of the regression model, and misspecification of the ME distribution within a general nonlinear regression setting. The basis of our framework is *Bayesian nonparametrics*, which provides the ability to model unknown distributions and incorporate prior information via DPs, yielding the flexibility required for Bayesian total robustness. To explain this claim and situate our contribution, we now survey the existing Bayesian nonparametric work on robustness.

*Bayesian nonparametric* methods are widely used for flexible modelling and, among other benefits, can reduce sensitivity to model misspecification. DP mixtures flexibly describe unknown distributions, such as heavy-tailed residuals or random effects, thus reducing the dependence on distributional assumptions [Müller and Roeder, 1997, Neal, 2000, Lee et al.,

2020]. Gaussian process (GP) priors, meanwhile, allow decision makers to place nonparametric priors over functions, facilitating complex regression relations to be captured without fixing a functional form [Gramacy, 2020, Section 5];[Zhou et al., 2023]. The Bayesian semiparametric regression of Sarkar et al. [2014] combines B-spline mixtures with a DP mixture prior for the covariate density to target some classes of heteroscedastic ME but presumes the existence of replicated measurements and does not deliver theoretical guarantees. The closest prior work to ours is Dellaporta and Damoulas [2023], which addresses ME by placing a single DP prior on the latent-covariate distribution and infers model parameters via the posterior bootstrap. However, it offers no explicit guarantees when the regression model is misspecified. Our framework, by contrast, assigns a DP to the *joint* distribution of the latent covariate and response, explicitly disentangling ME uncertainty from regression-model misspecification while still preserving their dependence, and updates this joint posterior through latent-variable pseudo-sampling. This design yields posterior summaries that separate uncertainty arising from ME and from different aspects of model misspecification, and consistency guarantees that hold even when the scale of ME does not vanish, overcoming a key limitation of Dellaporta and Damoulas [2023].

We therefore present a *unified framework for total robustness* based on Bayesian nonparametric learning (NPL)[Lyddon et al., 2018, Fong et al., 2019] that simultaneously handles covariate ME, regression model misspecification, and ME distribution misspecification in general nonlinear models. Our framework is deliberately flexible: decision makers can tune prior strength and choose whether to sample latent covariates or to work directly with ME-prone observations. At the same time, our theory isolates the effect of each decision through generalisation bounds and convergence properties of the resulting estimators. This closes the long-standing robustness gap and offers a concise blueprint for trustworthy regression in complex, error-prone, and data-driven settings.

### 1.1. Problem setting

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a complete probability space on which all random variables are defined. For a fixed dimension $d \geqslant 1$, let $\mathcal{X}, \mathcal{W} \subseteq \mathbb{R}^d$ denote the spaces of latent covariates $X$ and their noisy observations $W$, and let $\mathcal{Y} \subseteq \mathbb{R}$ denote the outcome space. We observe i.i.d. pairs $\{(W_i, Y_i)\}_{i=1}^n \sim \mathbb{P}^0_{WY}$.

The covariates are generated by a mechanism involving a latent $X$ and one of two standard forms of ME (demonstrated in Figure 1; application examples listed in Table 1):

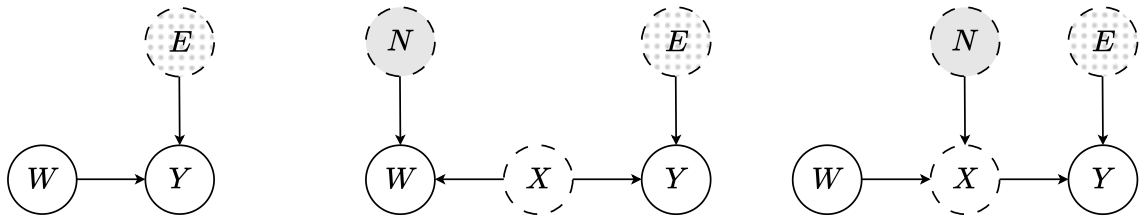
$$\textit{Classical ME:} \quad X_i \sim P_X^0, \quad N_i \sim F_N^0, \quad N_i \perp X_i, \quad W_i = X_i + N_i, \qquad \text{(Fig. 1b)}$$

$$\textit{Berkson ME:} \quad W_i \sim P_W^0, \quad N_i \sim F_N^0, \quad N_i \perp W_i, \quad X_i = W_i + N_i. \qquad \text{(Fig. 1c)}$$

The response relates to the latent covariate via a nonlinear regression function,

$$Y_i = g^0(X_i) + E_i, \qquad E_i \sim F_E^0, \quad E_i \perp (X_i, N_i).$$

Throughout, $P_\cdot^0$ denotes the unknown data-generating law, whereas $P_\cdot$ (without superscript) denotes the working model. Departures from $P_\cdot$ to $P_\cdot^0$ can arise at three distinct levels:




(a) Ignoring ME (b) Classical ME (c) Berkson ME

Fig. 1: Graphical representation of ME in the covariate and model misspecification.

*ME mechanism.* The working ME density may deviate from the truth. A common example is scale miscalibration $f_{N,\tau}(u) = \tau^{-1} f_0(u/\tau)$. See Yi and Yan [2021] and Dellaporta and Damoulas [2023] for examples.

*Regression function.* The parametric family chosen by the decision maker $\{g(\cdot, \theta) : \theta \in \Theta\}$, where $g : \mathcal{X} \times \Theta \to \mathbb{R}$ is a nonlinear parametric regression function indexed by $\theta \in \Theta \subseteq \mathbb{R}^p$,

| **ME type** | **Example** | **Literature** |
|---|---|---|
| Classical | Economic study of Engel curves<br>$X$ = true household expenditure<br>$W$ = self-reported expenditure from surveys<br>$Y$ = budget share of some commodity (e.g. food) | Hausman et al. [1995] |
| | Effect of potassium intake on health outcomes<br>$X$ = true long-term intake of potassium<br>$W$ = potassium intake converted from self-reported diet<br>$Y$ = systolic blood pressure | Curley [2022] |
| Berkson | Relationship between body fat level and risk of diabetes<br>$X$ = true body fat percentage<br>$W$ = BMI predicted body fat percentage<br>$Y$ = blood sugar level (HbA1c) | Haber et al. [2021] |
| | Effect of air pollution on respiratory health<br>$X$ = true exposure to pollution for each individual<br>$W$ = state-level average of pollution measure<br>$Y$ = respiratory health indicator | [Schennach, 2013, Section 6] |

Table 1: Applications that involve classical or Berkson ME. For each example we specify the latent covariate $X$, its noisy observation $W$ and the response $Y$.

need not contain the regression function: $g^0(\cdot) \notin \big\{g(\cdot,\theta) : \theta \in \Theta\big\}$.

*Outcome noise.* Heavy tails, contamination, or heteroskedasticity may render the working distribution $F_E$ different from the true $F_E^0$. A well-known class of outcome noise misspecification is the Huber contamination model [Huber, 1992], where $F_E^0 = (1-\eta)F_E + \eta Q_E$ with contamination ratio $\eta \in [0, 1)$.

Table 2 gathers a set of notable references on regression with ME, organised by the error mechanism (classical or Berkson), the regression type (linear vs. nonlinear) and robustness to model misspecification. Our goal is to find $\theta_0 \in \Theta$ such that $g(\cdot,\theta_0)$ recovers $g^0(\cdot)$ as accurately as possible, despite the existence of ME and the misspecification coming from all aspects of the model. In Section 2, we formalise this by defining the *optimal* estimator in the Maximum Mean Discrepancy (MMD) sense, $\theta_0 = \arg\min_{\theta\in\Theta} \mathrm{MMD}\big(\mathbb{P}^0_{XY}, \mathbb{P}^\theta_{XY}\big)$, where $\mathbb{P}^\theta_{X,Y}$ denotes the joint law of $(X, Y)$ induced by the working regression function $g(\cdot,\theta)$ and the outcome noise model $F_E$. The MMD is the distance between kernel mean embeddings of probability measures in a Reproducing Kernel Hilbert Space (RKHS). With characteristic kernels, the MMD is a robust metric that has been a popular choice as an optimisation target in recent robustness literature. It limits the influence of outliers under bounded kernels, and

admits unbiased U-statistic estimators with straightforward stochastic gradient calculations [Gretton et al., 2012, Briol et al., 2019, Alquier and Gerber, 2024, Chérief-Abdellatif and Näf, 2025].

| **Method** | **Error type** | **Regression type** | **RF** | **RN** | **MEM** |
|---|---|---|---|---|---|
| Deming [1943] | C | Linear | × | × | × |
| Berkson [1950] | B | Linear | × | × | × |
| Zamar [1989] | C | Linear | × | × | ✓ |
| Nakamura [1990] | C | Nonlinear (GLM) | × | × | × |
| Cook and Stefanski [1994] | C | Nonlinear (parametric) | × | × | × |
| Berry et al. [2002] | C | Nonlinear (splines) | ✓ | × | × |
| Schennach [2013] | B | Nonlinear (Instrumental Variable) | × | × | ✓ |
| Zhou et al. [2023] | C | Nonlinear (Gaussian Process) | ✓ | × | ✓ |
| Dellaporta and Damoulas [2023] | C *or* B | Nonlinear (parametric) | × | × | ✓ |
| **Present paper (2025)** | C *or* B | Nonlinear (parametric) | ✓ | ✓ | ✓ |

Table 2: Representative methods for regression with ME. C = classical ME; B = Berkson ME. RF = Target regression–function misspecification; RN = Target response-noise misspecification; MEM = Target misspecification of the ME distribution.

### *1.2. Main contributions*

We develop a unified framework that *simultaneously* corrects Berkson or classical ME in $X$ under the joint misspecification of (i) the regression model, (ii) the response-noise law, and (iii) the ME distribution. The framework is highly flexible: prior strength can be tuned based on confidence levels, and the analyst may either pseudo-sample latent $X$ or work directly with ME-prone observations $W$ depending on the scale of ME and the reliability of the pseudo-sampling procedure. We provide a thorough theoretical assessment via finite-sample generalisation bounds that offer interpretable guarantees through a decomposition of excess risk. We also establish consistency of the NPL estimator across variants of the framework. Practically, we implement a posterior bootstrap that combines Hamiltonian Monte Carlo (HMC)-based pseudo-sampling of latent $X$ with gradient-based MMD minimisation. In simulations and two real-data studies, the method yields lower estimation error and greater stability to misspecification and increasing ME than recent robust Bayesian and frequentist competitors.

### *1.3. Organisation of paper*

Section 2 details the proposed framework, combining latent-variable representations of Berkson or classical error with an MMD objective and an HMC sampler. Section 3 establishes generalisation bounds for different methods within our framework, as well as consistency properties for our estimators under different prior and pseudo-sampling settings. Sections 4 and 5 report results on synthetic and real data, respectively, and Section 6 concludes with limitations and future directions. Code to reproduce all results in this paper is available at `https://github.com/MengqiChenMC/tot_robust_code`.

## 2. Methodology

Our methodology follows a single principle. We retain parametric structure only for the inferential target of the regression function $g(\cdot, \theta)$, and we model all remaining components nonparametrically. In the spirit of Bayesian NPL, we place a DP prior on the unknown joint law $\mathbb{P}_{X,Y}$, covering both Berkson and classical ME regimes without fixing a parametric likelihood. The regression parameter $\theta$ enters only through $g(\cdot, \theta)$. It is learned via minimising the MMD between the nonparametric DP posterior for the joint distribution of $(X, Y)$ and the $\theta$-implied joint distribution of $(X, Y)$. The latent covariate values required to update the DP are handled by posterior pseudo-sampling. This gives us a coherent MMD-based Bayesian framework that remains robust when ME is non-negligible and when the ME mechanism or the outcome model is misspecified.

### *2.1. MMD-Based target for NPL*

We begin by recalling the definition of the MMD, our chosen loss, from Gretton et al. [2012].

**Definition 1** (MMD with characteristic kernel)**.** Given an RKHS $(\mathcal{H}, k)$ with characteristic kernel $k$, the *Maximum Mean Discrepancy* between two probability measures $P$ and $Q$ on $\mathcal{X}$ is $\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}$, where $\mu_P(\cdot) := \int k(\cdot, x)\mathrm{d}P(x)$.

Here, $k$ being characteristic means that $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$ [Gretton et al., 2012]. This is satisfied by common kernel choices (e.g. Gaussian kernels, Matérn kernels, Laplace kernels). Therefore, minimising MMD aligns two distributions without requiring a correctly specified likelihood. MMD-based loss functions are a popular choice in the robustness literature: by comparing distributions in feature space, it downweights atypical observations, maintaining reliability under heavy tails, outliers, or model misspecification [Briol et al., 2019, Alquier and Gerber, 2024]. These properties make MMD a natural choice for our setting, where both the regression model and the ME mechanism can be misspecified.

To formalise our loss target, assume hypothetically that we know the true conditional laws $\mathbb{P}^0_{X|w}$ (capturing ME) and $\mathbb{P}^0_{Y|x}$ (capturing the regression model). Then, we could form

$$\mathbb{P}^0_{XY} = \int_{\mathcal{W}} \mathbb{P}^0_{X,Y\,|\,w}\, dF^0_W(w) = \int_{\mathcal{W}} \mathbb{P}^0_{X\,|\,w} \times \mathbb{P}^0_{Y|X}\, F^0_W(\mathrm{d}w),$$

where $F^0_W$ is the marginal distribution of the observed $w$ values. Suppose we posit a parametric family $\{g(\cdot, \theta) : \theta \in \Theta\}$ for the regression function. Write $\mathbb{P}_{g(x,\theta)}(Y)$ for the $g(\cdot, \theta)$–induced conditional law of $Y$ given $X = x$, i.e. $\mathbb{P}_{g(x,\theta)}(\,\cdot\,) = \mathrm{Law}\{\mathrm{Y} \mid \mathrm{X} = \mathrm{x}; \theta\}$. The resulting model-implied joint distribution of $(X, Y)$ is

$$\mathbb{P}^\theta_{X,Y} = \mathbb{P}^0_X \times \mathbb{P}_{g(X,\theta)} = \int_{\mathcal{W}} \mathbb{P}^0_{X|w} \times \mathbb{P}_{g(X,\theta)}\, F^0_W(\mathrm{d}w).$$

We then define the optimal $\theta_0$ in the MMD sense: $\theta_0 = \arg\min_{\theta\in\Theta} \mathrm{MMD}\Big(\mathbb{P}^0_{XY}, \mathbb{P}^\theta_{XY}\Big)$. In reality, neither $\mathbb{P}^0_{X,Y|w}$ nor $\mathbb{P}^0_{X|w}$ is known. Therefore, we adopt a DP prior on these laws, leading to a Bayesian NPL procedure, which we will explain in detail in the next section.

### *2.2. DP-Based Bayesian NPL Framework*

To model the latent covariates $X$ and the response $Y$ without restrictive parametric assumptions, we place a DP prior, $\mathrm{DP}(c, F)$, on the joint law $\mathbb{P}_{X,Y|w}$ [Ferguson, 1973].

**Definition 2.** For every measurable partition $(A_1, \ldots, A_k)$ of the sample space, a random

measure $P \sim \mathrm{DP}(c, F)$ satisfies $(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(cF(A_1), \ldots, cF(A_k))$.

The concentration parameter $c > 0$ determines how tightly draws of $P$ concentrate around the base measure $F$. A key property for our framework is *conjugacy*: after observing data $z_{1:m}$, the posterior remains a DP, $P \mid z_{1:m} \sim \mathrm{DP}\left(c+m, \frac{c}{c+m}F + \frac{1}{c+m}\sum_{j=1}^{m}\delta_{z_j}\right)$. Hence each posterior draw simply reweights the empirical atoms and the prior.

### 2.2.1. General Formulation

Let $\{w_i\}_{i=1}^n$ be the observed noisy covariates, with unknown true $\{x_i\}_{i=1}^n$. For each $w_i$, we define a base measure $\mathbb{Q}_{XY,i}$ that reflects any initial beliefs about $(X, Y)$ (discussed in Section 2.2.2). Given data $\{(\tilde{x}_{i,j}, y_i)\}_{j=1}^m$, where $\{\tilde{x}_{i,j}\}_{j=1}^m$ are posterior pseudo-samples of the latent true covariate $X$ (discussed in Section 2.2.3) from the distribution of the unobserved $x_i$ given $y_i$ and $w_i$, we propose the DP framework by conjugacy:

$$\text{Prior: } \mathbb{P}_{X,Y|w_i} \sim \mathrm{DP}\big(c, \mathbb{Q}_{XY,i}\big), \tag{1}$$

$$\text{Posterior: } \mathbb{P}_{X,Y|w_i} \mid \big\{(\tilde{x}_{i,j}, y_i)\big\}_{j=1}^m \sim \mathrm{DP}\Big(c+m, \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{1}{c+m}\sum_{j=1}^{m}\delta_{(\tilde{x}_{i,j}, y_i)}\Big). \tag{2}$$

Each posterior realisation from the posterior DP (2) is a distribution over $(X, Y)$.

### 2.2.2. Constructing the Prior Centring Measure $\mathbb{Q}_{XY,i}$

The joint prior centring measures $\mathbb{Q}_{XY,i}$ encodes prior knowledge about $(X, Y)$. In practice, this could be a prior built upon historical data or domain-specific knowledge. In the absence of this prior information, we can often define $\mathbb{Q}_{XY,i}$ via three separate components - (a) a prior on $\theta$, (b) a prior on $X|W$, and (c) a prior for $Y|X$:

(a) Prior on $\theta$: suppose $\theta \sim f(\theta)$, e.g. a normal distribution around some initial estimate or uniform distribution in some compact parameter space $\Theta$.

(b) Prior on $X|W$: We denote the marginal prior for $\mathbb{P}_{X|w_i}$ as $\mathbb{Q}_{X,i}$, which needs to be defined differently in the classical and Berkson ME cases. Denote $F_N$ as our assumed

ME distribution, noting that it is not necessarily the same as the true ME $F_N^0$.

(i) *Berkson:* We have $X = W + N,\ N \perp\!\!\!\perp W,\ N \sim F_N$, so $\mathbb{Q}_{X,i}$ is simply

$$\mathbb{Q}_{X,i}(x) = F_N(x - w_i).$$

(ii) *Classical:* We have $W = X + N,\ N \perp\!\!\!\perp X,\ N \sim F_N$. We need to further assume the marginal distribution of $X \sim \mathbb{P}_X$, which also need not equal the true $\mathbb{P}_X^0$. Then

$$\mathbb{Q}_{X,i}(x) = \frac{\mathbb{P}_{W \mid X=x}(w_i)P_X(x)}{\int_{\mathcal{X}} \mathbb{P}_{W \mid X=t}(w_i)P_X(t)dt} = \frac{F_N(w_i - x)P_X(x)}{\int_{\mathcal{X}} F_N(w_i - t)P_X(t)dt}.$$

(c) Prior for $\mathbb{P}_{Y|X}^{\text{prior}}$: the integral $\mathbb{P}_{Y|X}^{\text{prior}} = \int P\big(Y \mid X, \theta\big)\, f(\theta)\, d\theta$ often lacks closed-form for nonlinear $g(x, \theta)$. We approximate it by a Monte Carlo procedure.

Hence, $\mathbb{Q}_{XY,i}$ can be expressed as

$$\mathbb{Q}_{XY,i} \;=\; \underbrace{\mathbb{Q}_{X,i}}_{\substack{\text{prior for } X \text{ around } w_i, \text{ depending on classical or Berkson ME}}} \;\times\; \underbrace{\int P\big(Y \mid X, \theta\big)\, f(\theta)\, d\theta}_{\text{for } Y|X}.$$

This ensures $\mathbb{Q}_{XY,i}$ is sufficiently rich to capture a broad range of $(X, Y)$ configurations.

This prior construction also helps us define the corresponding marginal DP for $P_{X \mid w_i}$:

$$\text{Prior:} \quad \mathbb{P}_{X|w_i} \sim \text{DP}\Big(c,\, \mathbb{Q}_{X,i}\Big), \tag{3}$$

$$\text{Posterior:} \quad \mathbb{P}_{X|w_i} \;\Big|\; \{\tilde{x}_{i,j}\}_{j=1}^m \sim \text{DP}\left(c + m,\, \frac{c}{c+m}\, \mathbb{Q}_{X,i} + \frac{1}{c+m} \sum_{j=1}^m \delta_{\tilde{x}_{i,j}}\right). \tag{4}$$

It is easy to see that the base measures in the marginal DPs 3 and 4 are the $X$-marginal of the base measures in the joint DPs 1 and 2.

### *2.2.3. Sampling latent covariates for NPL updates*

To facilitate the nonparametric learning procedure, we require samples from the joint distribution $P_{X,Y|w_i}^0$. Since $X$ is unobserved, we employ a *pseudo-sampling* procedure to acquire

plausible realisations of the latent covariate $X_i$ given observed data $(w_i, y_i)$.

Let $\mathcal{L}(\theta, X; W, Y)$ denote the negative log-likelihood function that defines the joint model for $(X, Y, \theta)$, and let $f(\theta)$ be a prior density on the parameter $\theta$. The resulting posterior over the parameters and latent variables takes the form

$$P\big(x_{1:n},\, \theta \;\big|\; w_{1:n}, y_{1:n}\big) \propto \exp\Big\{-\mathcal{L}\big(\theta,\, x_{1:n};\, w_{1:n}, y_{1:n}\big)\Big\} f(\theta). \tag{5}$$

In general, the conditional distribution of $X_i$ given $(w_i, y_i)$ is not available in closed form due to the nonlinear structure of the outcome model $g(\theta, X)$. We therefore run Markov chain Monte Carlo (MCMC) schemes that target the joint posterior in (5). The DP update requires *independent* draws from the posterior–predictive kernel

$$\Psi_n(\mathrm{d}x \mid w_i, y_i) := \int_\Theta \Pi\big(\mathrm{d}x \mid \theta, w_i, y_i\big)\, \Pi_n\big(\mathrm{d}\theta \mid w_{1:n}, y_{1:n}\big), \tag{6}$$

so we generate $\tilde{x}_{i,j} \overset{\text{iid}}{\sim} \Psi_n(\cdot \mid w_i, y_i)$ via *posterior predictive sampling*: draw $\theta_{ij} \overset{\text{iid}}{\sim} \Pi_n(\theta \mid w_{1:n}, y_{1:n})$ and then $\tilde{x}_{i,j} \sim \Pi(\mathrm{d}x \mid \theta_{ij}, w_i, y_i)$. Classic examples include: posterior predictive checks, which sample $\theta$ and simulate replicated data for model checking [Gelman et al., 1996]; and multiple imputation, which repeatedly draws missing/latent values from the posterior predictive conditional on $\theta$ and combines analyses across imputations [Rubin, 1987].

As $n \to \infty$, the posterior $\Pi_n(d\theta \mid w_{1:n}, y_{1:n})$ concentrates around the parameter value minimising the Kullback–Leibler divergence to the data-generating model, effectively transferring the best information the observed data carries on the model parameter into the pseudo-sampling procedure. Consequently, the independence requirement $\theta_{ij} \overset{\text{iid}}{\sim} \Pi_n(\theta \mid w_{1:n}, y_{1:n})$ is asymptotically immaterial: under contraction, the bias due to the dependency of $\{x_{i,j}\}_{i=1,j=1}^{n,\ m}$ introduced by the $\theta$–mixture in (6) vanishes. For small $n$, however, near-independent draws of $\theta$ can improve exploration of a potentially dispersed posterior. A detailed theoretical assessment of this procedure is provided in Section 3.1, and we discuss dropping the conditional independence requirement in Appendix E.3.

To implement the joint posterior sampling in (5) we use HMC: its gradient–informed proposals typically yield high effective sample sizes with low autocorrelation, and independent chains parallelise naturally with modest additional cost. In theory, one could obtain independent posterior–predictive draws by running $n \times m$ independent chains, but this is computationally onerous. In practice, we run a small number ($< 10$) of well–mixed chains targeting $\Pi_n(\theta \mid w_{1:n}, y_{1:n})$ and retain sparsely spaced post–burn-in states so that autocorrelation of the retained $\theta$ is negligible. Convergence diagnostics and effective sample sizes are reported in Appendix E.1. A sensitivity analysis in Appendix E.2 shows that the resulting distributions of $\{\tilde{X}_{ij}\}$ are empirically indistinguishable from those obtained by the theoretical construction of $n \times m$ independent chains.

The pseudo-sampling scheme can be less desirable or infeasible in some settings. For example, when the scale of ME is small, acquiring information about the latent covariate $X_i$ from $(w_i, y_i)$ may not justify the extra computation. When the parameter space is high-dimensional or the model is severely misspecified, reaching stationarity can be difficult, and the information may be obscured and unreliable. In such cases, we can set $m \equiv 1$ and replace the pseudo-samples $\tilde{x}_{ij}$ with $w_i$, so that the DP posteriors are updated by $(w_i, y_i)$. This update does not coincide with the true joint distribution $\mathbb{P}^0_{XY}$ and serves as a compromise. We call this the *no–pseudo–sampling variant* of our framework. Section 3.1 provides detailed theoretical assessments of this variant relative to the pseudo-sampling scheme.

### 2.2.4. MMD target with DP posteriors

Now we have the DP posterior estimates of our target distributions $\mathbb{P}^0_{X,Y|w_i}$ and $\mathbb{P}^0_{X|w_i}$, which we denote as $\mathbb{P}^{\mathrm{DP}}_{X,Y|w_i}$ (from 2) and $\mathbb{P}^{\mathrm{DP}}_{X|w_i}$ (from 4) respectively. Collecting these DP posteriors from each group and further representing $F_W$ as the empirical distribution of $W$: $F_W = \frac{1}{n}\sum_{i=1}^n \delta_{w_i}$, we formally define the DP counterpart of our MMD target as

$$\hat{\theta}_n = \arg\min_{\theta=\Theta} \mathrm{MMD}_k \left( \frac{1}{n}\sum_{i=1}^n \mathbb{P}^{\mathrm{DP}}_{X,Y|w_i}, \left(\frac{1}{n}\sum_{i=1}^n \mathbb{P}^{\mathrm{DP}}_{X|w_i}\right) \mathbb{P}_{g(X,\theta)} \right). \tag{7}$$

*2.2.5. Posterior bootstrap implementation.*

For each bootstrap replicate $b = 1, \ldots, B$ we independently draw random measures $\{\mathbb{P}^{\mathrm{DP},(b)}_{X,Y|w_i}\}_{i=1}^n$ and $\{\mathbb{P}^{\mathrm{DP},(b)}_{X|w_i}\}_{i=1}^n$ from (2) and (4) respectively. We then solve

$$\hat{\theta}_{n,b} = \underset{\theta=\Theta}{\arg\min}\, \mathrm{MMD}_k \left( \frac{1}{n}\sum_{i=1}^n \mathbb{P}^{\mathrm{DP},(b)}_{X,Y|w_i}, \left(\frac{1}{n}\sum_{i=1}^n \mathbb{P}^{\mathrm{DP},(b)}_{X|w_i}\right) \mathbb{P}_{g(X,\theta)} \right).$$

The collection $\{\hat{\theta}_b\}_{b=1}^B$ forms an empirical approximation to the posterior implied by our MMD loss, thereby placing the proposed method within the Bayesian NPL framework. Algorithm 1 in Appendix A demonstrates our bootstrapping procedure.

## 3. Theoretical Assessments

This section studies the estimator $\hat{\theta}_n$ defined above in (7). We provide two types of results. First, in Section 3.1, we derive generalisation bounds for the excess risk under model misspecification. The bounds separate three contributions: statistical fluctuation, prior–data discrepancy, and pseudo–sample discrepancy, which are weighted by the DP centring parameter $c$ and the number of pseudo-samples per observation $m$. Second, we prove consistency in Section 3.2: under regularity and identifiability conditions, $\hat{\theta}_n$ converges in distribution to the minimiser of the limiting MMD, which coincides with the data–generating parameter under correct model specification. Proofs are supplied in Appendix B.

We now introduce the construction used in both results and define the notation. Let $\mathcal{D} = \{(W_i, Y_i)\}_{i=1}^n$ be the observed sample, drawn i.i.d. from $p^0_{WY}$. Fix a pseudo–sample size $m \geqslant 1$. Given $\mathcal{D}$, define the posterior–predictive kernel $\Psi_n(\mathrm{d}x \mid w, y) := \int_\Theta \Pi_\theta(\mathrm{d}x \mid w, y)\Pi_n(\mathrm{d}\theta \mid \mathcal{D})$, where $\Pi_n(\mathrm{d}\theta \mid \mathcal{D})$ is the (possibly misspecified) posterior for $\theta$. For each $i$, independently draw pseudo–samples $\tilde{X}_{i1}, \ldots, \tilde{X}_{im} \mid \mathcal{D} \overset{\text{iid}}{\sim} \Psi_n(\cdot \mid W_i, Y_i)$ and collect them in $S := \{\tilde{X}_{ij}\}_{i=1,j=1}^{n,\ m}$. They feed the DP update by supplying atoms for the posteriors of $(X, Y) \mid W_i$ and $X \mid W_i$.

Let $\{\mathbb{Q}_{XY,i}, \mathbb{Q}_{X,i}\}_{i=1}^n$ be prior centring measures. Given $(\mathcal{D}, S)$, draw independent realisa-

tions from the joint and marginal DP posteriors (2) and (4):

$$\begin{aligned}\mathbb{P}^{\mathrm{DP}}_{X,Y|W_i} &\sim \mathrm{DP}\left(c+m, \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{1}{c+m}\sum_{k=1}^{m}\delta_{(\tilde{X}_{ik},Y_i)}\right),\\ \mathbb{P}^{\mathrm{DP}}_{X|W_i} &\sim \mathrm{DP}\left(c+m, \frac{c}{c+m}\,\mathbb{Q}_{X,i} + \frac{1}{c+m}\sum_{k=1}^{m}\delta_{\tilde{X}_{ik}}\right).\end{aligned} \tag{8}$$

Finally, we define the DP-based MMD minimisation target and the resulting estimator $\hat{\theta}_n$

$$M_n(\theta) := \mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}^{\mathrm{DP}}_{X,Y|W_i}, \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}^{\mathrm{DP}}_{X|W_i}\right)\mathbb{P}_{g(X,\theta)}\right), \qquad \hat{\theta}_n := \arg\min_{\theta\in\Theta} M_n(\theta).$$

To facilitate theoretical assessments, we define the following notation for the average prior measures and empirical measures for the pseudo–samples:

$$\mathbb{P}^{\mathrm{prior}}_{XY} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{Q}_{XY,i},\ \mathbb{P}^{\mathrm{prior}}_{X} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{Q}_{X,i}; \quad \mathbb{P}^{\mathrm{pseudo}}_{XY} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(\tilde{X}_{ij},Y_i)},\ \mathbb{P}^{\mathrm{pseudo}}_{X} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{\tilde{X}_{ij}}.$$

We also define the average base measures in the DP posteriors (8):

$$\mathbb{P}^{\mathrm{base}}_{XY} := \frac{c}{c+m}\,\mathbb{P}^{\mathrm{prior}}_{XY} + \frac{m}{c+m}\,\mathbb{P}^{\mathrm{pseudo}}_{XY}, \quad \mathbb{P}^{\mathrm{base}}_{X} := \frac{c}{c+m}\,\mathbb{P}^{\mathrm{prior}}_{X} + \frac{m}{c+m}\,\mathbb{P}^{\mathrm{pseudo}}_{X}.$$

All randomness is defined on $\Pr = \Pr_{\mathcal{D},S}\,\Pr_{\mathrm{DP}|\mathcal{D},S}$, where $\Pr_{\mathcal{D},S}$ is the joint law of $(\mathcal{D}, S)$ under the DGP and the pseudo–sampling scheme, and $\Pr_{\mathrm{DP}|\mathcal{D},S}$ is the conditional law of the DP realisations given $(\mathcal{D}, S)$. We write

$$\mathbb{E}_{\mathcal{D},S}[\cdot] := \mathbb{E}_{\Pr_{\mathcal{D},S}}[\cdot], \qquad \mathbb{E}_{\mathrm{DP}}[\cdot \mid \mathcal{D}, S] := \mathbb{E}_{\Pr_{\mathrm{DP}|\mathcal{D},S}}[\cdot],$$

Unless noted otherwise, expectations are taken with respect to Pr.

If the decision maker chooses not to perform pseudo-sampling as discussed at the end of Section 2.2.3, they can set $m \equiv 1$ and replace the pseudo samples by the observed covariates, which gives $\mathbb{P}^{\mathrm{base}}_{XY,i} = \frac{c}{c+1}\,\mathbb{Q}_{XY,i} + \frac{1}{c+1}\delta_{(W_i,Y_i)}$ and $\mathbb{P}^{\mathrm{base}}_{X,i} = \frac{c}{c+1}\,\mathbb{Q}_{X,i} + \frac{1}{c+1}\delta_{W_i}$. The empirical laws corresponding to $\mathbb{P}^{\mathrm{pseudo}}_{XY}$ and $\mathbb{P}^{\mathrm{pseudo}}_{X}$ reduce to $\hat{\mathbb{P}}^{n}_{WY} = \frac{1}{n}\sum_{i=1}^{n}\delta_{(W_i,Y_i)}$ and $\hat{\mathbb{P}}^{n}_{W} = \frac{1}{n}\sum_{i=1}^{n}\delta_{W_i}$. All other notation remains unchanged. This formulation is investigated in Section 3.1.3.

We impose two standing conditions that apply throughout the theory section.

G1 The MMD kernel on $\mathcal{X} \times \mathcal{Y}$ is $k\big((x,y),(x',y')\big) = k_X(x,x')\,k_Y(y,y')$, where $k_X, k_Y$ are measurable, positive–definite, and characteristic, with $\sup_{x,x'} k_X(x,x') = \kappa_X < \infty$ and $\sup_{y,y'} k_Y(y,y') = \kappa_Y < \infty$. Consequently $k$ is measurable, positive–definite, and characteristic on $\mathcal{X} \times \mathcal{Y}$ with $\kappa := \sup k(\cdot,\cdot) = \kappa_X \kappa_Y < \infty$.

G2 Let $\mathcal{H}_X, \mathcal{H}_Y$ be the RKHSs of $k_X, k_Y$. For each $\theta \in \Theta$, the conditional mean embedding $\mu_\theta(x) := \int_{\mathcal{Y}} k_Y(y,\cdot)\, \mathbb{P}_{g(x,\theta)}(\mathrm{d}y) \in \mathcal{H}_Y$ extends to a bounded linear operator $\mathcal{C}_\theta : \mathcal{H}_X \to \mathcal{H}_Y$ with $\sup_{\theta\in\Theta} \|\mathcal{C}_\theta\|_{\mathrm{op}} = \Lambda < \infty$, where $\|\cdot\|_{\mathrm{op}}$ is the operator norm.

*Remark* 1. (a) Common bounded, characteristic kernels such as the Gaussian, Laplace, Matérn and rational–quadratic kernels all satisfy G1.

(b) Assumption G2 is commonly assumed in robust regression methods involving kernel mean embeddings [Alquier and Gerber, 2024, Dellaporta and Damoulas, 2023]. We need it to apply Lemma 2 of [Alquier and Gerber, 2024]: under G2, we have

$$\mathrm{MMD}_k\left(\mathbb{P}_X\, \mathbb{P}_{g(\theta,x)}, \mathbb{P}'_X\, \mathbb{P}_{g(\theta,x)}\right) \leqslant \|\mathcal{C}_\theta\|_{\mathrm{op}}\, \mathrm{MMD}_{k_X^2}\left(\mathbb{P}_X, \mathbb{P}'_X\right).$$

Here $\mathrm{MMD}_{k_X^2}(\mathbb{P}_X, \mathbb{P}'_X)$ denotes the MMD computed with kernel $k_X^2 := k_X \otimes k_X$ on $\mathcal{H}_{k_X^2} := \mathcal{H}_{k_X} \otimes \mathcal{H}_{k_X}$: it compares the mean embeddings of $\mathbb{P}_X$ and $\mathbb{P}'_X$ in the tensor product space $\mathcal{H}_{k_X^2}$; equivalently, it is the MMD (with kernel $k_X^2$) between the pushforward laws of $(X,X)$ with $X \sim \mathbb{P}_X$ and of $(X',X')$ with $X' \sim \mathbb{P}'_X$ under the diagonal map $x \mapsto (x,x)$. This lemma links the MMD between joint distributions and marginals, which is necessary when $\mathbb{P}_X$ is unknown or misspecified. We refer the reader to [Alquier and Gerber, 2024] for a detailed discussion on the existence and boundedness of $\mathcal{C}_\theta$.

### *3.1. Generalisation Bound*

We present the general version of the generalisation error bound that applies to both Berkson and classical ME settings.

**Theorem 1** (Generalisation error bound)**.** *Under assumptions G1-G2:*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},S}\left[\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X\,\mathbb{P}_{g(X,\hat{\theta}_n)}\right) \mid \mathcal{D}, S\right]\right] - \inf_{\theta\in\Theta}\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X\,\mathbb{P}_{g(X,\theta)}\right) \\
&\leqslant \underbrace{\frac{2\sqrt{\kappa}(1+\Lambda)}{\sqrt{n(c+m+1)}}}_{\textit{statistical fluctuation}} + \underbrace{\frac{2c}{c+m}\,\mathbb{E}_{\mathcal{D},S}\left[\Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathbb{P}^0_X, \mathbb{P}^{\mathrm{prior}}_X\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^{\mathrm{prior}}_{XY}\right)\right]}_{\textit{prior model discrepancy, vanishes as } c/m\to 0} \\
&\qquad + \underbrace{\frac{2m}{c+m}\,\mathbb{E}_{\mathcal{D},S}\left[\Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathbb{P}^0_X, \mathbb{P}^{\mathrm{pseudo}}_X\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^{\mathrm{pseudo}}_{XY}\right)\right]}_{\textit{pseudo-sample distribution discrepancy}}.
\end{aligned}$$

The first term, $2\sqrt{\kappa}(1+\Lambda)\{n(c+m+1)\}^{-1/2}$, arises from DP variations.

The second term measures how far the chosen centring measures $(\mathbb{Q}_{XY,i}, \mathbb{Q}_{X,i})$ are from the true DGP. Its weight is $c/(c+m)$. Thus we can borrow strength from a well-calibrated prior by taking a moderate or large value of $c$. Conversely, if historical information is unreliable, we can downweight it by letting $c/m \to 0$, after which this term vanishes.

The final line quantifies the error introduced by the latent covariate pseudo-sampling scheme. Its weight is $m/(c+m)$. We will show in Theorem 2 that, for any sequence $M_n \to \infty$,

$$\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_k\left(\mathbb{P}^{\mathrm{pseudo}}_{XY}, \mathbb{P}^0_{XY}\right)\right] = O\left(\frac{M_n}{\sqrt{n}} + \sqrt{1-e^{-\mathrm{KL}*}} + r_n\right),$$

with an analogous bound for the $X$-marginal. The term $\sqrt{1-e^{-\mathrm{KL}*}}$, which will be formally defined in Theorem 2, measures the *total misspecification*: it is zero when the working model is correct and otherwise gives the best error attainable under the chosen family.

*Remark* 2. Theorem 2 yields a remainder term $r_n \to 0$. Since the misspecified Bernstein–von Mises theorem [Kleijn and Van der Vaart, 2012] required for this bound is asymptotic, no general rate for $r_n$ is available. We therefore retain $r_n$ explicitly in the bound.

If one adopts the no-pseudo-sampling variant (end of Subsection 2.2.3) and $\mathbb{P}^{\mathrm{pseudo}}_{XY}$ reduces to $\hat{\mathbb{P}}^n_{WY} := \sum_{i=1}^n \delta_{(W_i,Y_i)}$, we will show a replacement bound in Lemma 1 that depends solely

on the true ME distribution:

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k(\hat{\mathbb{P}}^n_{WY}, \mathbb{P}^0_{XY})\right] = O\left(\frac{1}{\sqrt{n}} + \sqrt{\mathrm{MMD}_{k_X}(F^0_N, \delta_0)}\right) \tag{9}$$

These three pieces provide a clear strategy. A decision-maker can start with a relatively informative prior (moderate $c$) to exploit domain knowledge. If such knowledge is unreliable, reducing $c$ or increasing the pseudo-sample size $m$ shifts weight to the empirical component.

### *3.1.1. Pseudo-sampling Bounds: Classical ME*

We now control the pseudo–sample discrepancy term in Theorem 1 under *classical* ME. Let the true DGP under classical ME be

$$W = X + N,\ Y = g_0(X) + E,\ X \sim P^0_X,\ N \sim F^0_N,\ E \sim F^0_E,\ N, E \perp\!\!\!\perp X,\ N \perp\!\!\!\perp E. \tag{10}$$

The joint density of the observed $(W, Y)$ is $p^0_{WY}(w, y) = \int_{\mathcal{X}} p^0_X(x) f^0_N(w - x) f^0_E\big(y - g_0(x)\big)\mathrm{d}x$.

We work under a misspecified model

$$W = X + N,\ Y = g(X, \theta) + E,\ X \sim P_X,\ N \sim F_N,\ E \sim F_E,\ N, E \perp\!\!\!\perp X,\ N \perp\!\!\!\perp E,$$

and $g_0(\cdot) \notin \{g(\cdot, \theta) : \theta \in \Theta\}$. Here $\Theta \subset \mathbb{R}^d$ is compact. For each $\theta$ the induced density of $(W, Y)$ is $p^\theta_{WY}(w, y) = \int_{\mathcal{X}} p_X(x) f_N(w-x) f_E\big(y-g(x,\theta)\big)\mathrm{d}x$. Define the pseudo-true parameter by $\theta^* := \arg\min_{\theta\in\Theta} \mathrm{KL}\big(p^0_{WY} \| p^\theta_{WY}\big)$. We make the following assumptions:

A1 The family $\{p^\theta_{WY} : \theta \in \Theta\}$ satisfies the conditions of [Kleijn and Van der Vaart, 2012, Theorem 3.1] around $\theta^*$, ensuring posterior convergence.

A2 There exists a neighbourhood $\Theta_\rho$ of $\theta^*$ such that, for every $(w, y)$ and all $\theta_1, \theta_2 \in \Theta_\rho$, we have $\mathrm{MMD}_k\left(\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\right) \leqslant L(w, y)\,\|\theta_1 - \theta_2\|$ with $\mathbb{E}_{p^0_{WY}}\, L(W, Y) < \infty$.

*Remark* 3. A1 requires Bernstein-von Mises-type assumptions for posterior contraction, which collect the LAN, smoothness and integrability conditions of Kleijn and Van der Vaart [2012,

Theorem 3.1]. Assumption A2 is a local stability condition that transfers this contraction to the posterior predictive. Appendix C relists sufficient conditions for A1–A2 and provides practical scenarios where they hold.

**Theorem 2.** *For each $x \in \mathcal{X}$ write $p^0_{Y|x} := f^0_E(y - g_0(x))$ and $p^{\theta^*}_{Y|x} := f_E\big(y - g(\theta^*, x)\big)$. Let*

$$\mathrm{KL}_X := \mathrm{KL}\big(p^0_X \| p_X\big), \quad \mathrm{KL}_N := \mathrm{KL}\big(f^0_N \| f_N\big), \quad \mathrm{KL}_E := \mathbb{E}_{X\sim p^0_X}\,\mathrm{KL}\big(p^0_{Y|X} \| p^{\theta^*}_{Y|X}\big),$$

*Denote $\mathrm{KL}_* := \mathrm{KL}_X + \mathrm{KL}_N + \mathrm{KL}_E$ and recall that $\kappa = \kappa_X \kappa_Y$. Under conditions G1 and A1-A2, for all $n, m \geqslant 1$, and any sequence $M_n$ such that $M_n \to \infty$,*

$$\mathbb{E}_{\mathcal{D},S}\Big[\mathrm{MMD}_k\big(\mathbb{P}^{\mathrm{pseudo}}_{XY}, \mathbb{P}^0_{XY}\big)\Big] \leqslant \frac{4\sqrt{\kappa}}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + 2\sqrt{\kappa}\sqrt{1 - \exp(-\,\mathrm{KL}_*)} + \sqrt{\kappa} r_n,$$

*where $r_n$ depends on $M_n$ and satisfies $0 \leqslant r_n \leqslant 1$ and $r_n \to 0$.*

The pseudo–true parameter $\theta^* := \arg\min_{\theta\in\Theta} \mathrm{KL}\big(p^0_{WY} \| p^\theta_{WY}\big)$ is the value around which the misspecified posterior $\Pi_n(\mathrm{d}\theta \mid \mathcal{D})$ concentrates by the misspecified Bernstein–von Mises theorem; it is the best estimator attainable from the information in the error–prone pairs $(W_i, Y_i)$ within the working family $\{p^\theta_{WY}\}$ in a Bayesian posterior. Theorem 2 shows how this information is conveyed to the latent covariate distribution through the pseudo–sampling kernel $\Psi_n(\mathrm{d}x \mid w, y) = \int \Pi_\theta(\mathrm{d}x \mid w, y)\, \Pi_n(\mathrm{d}\theta \mid \mathcal{D})$: as $n$ increases, $\Psi_n(\cdot \mid w, y)$ tracks $\Pi_{\theta^*}(\cdot \mid w, y)$ and the resulting pseudo–sample joint distribution $\mathbb{P}^{\mathrm{pseudo}}_{XY}$ approaches the true joint distribution $\mathbb{P}^0_{XY}$ up to the term $\sqrt{1 - \exp(-\,\mathrm{KL}_*)}$ summarising total misspecification.

Our framework employs the ME-contaminated pairs $(W_i, Y_i)$ only to form a Bayesian predictive law for $X_i \mid (W_i, Y_i)$. It does not commit to the parametric model when estimating $\theta$. The pseudo–samples $\{\tilde{X}_{ij}\}$ propagate the information about $\theta$ learned from the data into the nonparametric stage by updating the DP priors. The final estimator is the $\theta$ that best aligns the DP–updated joint distribution with the model–implied joint under the MMD.

This construction is modular: any (generalised) posterior for $\theta$ that contracts to a limiting pseudo–truth can be used inside $\Psi_n(\cdot \mid w, y)$ (e.g., MMD–Bayes [Chérief-Abdellatif and

Alquier, 2020] or $\alpha$-posteriors [Medina et al., 2022]) and produce pseudo-samples $\{\tilde{X}_{ij}\}$ corresponding to the generalised posterior. Under Assumption A2, the pseudo–samples inherit the required stability, and extending the bounds to a generalised posterior amounts to verifying the analogue of Assumption A1 (posterior contraction under misspecification).

The quantity $\sqrt{1-\exp(-\mathrm{KL}_*)}$ in the bound provides a summary of *total* misspecification of the working model relative to the DGP, with $\mathrm{KL}_* = \mathrm{KL}_X + \mathrm{KL}_N + \mathrm{KL}_E = 0$ if and only if the latent covariate law, the ME distribution, and the outcome model are correctly specified. A sharper alternative is to replace $\sqrt{1-e^{-\mathrm{KL}_*}}$ by $\mathrm{MMD}_k\big(\Pi^{XY}_{\theta*}, \mathbb{P}^0_{XY}\big)$, where

$$\Pi^{XY}_{\theta*}(\mathrm{d}x, \mathrm{d}y) = \int_{\mathcal{W}\times\mathcal{Y}} \Pi_{\theta*}(\mathrm{d}x \mid w, y)\, \delta_y(\mathrm{d}y)\, p^0_{WY}(\mathrm{d}w, \mathrm{d}y).$$

Although this alternative does not explicitly separate the contributions of the different sources of misspecification, it avoids relying on the KL divergence and does not collapse to a trivial bound even when some component KLs are infinite.

**Proposition 1** (Classical ME: Marginal–$X$)**.** *Define the kernel $k_X^2 := k_X \otimes k_X$ with RKHS $\mathcal{H}_{k_X^2} := \mathcal{H}_{k_X} \otimes \mathcal{H}_{k_X}$. Under assumptions G1 and A1-A2, for all $n, m \geqslant 1$ and every $M_n \to \infty$,*

$$\mathbb{E}_{\mathcal{D},S}\big[\mathrm{MMD}_{k_X^2}(\mathbb{P}_X^{\mathrm{pseudo}}, \mathbb{P}_X^0)\big] \leqslant \frac{4\kappa_X}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + 2\kappa_X\sqrt{1-\exp(-\mathrm{KL}_*)} + \kappa_X r_n.$$

### *3.1.2. Pseudo-sampling Bounds: Berkson ME*

Next, we demonstrate the *Berkson* ME counterpart for Theorem 2. Let the true DGP be

$$X = W + N, \quad Y = g_0(X) + E, \quad N \sim F_N^0, \quad E \sim F_E^0, \quad N, E \perp\!\!\!\perp W,\ N \perp\!\!\!\perp E, \tag{11}$$

where the covariates $W_1, \ldots, W_n$ are i.i.d. draws from a *known and fixed* distribution $p^0_W(w)$. This is a standard assumption in Berkson ME, where we often have $p^0_W(w) = \frac{1}{n}\sum_{i=1}^n \delta_{w_i}$, which consists of fixed design values assigned by experts [Delaigle et al., 2006].

For $\theta \in \Theta \subset \mathbb{R}^d$ consider the (misspecified) model

$$X = W + N,\ Y = g(X,\theta) + E,\ N \sim F_N,\ E \sim F_E,\ N \perp\!\!\!\perp W,\ E \perp\!\!\!\perp (N, W).$$

Then we can define the true and model-implied joint densities of $(W, Y)$ as

$$p^0_{WY}(w,y) = \int_{\mathcal{X}} p^0_W(w) f^0_N(x-w) f^0_E(y-g^0(x))\mathrm{d}x,\ p^\theta_{WY}(w,y) = \int_{\mathcal{X}} p^0_W(w) f_N(x-w) f_E(y-g(\theta,x))\mathrm{d}x$$

Define the pseudo–true parameter $\theta^* := \arg\min_{\theta\in\Theta} \mathrm{KL}(p^0_{WY} \| p^\theta_{WY})$. We set

$$\mathrm{KL}_N := \mathrm{KL}(f^0_N \| f_N),\ \mathrm{KL}_E := \int_{\mathbb{R}} \int_{\mathcal{X}} p^0_W(w) f^0_N(w-x)\, \mathrm{KL}\big(p^0_{Y|x} \| p^{\theta^*}_{Y|x}\big)\mathrm{d}x\mathrm{d}w,\ \mathrm{KL}_* := \mathrm{KL}_N + \mathrm{KL}_E\,.$$

**Proposition 2** (Berkson ME)**.** *Under Assumptions G1 and A1-A2, for all $n, m \geqslant 1$ and any sequence $M_n \to \infty$,*

$$\mathbb{E}_{\mathcal{D},S}\Big[\mathrm{MMD}_k\big(\mathbb{P}^{\mathrm{pseudo}}_{XY}, \mathbb{P}^0_{XY}\big)\Big] \leqslant \frac{4\sqrt{\kappa}}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + 2\sqrt{\kappa}\sqrt{1-\exp(-\mathrm{KL}_*)} + \sqrt{\kappa} r_n,$$

$$\mathbb{E}_{\mathcal{D},S}\Big[\mathrm{MMD}_{k^2_X}\big(\mathbb{P}^{\mathrm{pseudo}}_{X}, \mathbb{P}^0_{X}\big)\Big] \leqslant \frac{4\kappa_X}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + 2\kappa_X\sqrt{1-\exp(-\mathrm{KL}_*)} + \kappa_X r_n,$$

*where $0 \leqslant r_n \leqslant 1$ and $r_n \to 0$.*

### 3.1.3. Bounds without Pseudo-Sampling

When we replace the pseudo samples $\{\tilde{X}_{ij}\}^{n,\ m}_{i=1,j=1}$ by the observed covariates $\{W_i\}^n_{i=1}$ with $m \equiv 1$, the following replace the pseudo-sample discrepancy term in Theorem 1 or Proposition 2.

**Lemma 1.** *Let $(X, W, Y)$ be jointly distributed random variables on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ satisfying the classical ME model* (10) *or Berkson ME model* (11)*, with $N \sim F^0_N$. Define the empirical measures $\hat{\mathbb{P}}^n_W := \frac{1}{n}\sum_{i=1}^n \delta_{W_i}$. $\hat{\mathbb{P}}^n_{WY} := \frac{1}{n}\sum_{i=1}^n \delta_{(W_i,Y_i)}$.*

*We assume that $k_X$ is translation-invariant, i.e. that there exists a positive-definite function $\psi$ on $\mathcal{X}$ such that $k_X(x,x') = \psi(x-x')\ \forall x, x' \in \mathcal{X}$, then $\psi(0) = \kappa_X$ by positive-*

*definiteness of $k_X$. Under assumption G1, we have*

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \hat{\mathbb{P}}^n_{WY})\right] \leqslant \sqrt{2}\kappa_Y^{1/2}\kappa_X^{1/4}\sqrt{\mathrm{MMD}_{k_X}(F^0_N, \delta_0)} + \frac{\sqrt{\kappa}}{\sqrt{n}}, \tag{12}$$

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_{k_X^2}(\mathbb{P}^0_X, \hat{\mathbb{P}}^n_W)\right] \leqslant \sqrt{2\kappa_X\,\mathrm{MMD}_{k_X^2}(F^0_N, \delta_0)} + \frac{\kappa_X}{\sqrt{n}}. \tag{13}$$

### *3.2. Consistency*

We next establish consistency of $\hat{\theta}_n$ for the pseudo–true value $\theta^\dagger$, the unique minimiser of the limiting loss $M(\theta)$. Before specialising to our ME settings, we first establish a general consistency guarantee for MMD-based NPL procedures, a setting that is popular [e.g. Dellaporta et al., 2022, Fazeli-Asl et al., 2023, Dellaporta and Damoulas, 2023] but for which asymptotic results for the NPL estimator are limited. The forms of the limiting measures $\mathbb{P}^\infty_{XY}$ and $\mathbb{P}^\infty_X$, which depend only on the DP base measures, are given explicitly in Proposition 3.

**Theorem 3.** *Assume G1-G2. Furthermore, assume that there exist fixed probability measures $\mathbb{P}^\infty_{XY}$ and $\mathbb{P}^\infty_X$, such that*

$$\mathrm{MMD}_k\left(\mathbb{P}^{\mathrm{base}}_{XY}, \mathbb{P}^\infty_{XY}\right) \xrightarrow{\Pr} 0, \qquad \mathrm{MMD}_{k_X^2}\left(\mathbb{P}^{\mathrm{base}}_X, \mathbb{P}^\infty_X\right) \xrightarrow{\Pr} 0 \qquad \text{as } n \to \infty. \tag{14}$$

*Define $M(\theta) := \mathrm{MMD}_k(\mathbb{P}^\infty_{XY}, \mathbb{P}^\infty_X \mathbb{P}_{g(X,\theta)})$. If $\Theta$ is compact, $M(\theta)$ is continuous, and $M(\theta)$ attains its unique minimum at an interior point $\theta^\dagger$, then $\hat{\theta}_n \xrightarrow{\Pr} \theta^\dagger$.*

Condition (14) requires that the DP base measures concentrate to fixed limits. The next proposition establishes explicit forms of $\mathbb{P}^\infty_{XY}$ and $\mathbb{P}^\infty_X$ for pseudo-sampling or non-pseudo-sampling schemes under different $(c, m)$-parameter settings. We first state an assumption on the convergence of the constructed prior centring measures:

(D) There exists fixed distributions $\mathbb{Q}^\infty_{XY}$ and $\mathbb{Q}^\infty_X$ such that $\mathrm{MMD}_k(\frac{1}{n}\sum_{i=1}^n \mathbb{Q}_{XY,i}, \mathbb{Q}^\infty_{XY}) \xrightarrow{\Pr} 0$ and $\mathrm{MMD}_{k_X^2}(\frac{1}{n}\sum_{i=1}^n \mathbb{Q}_{X,i}, \mathbb{Q}^\infty_X) \xrightarrow{\Pr} 0$.

This requirement is automatically met when each centring measure $\mathbb{Q}_{XY,i}, \mathbb{Q}_{X,i}$ is a fixed,

data-independent choice (for example, a historical distribution or an expert prior). When the centring measures depend on the data, e.g., around some naive estimator, the assumption still holds as long as the preliminary estimator contracts to a point estimate as $n \to \infty$.

**Proposition 3** (Sufficient conditions for (14))**.**

*(3.a) If the DP base measures are constructed via the pseudo-sampling scheme and assumptions A1-A2 are satisfied, recall that* $\theta^* = \arg\min_{\theta\in\Theta} \mathrm{KL}(p^0_{WY} \| p^\theta_{WY})$ *and define*

$$\begin{aligned} \Pi^{XY}_{\theta^*}(\mathrm{d}x, \mathrm{d}y) &:= \int_{\mathcal{W}\times\mathcal{Y}} \Pi_{\theta^*}(\mathrm{d}x \mid w, y)\delta_y(\mathrm{d}y)p^0_{WY}(\mathrm{d}w, \mathrm{d}y), \\ \Pi^{X}_{\theta^*}(\mathrm{d}x) &:= \int_{\mathcal{W}\times\mathcal{Y}} \Pi_{\theta^*}(\mathrm{d}x \mid w, y)p^0_{WY}(\mathrm{d}w, \mathrm{d}y). \end{aligned} \tag{15}$$

*If* $c, m$ *are finite constants and (D) holds, then 14 holds with*

$$\mathbb{P}^\infty_{XY} = \frac{c}{c+m}\,\mathbb{Q}^\infty_{XY} + \frac{m}{c+m}\Pi^{XY}_{\theta^*}, \qquad \mathbb{P}^\infty_{X} = \frac{c}{c+m}\,\mathbb{Q}^\infty_{X} + \frac{m}{c+m}\Pi^{X}_{\theta^*}. \tag{16}$$

*Otherwise, if* $c/m \to 0$ *as* $n \to \infty$*, then 14 holds with* $\mathbb{P}^\infty_{XY} = \Pi^{XY}_{\theta^*}$, $\mathbb{P}^\infty_{X} = \Pi^{X}_{\theta^*}$.

*(3.b) If the DP base measures are constructed without pseudo-sampling,* $c$ *is a finite constant and (D) holds, then 14 holds with*

$$\mathbb{P}^\infty_{XY} = \frac{c}{c+1}\,\mathbb{Q}^\infty_{XY} + \frac{1}{c+1}\,\mathbb{P}^0_{WY}, \qquad \mathbb{P}^\infty_{X} = \frac{c}{c+1}\,\mathbb{Q}^\infty_{X} + \frac{1}{c+1}\,\mathbb{P}^0_{W}\,. \tag{17}$$

*Otherwise, if* $c \to 0$ *as* $n \to \infty$*, then* $\mathbb{P}^\infty_{XY} = \mathbb{P}^0_{WY}$, $\mathbb{P}^\infty_{X} = \mathbb{P}^0_{W}$.

The proposition shows that condition (D) is only required when the prior weight ratio $c/m$ does not vanish. If the ratio $c/m \to 0$ with $n$, the prior contribution is asymptotically negligible and (D) can be discarded.

*Remark* 4. We have focused on estimating $\theta$, which parametrises $g(\cdot)$. The construction extends to a joint parameter $\eta = (\theta, s) \in \Theta \times \mathcal{S}$, where $s$ parametrises $f_E(\cdot\,; s)$. All statements in Sections 3.1–3.2 then hold with $(\hat{\theta}_n, \theta^*, \theta^\dagger)$ replaced by $(\hat{\eta}_n, \eta^*, \eta^\dagger)$, provided A1–A2 are imposed on the joint family $\{p^\eta_{WY} : \eta \in \Theta \times \mathcal{S}\}$.

## 4. Synthetic Experiments

We consider the sigmoid regression model $g(x, \theta) = \theta_1 / \left[1 + \exp\left\{ -\theta_2(x - \theta_3)\right\}\right]$. It is a popular choice in synthetic experiments with nonlinear regression and is widely used in practical applications where one observes threshold behaviour [Yin et al., 2003, Klimstra and Zehr, 2008]. We study both Berkson and classical ME:

$$\text{Berkson: } X = W + N, \qquad \text{Classical: } W = X + N, \qquad F_N^0 = \mathcal{N}(0, \sigma_N^2).$$

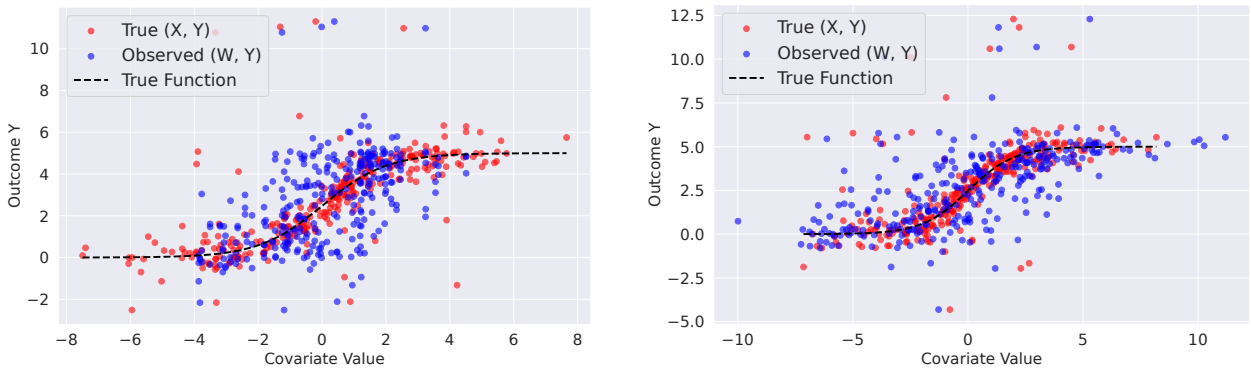
To introduce misspecification we contaminate the outcome noise using a Huber–type mixture:

$$F_E^0 = (1 - \varepsilon)\mathcal{N}(0, \sigma_E^2) + \varepsilon\mathcal{N}(0, \eta_E^2\sigma_E^2). \tag{18}$$

Figure 2 illustrates samples under both ME mechanisms with $\theta = (5, 1, 0.02)$, and $\sigma_E = 0.5$, $(\varepsilon, \eta_E) = (0.1, 9)$, $\sigma_N = 2$. We compare four estimators: *NPL–HMC* (ours); *R–MEM* [Dellaporta and Damoulas, 2023], a robust approach for ME models that places a single DP prior on the conditional law of $\mathbb{P}_{X|w_i}$, updates this prior using the observed $\{w_i\}_{i=1}^n$ values, and performs parameter inference via a posterior bootstrap; *NLS* (nonlinear least squares fitted to $(W, Y)$); and *HMC* (posterior mean from HMC chains). Performance is summarised by the root mean squared error (RMSE) over 100 independent replications. Figures 3a–3b report RMSE as the ME scale $\sigma_N$ increases, under *both* model and ME–distribution misspecification.

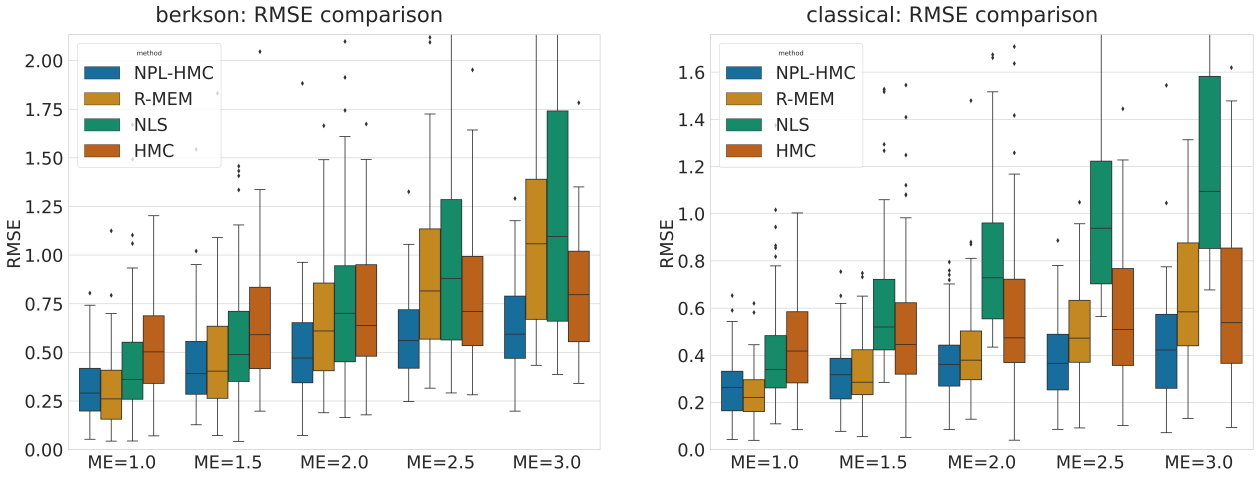
$$\begin{aligned} \text{DGP}: &\qquad N \sim \mathcal{N}(0, \sigma_N^2), \qquad E \sim (1 - \varepsilon)\,\mathcal{N}(0, \sigma_E^2) + \varepsilon\,\mathcal{N}(0, \eta_E^2\sigma_E^2) \\ \text{Model}: &\qquad N \sim \mathcal{N}(0, \tau_N^2\sigma_N^2), \qquad E \sim \mathcal{N}(0, \tau_E^2\sigma_E^2) \end{aligned}$$

We fix $n = 300$, $(\varepsilon, \eta_E) = (0.1, 9)$, and $(\tau_N, \tau_E) = (0.7, 2)$. In both Berkson and classical settings, NPL–HMC attains the lowest median RMSE and interquartile ranges for moderate–to–large ME ($\geqslant 2$), while remaining comparable to R–MEM for small ME. The advantage of NPL-HMC widens as ME increases: all competing methods degrade markedly with larger ME scales, while the RMSE of NPL-HMC estimators remains stable. Implementation and




(a) Berkson ME: $(X, Y)$ (red) and $(W, Y)$ (blue). (b) Classical ME: $(X, Y)$ (red) and $(W, Y)$ (blue).

Fig. 2: Illustrative samples for the sigmoid model under ME and 10% Huber contamination.




(a) Berkson ME. (b) Classical ME.

Fig. 3: RMSE comparison for the sigmoid model under misspecification. ME denotes $\sigma_N$

additional set-up details are in Appendix D.

## 5. Real-World Experiments

### *5.1. Berkson ME: California school test scores*

We analyse a Berkson ME setting using the `CASchools` data ($n = 420$ districts, 1998–99) from the `AER` R package [Kleiber and Zeileis, 2008], originally compiled by Stock and Watson [2007]. The response is the average Stanford 9 score $Y_i = \frac{1}{2}(\texttt{math9}_i + \texttt{read9}_i)$ and the regressor of interest is district income. Following Dellaporta and Damoulas [2023], we mimic grouped income by partitioning the empirical support of `avg_income` into 15 equal–width bins and replacing each observation by its within–bin mean $W_i$. This induces a Berkson structure

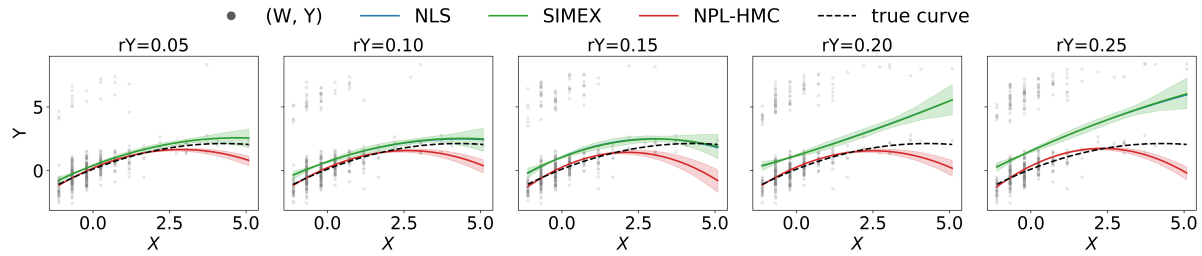



Fig. 4: Estimated curves under five contamination ratios $r_Y$. NLS (blue), SIMEX (green), NPL-HMC (red); 95% bands are shaded, the dashed line is the oracle fit based on latent $X$.

| | $\theta$-RMSE | | | $Y$-RMSE | | |
|---|---|---|---|---|---|---|
| $r_Y$ | NLS | SIMEX | **NPL-HMC** | NLS | SIMEX | **NPL-HMC** |
| 0.05 | 0.091 | 0.104 (0.024) | **0.017 (0.004)** | 0.729 | 0.732 (0.010) | **0.677 (0.004)** |
| 0.10 | 0.391 | 0.413 (0.048) | **0.031 (0.006)** | 0.898 | 0.903 (0.017) | **0.701 (0.008)** |
| 0.15 | 0.990 | 1.003 (0.118) | **0.075 (0.010)** | 1.132 | 1.137 (0.022) | **0.724 (0.022)** |
| 0.20 | 1.314 | 1.358 (0.081) | **0.046 (0.008)** | 1.361 | 1.369 (0.019) | **0.698 (0.008)** |
| 0.25 | 2.279 | 2.312 (0.239) | **0.156 (0.015)** | 1.645 | 1.650 (0.025) | **0.706 (0.004)** |

Table 3: Coefficient RMSE ($\theta$-RMSE) and prediction RMSE ($Y$-RMSE) by contamination ratio $r_Y$. For each $r_Y$, the dataset is fixed across methods; entries report mean (s.d.).

$X_i = W_i + \nu_i$ with $\nu_i \overset{\text{iid}}{\perp\!\!\!\perp} W_i$ representing within–bin variability. After standardising $Y$ and $X$, we fit $Y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. To assess robustness, we construct for each contamination ratio $r_Y \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ a contaminated dataset by shifting a proportion $r_Y$ of responses upward by six standard deviations (SD). We compare our NPL-HMC estimators with two other methods. Nonlinear least squares (NLS) serves as a baseline: under Berkson ME it targets $\mathbb{E}(Y \mid W)$ and does not recover $\mathbb{E}(Y \mid X)$ in general. SIMEX [Cook and Stefanski, 1994] is implemented for Berkson ME by adding synthetic normal noise to $W$ at known multiples of the ME and extrapolating to the error–free limit.

Figure 4 shows fitted curves for the five contamination ratios. NLS and SIMEX rise steeply due to the impact of outliers once the contaminated ratio exceeds 15%. NPL-HMC predicted curves remain closer to the oracle curve (dashed) despite increasing contamination, especially for lower to mid $X$ values, where most of the data are. Table 3 reports RMSEs for (i) the parameter estimate $\hat{\theta}$, compared with an oracle estimator using the true latent $X$; and (ii) predicted $Y$ computed using the true $X$ evaluated on uncontaminated $Y$.

### 5.2. Classical ME: Engel curves

We analyse a log–quadratic Engel curve for food expenditure using the Belgian household data assembled by Engel in the 1850s (235 households; distributed as `statsmodels::engel` in the Python package by Seabold and Perktold [2010]). The outcome is food expenditure $Y_i$ (francs) and the true covariate is household income $X_i$ [Engel, 1857]. The working model is $Y_i = \theta_0 + \theta_1 \log X_i + \theta_2 (\log X_i)^2 + \varepsilon_i,\ \varepsilon_i \perp \log X_i$. Income is observed with error via self–reports $W_i$. Motivated by evidence that misreporting is roughly proportional to income, we adopt a classical error model on the log scale,

$$\log W_i = \log X_i + N_i, \qquad N_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_N^2),$$

with $N_i$ independent of $(X_i, \varepsilon_i)$ [Kedir and Girma, 2003]. We set $\log X_i \sim \mathcal{N}(\mu, \sigma_X^2)$ and index the ME magnitude by the *error–variance ratio* on the log scale, $\rho = \sigma_N^2 / \sigma_X^2$.

In our application, the sample is small and the log–quadratic model is an approximation, so the error ratio cannot be identified precisely. Because Engel-curve elasticities inform economic and policy analysis, it is important that estimates are not overly sensitive to plausible choices of $\rho$. Estimated survey error varies across studies: Bound et al. [2001] and Aasness et al. [1993] place $\rho$ in the range $[0.1, 0.5]$, while Hausman et al. [1995] estimates $\rho$ values up to 0.72 in a generalised setting with total expenditure as $X_i$ and budget shares as $Y_i$. We therefore examine $\rho \in [0, 0.8]$, which covers the empirical ranges observed under different settings.

We first study how fitted curves change with $\rho$. We compare our NPL–HMC estimator with SIMEX. Figure 5 shows fitted curves with pointwise 95% bands for four different $\rho$ values. As $\rho$ increases, the SIMEX curves move substantially while their band size is roughly constant. By contrast, the NPL–HMC curves vary little with $\rho$, and their bands widen as $\rho$ grows. Thus NPL–HMC yields more stable point estimates and a cautious widening of uncertainty as the assumed error variance increases.

To summarise curve variation across $\rho$ we use $\widehat{S} = \left\{ \frac{1}{|\mathcal{R}|} \sum_{\rho \in \mathcal{R}} \sum_{x \in \mathcal{G}} w(x) \big[ g(\theta_\rho, x) - \bar{g}(x) \big]^2 \right\}^{1/2}$,

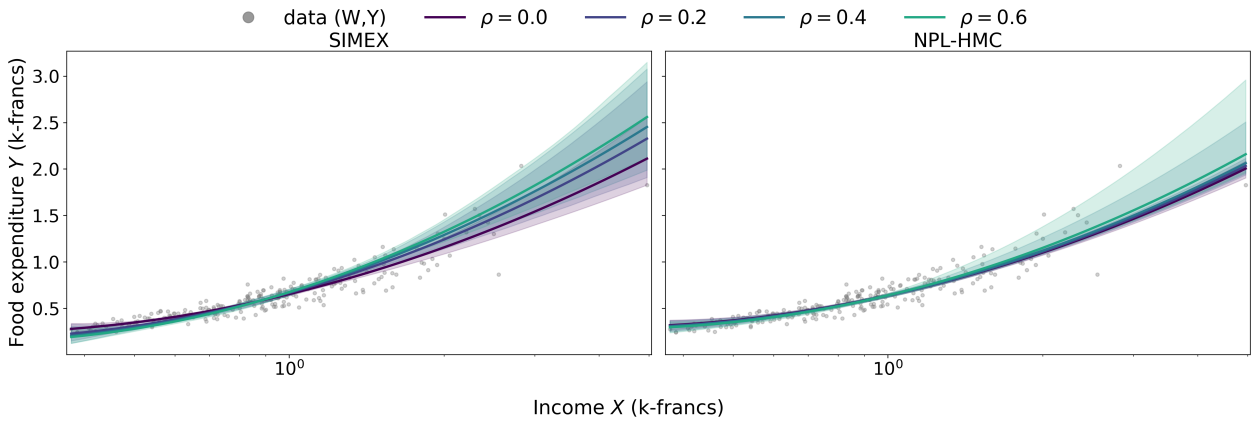



Fig. 5: Fitted Engel curves for increasing values of $\rho$ with pointwise 95% bands for SIMEX (left) and NPL–HMC (right).

the weighted deviation of fitted curves $g(\theta_\rho, x) = \theta_{0,\rho} + \theta_{1,\rho} \log x + \theta_{2,\rho} (\log x)^2$ over a set $\mathcal{R}$ of $\rho$ values on a grid $\mathcal{G}$, with weights $w$ given by the empirical histogram of $W$. We compute $\hat{S}$ with $\mathcal{R} = \{0, 0.1, \ldots, 0.8\}$, which is smaller for NPL–HMC (0.017) than for SIMEX (0.028).

We next assess sampling variability by repeated subsampling. For each $\rho \in \mathcal{R}$ we draw $M = 100$ independent subsamples of size $0.8n$ and re–estimate the curve. Table 4 reports, for each $\rho$, the SD across subsamples of $(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2)$. As a reference, we also report NLS, which ignores ME and is $\rho$-invariant: the SDs (multiplied by 100) are (0.53) for $\hat{\theta}_0$, (1.55) for $\hat{\theta}_1$, and (3.52) for $\hat{\theta}_2$. NPL–HMC has smaller SDs than SIMEX for almost all parameters and all $\rho$, and its SDs are comparable to the NLS baselines. To summarise sensitivity of parameter estimation across different $\rho$ values, Table 5 reports the mean (over subsamples) of the variance of $\hat{\theta}_i$ across $\rho$, defined as $\mathrm{Var}_\rho(\hat{\theta}_i) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Var}_\rho \left( \hat{\theta}_{i,\rho}^{(m)} \right)$, where $\hat{\theta}_{i,\rho}^{(m)}$ is the estimator for $\theta_i$ under ME ratio $\rho$ and for subsample index $m$. The across–$\rho$ variance is smaller for NPL–HMC than for SIMEX for each parameter, with the largest difference in $\hat{\theta}_1$, the *income–elasticity index* used in Engel–curve applications.

## 6. Conclusions

We have proposed a Bayesian NPL framework for nonlinear regression with covariate ME that remains valid under simultaneous misspecification of the regression model, the ME dis-

| | **NPL–HMC** | | | **SIMEX** | | |
|---|---|---|---|---|---|---|
| $\rho$ | $\mathrm{SD}(\hat{\theta}_0)$ | $\mathrm{SD}(\hat{\theta}_1)$ | $\mathrm{SD}(\hat{\theta}_2)$ | $\mathrm{SD}(\hat{\theta}_0)$ | $\mathrm{SD}(\hat{\theta}_1)$ | $\mathrm{SD}(\hat{\theta}_2)$ |
| 0.00 | 0.38 | 1.49 | 3.06 | 0.53 | 1.56 | 3.53 |
| 0.10 | 0.47 | 1.57 | 3.10 | 0.87 | 2.10 | 5.17 |
| 0.20 | 0.45 | 1.56 | 3.19 | 1.05 | 2.24 | 5.73 |
| 0.30 | 0.54 | 1.55 | 3.19 | 1.00 | 2.60 | 6.39 |
| 0.40 | 0.55 | 1.50 | 3.13 | 1.16 | 2.60 | 6.20 |
| 0.50 | 0.53 | 1.91 | 3.29 | 1.18 | 2.75 | 6.63 |
| 0.60 | 0.59 | 2.40 | 3.29 | 1.22 | 2.76 | 6.45 |
| 0.70 | 0.57 | 2.82 | 3.67 | 1.29 | 2.76 | 6.22 |
| 0.80 | 0.66 | 3.92 | 4.94 | 1.25 | 2.79 | 6.33 |
| **Mean** | **0.53** | **2.08** | **3.43** | **1.06** | **2.46** | **5.85** |

Table 4: Within–$\rho$ SDs of subsample estimates ($\times 100$).

| Method | $\mathrm{Var}_\rho(\hat{\theta}_0)$ | $\mathrm{Var}_\rho(\hat{\theta}_1)$ | $\mathrm{Var}_\rho(\hat{\theta}_2)$ |
|---|---|---|---|
| NPL–HMC | 0.20 | 15.46 | 9.11 |
| SIMEX | 0.97 | 47.10 | 15.57 |

Table 5: Across–$\rho$ variance of parameter estimates ($\times 10^4$), reported as the mean (over subsamples) of the per-subsample variance across different $\rho$.

tribution, and the outcome noise. We assign DP priors to the unknown joint distribution of $(X, Y)$ and infer $\theta$ by matching the DP posterior to the model–implied joint distribution via the MMD, while posterior pseudo–sampling transports information from $(W, Y)$ to $X$ in both Berkson and classical ME settings. This yields generalisation bounds that separate statistical fluctuation, prior model discrepancy, and pseudo–sample distribution discrepancy. The latter is controlled by a total misspecification term and an asymptotic remainder. A consistency result for the resulting estimator complements these bounds. Empirically, the framework achieves lower error and more stable estimates under misspecification than comparable Bayesian and frequentist methods in synthetic and real applications.

The pseudo-sampling step is generalisable: any (generalised) posterior for $\theta$ that contracts to a pseudo-true value can be used, and our risk decomposition still applies once the same contraction and stability conditions are verified. A limitation is that, in high-dimensional settings, the HMC step can mix slowly and be computationally expensive. Alternatives such as preconditioned/tempered or stochastic-gradient MCMC can be desirable, and variational approximations (e.g., mean-field or $\alpha$-variational inference [Blei et al., 2017, Yang et al.,

2020]) may be used to produce pseudo-samples at the price of approximation bias. More generally, our framework is compatible with modern ML components (e.g., Bayesian neural networks) as modelling and inference modules. Exploring these integrations is a promising direction for future work.

Several important theoretical questions remain open. First, the distributional properties of $\hat{\theta}_n$ beyond consistency are unknown. In particular, it is open whether it satisfies asymptotic normality or a Bernstein–von Mises–type limit. Settling these would enable interval estimation and hypothesis testing. Second, fully nonparametric modelling of $g$ via Gaussian process priors under ME is attractive, but current approaches typically assume a known ME law and lack guarantees under misspecification. In parallel, it would be useful to develop adversarial robustness guarantees and (near-)minimax rates under $\varepsilon$-contamination and ME, with rates that degrade explicitly with the level of total model misspecification.

## 7. Acknowledgements

MC is supported by the Warwick Statistics Centre for Doctoral Training and acknowledges funding from the University of Warwick. CD was supported from EPSRC grant [EP/Y022300/1]. TBB was supported by European Research Council Starting Grant 101163546. TD acknowledges support from a UKRI Turing AI acceleration Fellowship [EP/V02678X/1].

## References

J. Aasness, E. Biørn, and T. Skjerpen. Engel functions, panel data, and latent variables. *Econometrica*, pages 1395–1422, 1993.

P. Alquier and M. Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1): 71–92, 2024.

J. Berkson. Are there two regressions? *J. Am. Statist. Ass.*, 45(250):164–180, 1950.

S. M. Berry, R. J. Carroll, and D. Ruppert. Bayesian smoothing and regression splines for measurement error problems. *J. Am. Statist. Ass.*, 97(457):160–169, 2002.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *J. R. Statist. Soc. B*, 78(5):1103–1130, 2016.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *J. Am. Statist. Ass.*, 112(518):859–877, 2017.

J. Bound, C. Brown, and N. Mathiowetz. Measurement error in survey data. In *Handbook of Econometrics*,

volume 5, pages 3705–3843. Elsevier, 2001.

T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, and M. van Smeden. Measurement error is often neglected in medical literature: a systematic review. *Journal of clinical epidemiology*, 98: 89–97, 2018.

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

J. P. Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

C. R. B. Cabral, V. H. Lachos, and C. B. Zeller. Multivariate measurement error models using finite mixtures of skew-student t distributions. *J. Multiv. Anal.*, 124:179–198, 2014.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

B.-E. Chérief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR, 2020.

B.-E. Chérief-Abdellatif and J. Näf. Parametric MMD estimation with missing values: Robustness to missingness and data model misspecification. *arXiv preprint arXiv:2503.00448*, 2025.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Ass.*, 89(428):1314–1328, 1994.

B. Curley. A nonlinear measurement error model and its application to describing the dependency of health outcomes on dietary intake. *J. Appl. Statist*, 49(6):1485–1518, 2022. doi: 10.1080/02664763.2020.1870671. URL `https://doi.org/10.1080/02664763.2020.1870671`.

A. Delaigle and P. Hall. Methodology for non-parametric deconvolution when the error distribution is unknown. *J. R. Statist. Soc. B*, 78(1):231–252, 2016.

A. Delaigle, P. Hall, and P. Qiu. Nonparametric methods for solving the Berkson errors-in-variables problem. *J. R. Statist. Soc. B*, 68(2):201–220, 2006.

C. Dellaporta and T. Damoulas. Robust Bayesian inference for measurement error models. *arXiv preprint arXiv:2306.01468*, 2023.

C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR, 2022.

W. Deming. *Statistical Adjustment of Data*. Dover books on elementary and intermediate mathematics. J. Wiley & Sons, Incorporated, 1943. URL `https://books.google.co.uk/books?id=9awgAAAAMAAJ`.

E. Engel. Die produktions-und consumtionsverhaltnisse des königreichs sachsen, reprinted with engel (1895), anlage i, 1-54. *TER*, 4(2):216–225, 1857.

F. Fazeli-Asl, M. M. Zhang, and L. Lin. A semi-bayesian nonparametric estimator of the maximum mean discrepancy measure: Applications in goodness-of-fit testing and generative adversarial networks. *arXiv preprint arXiv:2303.02637*, 2023.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, pages 209–230, 1973.

E. Fong, S. Lyddon, and C. Holmes. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning*, pages 1952–1962. PMLR, 2019.

M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760, 1996.

A. Ghassami, A. Ying, I. Shpitser, and E. T. Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *International conference on artificial*

*intelligence and statistics*, pages 7210–7239. PMLR, 2022.

R. B. Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, 2012.

P. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 2017.

P. Gustafson. On the simultaneous effects of model misspecification and errors in variables. *Canadian Journal of Statistics*, 30(3):463–474, 2002.

P. Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman and Hall/CRC, 2003.

G. Haber, J. Sampson, and B. Graubard. Bias due to Berkson error: issues when using predicted values in place of observed covariates. *Biostatistics*, 22(4):858–872, 2021.

F. R. Hampel. The influence curve and its role in robust estimation. *J. Am. Statist. Ass.*, 69(346):383–393, 1974.

J. A. Hausman, W. K. Newey, and J. L. Powell. Nonlinear errors in variables estimation of some engel curves. *Journal of Econometrics*, 65(1):205–233, 1995.

Z. Hu, Z. T. Ke, and J. S. Liu. Measurement error models: from nonparametric methods to deep neural networks, 2020. URL `https://arxiv.org/abs/2007.07498`.

X. Huang. Dual model misspecification in generalized linear models with error in variables. In *New Developments in Statistical Modeling, Inference and Application: Selected Papers from the 2014 ICSA/KISS Joint Applied Statistics Symposium in Portland, OR*, pages 3–35. Springer, 2016.

Y. Huang. Corrected score with sizable covariate measurement error: pathology and remedy. *Statistica Sinica*, 24(1):357, 2014.

P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

A. M. Kedir and S. Girma. Quadratic food engel curves with measurement error: Evidence from a budget survey. *Oxford Bulletin of Economics and Statistics*, 2003.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. doi: 10.1007/978-0-387-77318-6. URL `https://CRAN.R-project.org/package=AER`.

B. J. Kleijn and A. W. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electron. J. Statist*, 2012.

M. Klimstra and E. P. Zehr. A sigmoid function is the best fit for the ascending limb of the hoffmann reflex recruitment curve. *Experimental brain research*, 186(1):93–105, 2008.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *J. Mach. Learn. Res.*, 23(132):1–109, 2022.

Y. Lee, T. Jeong, and H. Kim. A Bayesian nonparametric mixture measurement error model with application to spatial density estimation using mobile positioning data with multi-accuracy and multi-coverage. *Technometrics*, 62(2):173–183, 2020.

S. Lyddon, S. Walker, and C. C. Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in neural information processing systems*, 31, 2018.

T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Generalized Bayesian inference for discrete intractable likelihood. *J. Am. Statist. Ass.*, 119(547):2345–2355, 2024.

J. McIntyre and L. A. Stefanski. Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics*, 63(1):81–99, 2011.

M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez. On the robustness to misspecification of $\alpha$-posteriors and their variational approximations. *J. Mach. Learn. Res.*, 23(147):1–51, 2022.

P. Müller and K. Roeder. A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, 84(3):523–537, 1997.

T. Nakamura. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1):127–137, 1990.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

S. Roy and T. Banerjee. A flexible model for generalized linear regression with measurement error. *Annals of the Institute of Statistical Mathematics*, 58:153–169, 2006.

D. Rubin. Multiple imputation for nonresponse in surveys. New York, ny, 1987.

A. Sarkar, B. K. Mallick, and R. J. Carroll. Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors. *Biometrics*, 70(4):823–834, 2014.

S. M. Schennach. Regressions with Berkson errors in covariates—a nonparametric approach. *Ann. Statist.*, pages 1642–1668, 2013.

S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

J. Sethuraman and M. Ghosh. Some properties of GEM($\alpha$) distributions. *Sankhya A*, 86(Suppl 1):288–300, 2024.

L. A. Stefanski. The effects of measurement error on parameter estimation. *Biometrika*, 72(3):583–592, 1985.

J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Addison Wesley, Boston, 2007.

L. Wang. Estimation of nonlinear models with Berkson measurement errors. *Ann. Statist.*, 32(6):2559–2579, 2004.

Y. Yang, D. Pati, and A. Bhattacharya. $\alpha$-variational inference with statistical guarantees. *Ann. Statist.*, 48 (2):886–905, 2020.

G. Y. Yi and Y. Yan. Estimation and hypothesis testing with error-contaminated survival data under possibly misspecified measurement error models. *Canadian Journal of Statistics*, 49(3):853–874, 2021.

X. Yin, J. Goudriaan, E. A. Lantinga, J. Vos, and H. J. Spiertz. A flexible sigmoid function of determinate growth. *Annals of botany*, 91(3):361–371, 2003.

R. H. Zamar. Robust estimation in the errors-in-variables model. *Biometrika*, 76(1):149–160, 1989.

Y. Zhang, G. Qin, Z. Zhu, and J. Zhang. Robust estimation in linear regression models for longitudinal data with covariate measurement errors and outliers. *J. Multiv. Anal.*, 168:261–275, 2018.

S. Zhou, D. Pati, T. Wang, Y. Yang, and R. J. Carroll. Gaussian processes with errors in variables: Theory and computation. *J. Mach. Learn. Res.*, 24(87):1–53, 2023.

## Supplementary Material

Section A presents the DP posterior bootstrapping algorithm. Section B contains proofs for results in the main text. Section C lists sufficient conditions for assumptions A1-A2 and gives example scenarios where they hold. Section D provides implementation and additional set-up details for synthetic and real-world experiments. Section E includes diagnostics and sensitivity analysis for the HMC-based pseudo-sampling procedure. Furthermore, a detailed discussion of dropping the conditional independence requirement in the pseudo-sampling procedure (Section 2.2.3 in the main text) is included at the end.

## A. Algorithm

Below is our DP posterior bootstrapping algorithm as described in Section 2.2.5. We use the truncated stick-breaking procedure [Sethuraman, 1994] to approximate samples from the DP posterior:

$$\mathbb{P}^{\mathrm{DP}}_{X,Y|w_i} \sim \mathrm{DP}\Big(c+m, \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{m}{c+m}\sum_{j=1}^{m}\delta_{(\tilde{x}_{i,j},y_i)}\Big),$$

$$\text{take} \qquad \{(x^{(\mathrm{prior})}_{i,k}, y^{(\mathrm{prior})}_{i,k})\}_{k=1}^{T_{\mathrm{DP}}} \overset{\mathrm{iid}}{\sim} \mathbb{Q}_{XY,i}, \qquad \xi^i_{1:(T_{\mathrm{DP}}+m)} \sim \mathrm{Dir}\Big(\underbrace{\frac{c}{T_{\mathrm{DP}}},\ldots,\frac{c}{T_{\mathrm{DP}}}}_{T_{\mathrm{DP}}\text{ terms}}, \underbrace{1,\ldots,1}_{m\text{ terms}}\Big),$$

$$\text{then} \qquad \mathbb{P}^{\mathrm{DP}}_{X,Y|w_i} \approx \sum_{k=1}^{T_{\mathrm{DP}}} \xi^i_k \delta_{\left(x^{(\mathrm{prior})}_{i,k}, y^{(\mathrm{prior})}_{i,k}\right)} + \sum_{j=1}^{m} \xi^i_{T_{\mathrm{DP}}+j}\delta_{(\tilde{x}_{ij},y_i)}.$$

Here $T_{\mathrm{DP}}$ is a finite truncation limit: in all experiments, we fix $T_{\mathrm{DP}} = 100$.

## B. Proofs

### *B.1. Proof of Theorem 1*

Before proving Theorem 1, we first state and prove the following lemma

**Lemma 2** (Joint DP MMD Bound with concentration parameter $c+m$)**.** *Let $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ be a joint input-output space, and let $k : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ be a positive-definite kernel with associated Hilbert space $\mathcal{H}_k$. Suppose $\mathbb{P}^0_{XY}$ is a target ("true") joint distribution on $\mathcal{Z}$. For $i = 1,\ldots,n$, let $\mathbb{P}^i = \mathbb{P}^{\mathrm{DP}}_{X,Y\,|\,w_i}$ be drawn from a Dirichlet Process*

$$\mathrm{DP}(c+m, \mathbb{P}^{\mathrm{base}}_{XY,i}),$$

*where each base measure is*

$$\mathbb{P}^{\mathrm{base}}_{XY,i} = \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{1}{c+m}\sum_{k=1}^{m}\delta_{\left(\tilde{X}_{ij},Y_i\right)},$$

**Algorithm 1** DP Posterior Bootstrapping with Truncation and MMD Minimisation

**Input:** Pseudo-samples $\{\tilde{x}_{i,j}\}_{i=1,\dots,n;j=1,\dots,m}$, Observed $\{(w_i, y_i)\}_{i=1}^n$, DP concentration $c \geqslant 0$, base measures $\{\mathbb{Q}_{XY,i}\}$, DP truncation limit $T_{\mathrm{DP}}$, number of bootstrap loops $B_{\mathrm{boot}}$.

1: **for** $b = 1$ to $B_{\mathrm{boot}}$ **do**

2: **for** $i = 1$ to $n$ **do**

3: Draw $T_{\mathrm{DP}}$ atoms from the prior $\{(x_{i,k}^{(\mathrm{prior})}, y_{i,k}^{(\mathrm{prior})})\}_{k=1}^{T_{\mathrm{DP}}} \overset{\mathrm{iid}}{\sim} \mathbb{Q}_{XY,i}$.

4: Form combined set $\mathcal{S}_i^{(b)} = \left\{\left(x_{i,k}^{(\mathrm{prior})}, y_{i,k}^{(\mathrm{prior})}\right)\right\}_{k=1}^{T_{\mathrm{DP}}} \cup \left\{(\tilde{x}_{i,j}, y_i)\right\}_{j=1}^{m}$.

5: Sample weights $\xi_{1:(T_{\mathrm{DP}}+m)}^{(i,b)} \sim \mathrm{Dir}\left(\frac{c}{T_{\mathrm{DP}}}, \dots, \frac{c}{T_{\mathrm{DP}}}, 1, \dots, 1\right)$, construct DP measure:

$$\mathbb{P}_{X,Y|w_i}^{(b)} := \sum_{k=1}^{T_{\mathrm{DP}}} \xi_k^{(i,b)} \delta_{\left(x_{i,k}^{(\mathrm{prior})}, y_{i,k}^{(\mathrm{prior})}\right)} + \sum_{j=1}^{m} \xi_{T_{\mathrm{DP}}+j}^{(i,b)} \delta_{\left(\tilde{x}_{i,j}, y_i\right)}.$$

6: Similarly, construct marginal DP on $X|w_i$:

$$\mathbb{P}_{X|w_i}^{(b)} := \sum_{k=1}^{T_{\mathrm{DP}}} \xi_k^{(i,b)} \delta_{x_{i,k}^{(\mathrm{prior})}} + \sum_{j=1}^{m} \xi_{T_{\mathrm{DP}}+j}^{(i,b)} \delta_{\tilde{x}_{i,j}}.$$

7: **end for**

8: Form $\mathcal{P}_{X,Y}^{\mathrm{DP},(b)} = \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{X,Y|w_i}^{(b)}$.

9: Calculate parametric counterpart $\widetilde{\mathcal{P}}_{X,Y}^{(\theta,b)} = \left(\frac{1}{n}\sum_{i=1}^n \mathbb{P}_{X|w_i}^{(b)}\right) \times \mathbb{P}_{g(\cdot,\theta)}$.

10: **Compute** $\hat{\theta}_b = \arg\min_{\theta\in\Theta}\left(\mathcal{P}_{X,Y}^{\mathrm{DP},(b)}, \widetilde{\mathcal{P}}_{X,Y}^{(\theta,b)}\right)$ via gradient descent.

11: **end for**

**Output:** Posterior bootstrap draws $\{\hat{\theta}_b\}_{b=1}^{B_{\mathrm{boot}}}$.

*and define the aggregated measure*

$$\mathcal{P}_{X,Y}^{\mathrm{DP}} = \frac{1}{n}\sum_{i=1}^n \mathbb{P}^i, \qquad \mathbb{P}_{XY}^{\mathrm{base}} = \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{XY,i}^{\mathrm{base}}, \qquad \mathbb{P}_{XY}^{\mathrm{prior}} = \frac{1}{n}\sum_{i=1}^n \mathbb{Q}_{XY,i}.$$

*Also define the pseudo-sample empirical distribution*

$$\mathbb{P}_{XY}^{\mathrm{pseudo}} = \frac{1}{n}\sum_{i=1}^n \Big[\frac{1}{m}\sum_{k=1}^m \delta_{\left(\tilde{X}_{ij}, Y_i\right)}\Big] = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{\left(\tilde{X}_{ij}, Y_i\right)}.$$

*Let $\mu(\mathcal{D}, S)$ be the joint distribution of $\mathbb{P} = (\mathbb{P}^1, \dots, \mathbb{P}^n)$ given the observed data $\mathcal{D} := \{(W_i, Y_i)\}_{i=1}^n$ and pseudo–samples $S := \{\tilde{X}_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$. Then the following statement holds*

$$\mathbb{E}_{\mathcal{D},S}\,\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k(\mathbb{P}_{XY}^0, \mathcal{P}_{X,Y}^{\mathrm{DP}}) \mid \mathcal{D}, S\right] \leqslant \frac{\sqrt{\kappa}}{\sqrt{n(c+m+1)}} + \frac{c}{c+m}\,\mathbb{E}_{\mathcal{D},S}\,\mathrm{MMD}_k(\mathbb{P}_{XY}^0, \mathbb{P}_{XY}^{\mathrm{prior}}) + \frac{m}{c+m}\,\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_k(\mathbb{P}_{XY}^0, \mathbb{P}_{XY}^{\mathrm{pseudo}})\right].$$

*Proof of Lemma 2.* Throughout this proof, we condition on $(\mathcal{D}, S)$ and treat them as fixed, until the final step where we take the expectation over all $(\mathcal{D}, S)$. Each $\mathbb{P}^i$ can be written

via Sethuraman's stick-breaking representation with concentration parameter $\alpha = c + m$ [Sethuraman, 1994]:

$$\mathbb{P}^i = \sum_{j=1}^{\infty} \xi_j^i \delta_{z_j^i},$$

where $\{\xi_j^i\} \sim \mathrm{GEM}(c+m)$, $\{z_j^i\}$ are i.i.d. draws from the base measure $\mathbb{P}_{XY,i}^{\mathrm{base}}$. Hence

$$\mathcal{P}_{X,Y}^{\mathrm{DP}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\infty} \xi_j^i \delta_{z_j^i}.$$

We use the triangle inequality:

$$\mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \mathcal{P}_{X,Y}^{\mathrm{DP}}\right) \leqslant \mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \mathbb{P}_{XY}^{\mathrm{base}}\right) + \mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{base}}, \mathcal{P}_{X,Y}^{\mathrm{DP}}\right),$$

where

$$\mathbb{P}_{XY}^{\mathrm{base}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{XY,i}^{\mathrm{base}} = \frac{c}{c+m} \mathbb{P}_{XY}^{\mathrm{prior}} + \frac{m}{c+m} \mathbb{P}_{X,Y}^{\mathrm{pseudo}}.$$

We will show the following two bounds:

(a) $\mathbb{E}_{\mathrm{DP}}\big[\mathrm{MMD}_k(\mathcal{P}_{X,Y}^{\mathrm{DP}}, \mathbb{P}_{XY}^{\mathrm{base}}) \mid \mathcal{D}, S\big] \leqslant \dfrac{1}{\sqrt{n(c+m+1)}}.$

(b) $\mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \mathbb{P}_{XY}^{\mathrm{base}}\right) \leqslant \dfrac{c}{c+m} \mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \mathbb{P}_{XY}^{\mathrm{prior}}\right) + \dfrac{m}{c+m} \mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \mathbb{P}_{X,Y}^{\mathrm{pseudo}}\right).$

*Bound on* $\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k(\mathcal{P}_{X,Y}^{\mathrm{DP}}, \mathbb{P}_{XY}^{\mathrm{base}}) \mid \mathcal{D}, S\right]$*:* We proceed in two sub-steps: (A) compute the squared MMD, (B) apply Jensen's inequality.

Rewrite

$$\mathrm{MMD}_k^2(\mathcal{P}_{X,Y}^{\mathrm{DP}}, \mathbb{P}_{XY}^{\mathrm{base}}) = \left\|\varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) - \varphi(\mathbb{P}_{XY}^{\mathrm{base}})\right\|_{\mathcal{H}_k}^2 = \left\langle \varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) - \varphi(\mathbb{P}_{XY}^{\mathrm{base}}), \varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) - \varphi(\mathbb{P}_{XY}^{\mathrm{base}}) \right\rangle_{\mathcal{H}_k}.$$

Express the embeddings:

$$\varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) = \int k(z,\cdot) d\mathcal{P}_{X,Y}^{\mathrm{DP}}(z) = \frac{1}{n} \sum_{i=1}^{n} \int k(z,\cdot) d\,\mathbb{P}^i(z),$$

$$\varphi(\mathbb{P}_{XY}^{\mathrm{base}}) = \int k(z,\cdot) d\,\mathbb{P}_{XY}^{\mathrm{base}}(z) = \frac{1}{n} \sum_{i=1}^{n} \int k(z,\cdot) d\,\mathbb{P}_{XY,i}^{\mathrm{base}}(z).$$

Denote $\varphi(\mathbb{P}^i) = \int k(z,\cdot) d\,\mathbb{P}^i(z), \varphi(\mathbb{P}_{XY,i}^{\mathrm{base}}) = \int k(z,\cdot) d\,\mathbb{P}_{XY,i}^{\mathrm{base}}(z)$. Then

$$\varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\mathbb{P}^i), \quad \varphi(\mathbb{P}_{XY}^{\mathrm{base}}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\mathbb{P}_{XY,i}^{\mathrm{base}}).$$

Hence

$$\varphi(\mathcal{P}_{X,Y}^{\mathrm{DP}}) - \varphi(\mathbb{P}_{XY}^{\mathrm{base}}) = \frac{1}{n} \sum_{i=1}^{n} \left(\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})\right).$$

Therefore,

$$\left\|\varphi(\mathcal{P}^{\mathrm{DP}}_{X,Y}) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY})\right\|^2_{\mathcal{H}_k} = \left\langle \frac{1}{n}\sum_{i=1}^{n}(\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})), \frac{1}{n}\sum_{\ell=1}^{n}(\varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})) \right\rangle_{\mathcal{H}_k}$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\left\langle \varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}), \varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})\right\rangle_{\mathcal{H}_k}.$$

We want its expectation w.r.t. $\mathbb{P} = (\mathbb{P}^1, \ldots, \mathbb{P}^n)$ under $\mathbb{E}_{\mathrm{DP}}[\cdot \mid \mathcal{D}, S]$. For brevity, we omit the conditioning on $\mathcal{D}, S$ for the rest of the proof and write $\mathbb{E}_{\mathrm{DP}}[\cdot] := \mathbb{E}_{\mathrm{DP}}[\cdot \mid \mathcal{D}, S]$. We have

$$\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}^2_k(\mathcal{P}^{\mathrm{DP}}_{X,Y}, \mathbb{P}^{\mathrm{base}}_{XY})\right] = \frac{1}{n^2}\sum_{i,\ell=1}^{n}\mathbb{E}_{\mathrm{DP}}\left[\left\langle \varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}), \varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})\right\rangle_{\mathcal{H}_k}\right]$$

Next, note that $\mathbb{P}^i$ and $\mathbb{P}^\ell$ are drawn independently from the Dirichlet processes $\mathrm{DP}(c + m, \mathbb{P}^{\mathrm{base}}_{XY,i})$ and $\mathrm{DP}(c + m, \mathbb{P}^{\mathrm{base}}_{XY,\ell})$, so:

(a) If $i \neq \ell$,

$$\mathbb{E}_{\mathrm{DP}}\left[\langle \varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}), \varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})\rangle\right] = \mathbb{E}_{\mathrm{DP}}\left[\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right] \cdot \mathbb{E}_{\mathrm{DP}}\left[\varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})\right],$$

since $\mathbb{P}^i$ is independent from $\mathbb{P}^\ell$ given $(\mathcal{D}, S)$. But $\mathbb{E}[\varphi(\mathbb{P}^i)] = \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})$ by definition of the DP's mean. Indeed, for any function $f$, $\mathbb{E}_{\mathbb{P}^i \sim \mu_i}\left[\int f(z)\mathrm{d}\,\mathbb{P}^i(z)\right] = \int f(z)\mathrm{d}\,\mathbb{P}^{\mathrm{base}}_{XY,i}(z)$. Hence

$$\mathbb{E}_{\mathrm{DP}}[\varphi(\mathbb{P}^i)] = \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}).$$

So

$$\mathbb{E}_{\mathrm{DP}}\left[\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right] = \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}) = 0.$$

Thus any cross-term with $i \neq \ell$ has expectation zero:

$$\mathbb{E}_{\mathrm{DP}}\left[\langle \varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}), \varphi(\mathbb{P}^\ell) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,\ell})\rangle\right] = 0.$$

(b) If $i = \ell$ we consider $\mathbb{E}\left[\|\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\|^2_{\mathcal{H}_k}\right]$.

By Sethuraman's construction, we write $\mathbb{P}^i = \sum_{j=1}^{\infty}\xi^i_j\delta_{z^i_j}$, where $\{\xi^i_j\}_{j=1}^{\infty} \sim \mathrm{GEM}(\alpha)$, and $z^i_j \overset{iid}{\sim} \mathbb{P}^{\mathrm{base}}_{XY,i}$. Define the kernel mean embeddings

$$\varphi(\mathbb{P}^i) = \int k(z, \cdot) d\,\mathbb{P}^i(z) = \sum_{j=1}^{\infty}\xi^i_j k\left(z^i_j, \cdot\right),$$

and

$$\varphi(\mathbb{P}^{\mathrm{base}}_{XY,i}) = \int k(z, \cdot) d\,\mathbb{P}^{\mathrm{base}}_{XY,i}(z).$$

We aim to bound

$$\mathbb{E}_{\mathrm{DP}}\left[\left\|\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right\|^2_{\mathcal{H}_k}\right].$$

First write

$$\left\|\varphi(\mathbb{P}^i) - \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right\|^2_{\mathcal{H}_k} = \left\|\varphi(\mathbb{P}^i)\right\|^2_{\mathcal{H}_k} - 2\left\langle \varphi(\mathbb{P}^i), \varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right\rangle_{\mathcal{H}_k} + \left\|\varphi(\mathbb{P}^{\mathrm{base}}_{XY,i})\right\|^2_{\mathcal{H}_k}.$$

Taking expectations over the random weights $\{\xi_j^i\}$ and atoms $\{z_j^i\}$ gives three terms:

$$\mathbb{E}_{\mathrm{DP}}\Big[\big\|\varphi(\mathbb{P}^i)\big\|_{\mathcal{H}_k}^2\Big], \quad \mathbb{E}_{\mathrm{DP}}\Big[\big\langle\varphi(\mathbb{P}^i), \varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})\big\rangle_{\mathcal{H}_k}\Big], \quad \mathbb{E}_{\mathrm{DP}}\Big[\big\|\varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})\big\|_{\mathcal{H}_k}^2\Big].$$

We denote these three pieces as (I), (II), and (III), respectively.

(a) (I): $\mathbb{E}_{\mathrm{DP}}[\|\varphi(\mathbb{P}^i)\|^2]$ Recall $\varphi(\mathbb{P}^i) = \sum_{j=1}^{\infty} \xi_j^i k(z_j^i, \cdot)$. Then:

$$\|\varphi(\mathbb{P}^i)\|_{\mathcal{H}_k}^2 = \Big\langle \sum_{j=1}^{\infty} \xi_j^i k(z_j^i, \cdot), \sum_{t=1}^{\infty} \xi_t^i k(z_t^i, \cdot) \Big\rangle_{\mathcal{H}_k} = \sum_{j=1}^{\infty}\sum_{t=1}^{\infty} \xi_j^i \xi_t^i \langle k(z_j^i, \cdot), k(z_t^i, \cdot)\rangle_{\mathcal{H}_k}.$$

By reproducing-kernel property, $\langle k(a, \cdot), k(b, \cdot)\rangle = k(a, b)$. Hence

$$\|\varphi(\mathbb{P}^i)\|_{\mathcal{H}_k}^2 = \sum_{j,t} \xi_j^i \xi_t^i k\left(z_j^i, z_t^i\right).$$

Taking expectation:

$$\text{(I)} = \mathbb{E}_{\mathrm{DP}}\Big[\|\varphi(\mathbb{P}^i)\|^2\Big] = \sum_{j,t} \mathbb{E}\big[\xi_j^i \xi_t^i\big]\, \mathbb{E}\big[k(z_j^i, z_t^i)\big],$$

where we used independence of $\xi_j^i$ from $z_j^i$, plus the i.i.d. property across all $j$. Define

$$\Gamma(\mathbb{P}_{XY,i}^{\mathrm{base}}) := \mathbb{E}_{z,z' \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}\big[k(z, z')\big], \quad \Psi(\mathbb{P}_{XY,i}^{\mathrm{base}}) := \mathbb{E}_{z \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}[k(z, z)]$$

Then

$$\begin{aligned} 2\Psi(\mathbb{P}_{XY,i}^{\mathrm{base}}) - 2\Gamma(\mathbb{P}_{XY,i}^{\mathrm{base}}) &= \mathbb{E}_{z \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}[k(z, z)] + \mathbb{E}_{z' \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}[k(z, z)] - 2\,\mathbb{E}_{z,z' \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}\big[k(z, z')\big] \\ &= \mathbb{E}_{z,z' \sim \mathbb{P}_{XY,i}^{\mathrm{base}}}\big[k(z, z) + k(z', z') - 2k(z, z')\big] \\ &= \|\phi(z) - \phi(z')\|_{\mathcal{H}_k}^2 \\ &\geqslant 0, \end{aligned}$$

where $\phi$ is the feature map. This means $\Psi(\mathbb{P}_{XY,i}^{\mathrm{base}}) \geqslant \Gamma(\mathbb{P}_{XY,i}^{\mathrm{base}})$. We separate diagonal ($j = t$) from off-diagonal:

$$\text{(I)} = \sum_{j=1}^{\infty} \mathbb{E}[(\xi_j^i)^2]\, \mathbb{E}[k(z_j^i, z_j^i)] + \sum_{j \neq t} \mathbb{E}[\xi_j^i \xi_t^i]\, \mathbb{E}[k(z_j^i, z_t^i)].$$

Hence

$$\text{(I)} = \sum_{j=1}^{\infty} \mathbb{E}[(\xi_j^i)^2] \Psi(\mathbb{P}_{XY,i}^{\mathrm{base}}) + \sum_{j \neq t} \mathbb{E}[\xi_j^i \xi_t^i] \Gamma(\mathbb{P}_{XY,i}^{\mathrm{base}}).$$

(b) Term (III): $\mathbb{E}_{\mathrm{DP}}[\|\varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})\|^2]$ Here,

$$\varphi(\mathbb{P}_{XY,i}^{\mathrm{base}}) = \int k(z, \cdot)\, d\,\mathbb{P}_{XY,i}^{\mathrm{base}}(z), \quad \|\varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})\|^2 = \iint k(z, z')\, d\,\mathbb{P}_{XY,i}^{\mathrm{base}}(z)\, d\,\mathbb{P}_{XY,i}^{\mathrm{base}}(z').$$

Since $\varphi(\mathbb{P}_{XY,i}^{\mathrm{base}})$ is fixed given $(\mathcal{D}, S)$ $\mathbb{P}^i$, we have

$$\text{(III)} = \mathbb{E}_{\text{DP}}\big[\|\varphi(\mathbb{P}^{\text{base}}_{XY,i})\|^2\big] = \iint k(z,z')d\,\mathbb{P}^{\text{base}}_{XY,i}(z)d\,\mathbb{P}^{\text{base}}_{XY,i}(z').$$

Define

$$\Gamma(\mathbb{P}^{\text{base}}_{XY,i}) = \mathbb{E}_{z,z'\sim\mathbb{P}^{\text{base}}_{XY,i}}[k(z,z')],$$

thus

$$\text{(III)} = \Gamma(\mathbb{P}^{\text{base}}_{XY,i}).$$

(c) Term (II): $\mathbb{E}_{\text{DP}}[-2\langle\varphi(\mathbb{P}^i),\varphi(\mathbb{P}^{\text{base}}_{XY,i})\rangle]$ We have

$$\langle\varphi(\mathbb{P}^i),\varphi(\mathbb{P}^{\text{base}}_{XY,i})\rangle = \Big\langle\sum_{j=1}^{\infty}\xi^i_j k(z^i_j,\cdot),\int k(z,\cdot)d\,\mathbb{P}^{\text{base}}_{XY,i}(z)\Big\rangle_{\mathcal{H}_k} = \sum_{j=1}^{\infty}\xi^i_j\int\big\langle k(z^i_j,\cdot),k(z,\cdot)\big\rangle_{\mathcal{H}_k}d\,\mathbb{P}^{\text{base}}_{XY,i}(z).$$

Again, $\langle k(a,\cdot),k(b,\cdot)\rangle = k(a,b)$. So

$$\langle\varphi(\mathbb{P}^i),\varphi(\mathbb{P}^{\text{base}}_{XY,i})\rangle = \sum_{j=1}^{\infty}\xi^i_j\int k(z^i_j,z)d\,\mathbb{P}^{\text{base}}_{XY,i}(z).$$

Taking expectation,

$$\text{(II)} = -2\,\mathbb{E}_{\text{DP}}\big[\langle\varphi(\mathbb{P}^i),\varphi(\mathbb{P}^{\text{base}}_{XY,i})\rangle\big] = -2\sum_{j=1}^{\infty}\mathbb{E}[\xi^i_j]\,\mathbb{E}\Big[\int k(z^i_j,z)d\,\mathbb{P}^{\text{base}}_{XY,i}(z)\Big].$$

But

$$\mathbb{E}\left[\int k(z^i_j,z)d\,\mathbb{P}^{\text{base}}_{XY,i}(z)\right] = \mathbb{E}\left[\mathbb{E}_{z\sim\mathbb{P}^{\text{base}}_{XY,i}}[k(z^i_j,z)]\right] = \mathbb{E}\left[\Gamma(\mathbb{P}^{\text{base}}_{XY,i})\right] = \Gamma(\mathbb{P}^{\text{base}}_{XY,i}),$$

Hence

$$\text{(II)} = -2\Gamma(\mathbb{P}^{\text{base}}_{XY,i})\sum_{j=1}^{\infty}\mathbb{E}[\xi^i_j].$$

Now for a GEM($\alpha$) distribution, we know that $\sum_{j=1}^{\infty}\mathbb{E}[\xi^i_j] = 1$ by [Sethuraman and Ghosh, 2024, Lemma 2(b)]. So:

$$\text{(II)} = -2\Gamma(\mathbb{P}^{\text{base}}_{XY,i})\cdot 1 = -2\Gamma(\mathbb{P}^{\text{base}}_{XY,i}).$$

Combining (I), (II), (III), we get

$$\mathbb{E}_{\text{DP}}\big[\|\varphi(\mathbb{P}^i)-\varphi(\mathbb{P}^{\text{base}}_{XY,i})\|^2\big] = \sum_{j=1}^{\infty}\mathbb{E}[(\xi^i_j)^2]\Psi(\mathbb{P}^{\text{base}}_{XY,i}) + \sum_{j\neq t}\mathbb{E}[\xi^i_j\xi^i_t]\Gamma(\mathbb{P}^{\text{base}}_{XY,i}) - 2\Gamma(\mathbb{P}^{\text{base}}_{XY,i}) + \Gamma(\mathbb{P}^{\text{base}}_{XY,i}). \tag{19}$$

$$= \sum_{j=1}^{\infty}\mathbb{E}[(\xi^i_j)^2]\Psi(\mathbb{P}^{\text{base}}_{XY,i}) + \Gamma(\mathbb{P}^{\text{base}}_{XY,i})\sum_{j\neq t}\mathbb{E}[\xi^i_j\xi^i_t] - \Gamma(\mathbb{P}^{\text{base}}_{XY,i}). \tag{20}$$

Now we recall this identity from a GEM($c+m$) distribution [Sethuraman and Ghosh, 2024, Lemma 3]:

$$\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]+\sum_{j\neq t}\mathbb{E}[\xi_j^i\xi_t^i]=\frac{1}{c+m+1}+\frac{c+m}{c+m+1}=1.$$

So we can write

$$\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\sum_{j\neq t}\mathbb{E}[\xi_j^i\xi_t^i]=\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\Big[1-\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\Big].$$

Thus:

$$\begin{aligned}
\mathbb{E}_{\text{DP}}\big[\|\varphi(\mathbb{P}^i)-\varphi(\mathbb{P}_{XY,i}^{\text{base}})\|^2\big] &= \sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\Psi(\mathbb{P}_{XY,i}^{\text{base}})+\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\Big[1-\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\Big]-\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\\
&= \sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\Psi(\mathbb{P}_{XY,i}^{\text{base}})-\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\\
&= \sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\Big[\Psi(\mathbb{P}_{XY,i}^{\text{base}})-\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\Big]\\
&\leqslant \kappa\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]
\end{aligned}$$

By [Sethuraman and Ghosh, 2024, Lemma 2(a), 2(b)], we have the identity

$$\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]=\mathbb{E}[\xi_1^i]=\frac{1}{c+m+1}$$

The last inequality follows from $\Psi(\mathbb{P}_{XY,i}^{\text{base}})-\Gamma(\mathbb{P}_{XY,i}^{\text{base}})\leqslant\Psi(\mathbb{P}_{XY,i}^{\text{base}})=\mathbb{E}_{z\sim\mathbb{P}_{XY,i}^{\text{base}}}[k(z,z)]\leqslant\kappa$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\text{DP}}\big[\text{MMD}_k^2(\mathcal{P}_{X,Y}^{\text{DP}},\mathbb{P}_{XY}^{\text{base}})\big] &= \frac{1}{n^2}\sum_{i,\ell=1}^{n}\mathbb{E}_{\text{DP}}\left[\big\langle\varphi(\mathbb{P}^i)-\varphi(\mathbb{P}_{XY,i}^{\text{base}}),\varphi(\mathbb{P}^\ell)-\varphi(\mathbb{P}_{XY,\ell}^{\text{base}})\big\rangle_{\mathcal{H}_k}\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}_{\text{DP}}\left[\|\varphi(\mathbb{P}^i)-\varphi(\mathbb{P}_{XY,i}^{\text{base}})\|^2\right]\\
&\leqslant \frac{\kappa}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{\infty}\mathbb{E}[(\xi_j^i)^2]\\
&= \frac{\kappa}{n^2}\cdot n\cdot\frac{1}{c+m+1}
\end{aligned}$$

By Jensen's inequality:

$$\mathbb{E}_{\text{DP}}[\text{MMD}_k(\mathcal{P}_{X,Y}^{\text{DP}},\mathbb{P}_{XY}^{\text{base}})]\leqslant\sqrt{\mathbb{E}_{\text{DP}}[\text{MMD}_k^2(\mathcal{P}_{X,Y}^{\text{DP}},\mathbb{P}_{XY}^{\text{base}})]}\leqslant\frac{\sqrt{\kappa}}{\sqrt{n(c+m+1)}}.\tag{21}$$

*Bound on* $\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{base}}_{XY})$*:* We next look at

$$\mathbb{P}^{\text{base}}_{XY,i} = \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{1}{c+m}\sum_{k=1}^{m}\delta_{(\tilde{X}_{i,j},Y_i)},$$

and

$$\mathbb{P}^{\text{base}}_{XY} = \frac{c}{c+m}\,\mathbb{P}^{\text{prior}}_{XY} + \frac{m}{c+m}\,\mathbb{P}^{\text{pseudo}}_{X,Y}.$$

By definition:

$$\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{base}}_{XY}) = \|\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{base}}_{XY})\|_{\mathcal{H}_k},$$

where

$$\varphi(\mathbb{P}^{\text{base}}_{XY}) = \int k(z,\cdot)d\,\mathbb{P}^{\text{base}}_{XY}(z) = \frac{c}{c+m}\varphi(\mathbb{P}^{\text{prior}}_{XY}) + \frac{m}{c+m}\varphi\left(\mathbb{P}^{\text{pseudo}}_{X,Y}\right).$$

Hence

$$\begin{aligned}\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{base}}_{XY}) &= \varphi\left(\mathbb{P}^0_{XY}\right) - \left[\frac{c}{c+m}\varphi(\mathbb{P}^{\text{prior}}_{XY}) + \frac{m}{c+m}\varphi\left(\mathbb{P}^{\text{pseudo}}_{X,Y}\right)\right] \\ &= \frac{c}{c+m}\left(\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{prior}}_{XY})\right) + \frac{m}{c+m}\left(\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi\left(\mathbb{P}^{\text{pseudo}}_{X,Y}\right)\right).\end{aligned}$$

Applying the triangle inequality:

$$\|\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{base}}_{XY})\|_{\mathcal{H}_k} \leqslant \frac{c}{c+m}\|\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{prior}}_{XY})\|_{\mathcal{H}_k} + \frac{m}{c+m}\|\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi\left(\mathbb{P}^{\text{pseudo}}_{X,Y}\right)\|_{\mathcal{H}_k}.$$

Thus

$$\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{base}}_{XY}) = \|\varphi\left(\mathbb{P}^0_{XY}\right) - \varphi(\mathbb{P}^{\text{base}}_{XY})\|_{\mathcal{H}_k} \leqslant \frac{c}{c+m}\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{prior}}_{XY}) + \frac{m}{c+m}\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{pseudo}}_{X,Y}).$$

Combining this with (21), we have, given $(\mathcal{D}, S)$

$$\mathbb{E}_{\text{DP}}\left[\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathcal{P}^{\text{DP}}_{X,Y})\right] \leqslant \frac{\sqrt{\kappa}}{\sqrt{n(c+m+1)}} + \frac{c}{c+m}\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^{\text{prior}}_{XY}) + \frac{m}{c+m}\mathrm{MMD}_k(\mathbb{P}^0_{X,Y}, \mathbb{P}^{\text{pseudo}}_{X,Y}).$$

Taking expectation over $(\mathcal{D}, S)$ completes the proof. □

By exactly the same argument as that in Lemma 2 but applied to the marginal DP on $\mathbb{P}_{X|W}$ defined in Section 2.2.2, we have the following corollary

**Corollary 1.**

$$\begin{aligned}\mathbb{E}_{\mathcal{D},S}\,\mathbb{E}_{\text{DP}}\left[\mathrm{MMD}_{k^2_X}(\mathbb{P}^0_X, \mathcal{P}^{\text{DP}}_X) \mid \mathcal{D}, S\right] \leqslant \frac{\sqrt{\kappa}}{\sqrt{n(c+m+1)}} &+ \frac{c}{c+m}\,\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_{k^2_X}(\mathbb{P}^0_X, \mathbb{P}^{\text{prior}}_X)\right] \\ &+ \frac{m}{c+m}\,\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_{k^2_X}(\mathbb{P}^0_X, \mathbb{P}^{\text{pseudo}}_X)\right].\end{aligned}$$

We are now able to prove the theorem

*Proof of Theorem 1.* We introduce the notation $\mathcal{P}^{\text{DP}}_{X,Y} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}^{\text{DP}}_{X,Y|w_i}$ and $\mathcal{P}^{\text{DP}}_{X} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}^{\text{DP}}_{X|w_i}$

for brevity. For any $\theta \in \Theta$, using the triangular inequality and the definition of $\hat{\theta}$, we have

$$\begin{aligned}
&\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \\
&\leqslant \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_{X,Y}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \\
&\leqslant \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_{X,Y}, \mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\hat{\theta}_n)}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \\
&\leqslant \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_{X,Y}, \mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\theta)}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\hat{\theta}_n)}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \\
&\leqslant 2\,\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\theta)}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\hat{\theta}_n)}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \\
&\leqslant 2\,\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X \mathbb{P}_{g(X,\theta)}\right) \\
&\quad + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\theta)}, \mathbb{P}^0_X \mathbb{P}_{g(X,\theta)}\right) + \mathrm{MMD}_k\left(\mathcal{P}^{\mathrm{DP}}_X \mathbb{P}_{g(X,\hat{\theta}_n)}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right)
\end{aligned}$$

Taking expectations and taking the infimum over $\theta \in \Theta$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},S}\left[\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X \mathbb{P}_{g(X,\hat{\theta}_n)}\right) \mid \mathcal{D}, S\right]\right] - \inf_{\theta\in\Theta} \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^0_X \mathbb{P}_{g(X,\theta)}\right) \\
&\leqslant 2\,\mathbb{E}_{\mathcal{D},S}\left[\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathcal{P}^{\mathrm{DP}}_{X,Y}\right) \mid \mathcal{D}, S\right]\right] + 2\Lambda 2\,\mathbb{E}_{\mathcal{D},S}\left[\mathbb{E}_{\mathrm{DP}}\left[\mathrm{MMD}_k\left(\mathbb{P}^0_X, \mathcal{P}^{\mathrm{DP}}_X\right) \mid \mathcal{D}, S\right]\right] \\
&\leqslant \frac{2\sqrt{\kappa}(1+\Lambda)}{\sqrt{n(c+m+1)}} + \frac{2c}{c+m}\left[\Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathbb{P}^0_X, \mathbb{P}^{\mathrm{prior}}_X\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^{\mathrm{prior}}_{XY}\right)\right] \\
&\qquad + \frac{2m}{c+m}\left[\Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathbb{P}^0_X, \mathbb{P}^{\mathrm{pseudo}}_X\right) + \mathrm{MMD}_k\left(\mathbb{P}^0_{XY}, \mathbb{P}^{\mathrm{pseudo}}_{XY}\right)\right]
\end{aligned}$$

The first inequality follows from [Alquier and Gerber, 2024, Lemma 2], and the final line follows from Lemma 2 and Corollary 1. □

### *B.2. Proof of Theorem 2*

Before proving the theorem, we first state and prove three lemmas relating to the MMD:

**Lemma 3** (TVD to MMD)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $P, Q$ two probability measures on it. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded, positive–definite kernel with*

$$K := \sup_{x\in\mathcal{X}} k(x,x) < \infty.$$

*Denote by $\mathcal{H}_k$ the reproducing–kernel Hilbert space (RKHS) associated with $k$, let*

$$\mathrm{MMD}_k(P,Q) := \sup_{\|f\|_{\mathcal{H}_k}\leqslant 1}\left|\int_{\mathcal{X}} f\,\mathrm{d}(P-Q)\right|, \qquad \|P-Q\|_{\mathrm{TV}} := \frac{1}{2}\int_{\mathcal{X}}\left|\mathrm{d}P - \mathrm{d}Q\right|$$

*be, respectively, the maximum mean discrepancy and the (half–$L^1$-normalised) total–variation distance. Then*

$$\mathrm{MMD}_k(P,Q) \leqslant 2\sqrt{K}\|P-Q\|_{\mathrm{TV}}.$$

*Proof.* For $x \in \mathcal{X}$ write $\phi(x) \in \mathcal{H}_k$ for the canonical feature map. If $f \in \mathcal{H}_k$, then by Cauchy–Schwarz

$$|f(x)| = \left|\langle f, \phi(x)\rangle_{\mathcal{H}_k}\right| \leqslant \|f\|_{\mathcal{H}_k}\|\phi(x)\|_{\mathcal{H}_k} = \|f\|_{\mathcal{H}_k}\sqrt{k(x,x)} \leqslant \|f\|_{\mathcal{H}_k}\sqrt{K}.$$

Hence

$$\sup_{x\in\mathcal{X}}|f(x)| \leqslant \sqrt{K}\|f\|_{\mathcal{H}_k}.$$

We have the IPM representation of the TVD

$$\|P-Q\|_{\mathrm{TV}} = \frac{1}{2}\sup_{\|g\|_\infty\leqslant 1}\Big|\int_{\mathcal{X}} g\mathrm{d}(P-Q)\Big|.$$

For any $f\in\mathcal{H}_k$ with $\|f\|_{\mathcal{H}_k}\leqslant 1$, we have $\|f/\sqrt{K}\|_\infty\leqslant 1$. Therefore,

$$\Big|\int_{\mathcal{X}} f\mathrm{d}(P-Q)\Big| = \sqrt{K}\Big|\int_{\mathcal{X}}\frac{f}{\sqrt{K}}\mathrm{d}(P-Q)\Big| \leqslant 2\sqrt{K}\|P-Q\|_{\mathrm{TV}}.$$

Since this holds for all $\|f\|_{\mathcal{H}_k}\leqslant 1$, taking the supremum over the unit ball gives

$$\mathrm{MMD}_k(P,Q) = \sup_{\|f\|_{\mathcal{H}_k}\leqslant 1}\Big|\int_{\mathcal{X}} f\mathrm{d}(P-Q)\Big| \leqslant 2\sqrt{K}\|P-Q\|_{\mathrm{TV}}.$$

□

**Lemma 4** (Joint–conditional MMD reduction)**.** *Let*

$$k_X : \mathcal{X}\times\mathcal{X}\longrightarrow[0,\kappa_X], \qquad k_Y : \mathcal{Y}\times\mathcal{Y}\longrightarrow[0,\kappa_Y], \qquad 0<\kappa_X,\kappa_Y<\infty,$$

*be bounded, measurable, non–negative, positive–definite kernels. Define the* product kernel

$$k\big((x,y),(x',y')\big) := k_X(x,x')\,k_Y(y,y'), \qquad (x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}.$$

*Fix a probability measure $P_Y^0$ on $\mathcal{Y}$. For each $y\in\mathcal{Y}$ let $P_{X|y}$ and $Q_{X|y}$ be probability measures on $\mathcal{X}$, measurable in $y$. Form the* joint *laws*

$$\tilde{P}(dx,dy) := P_Y^0(dy)\,P_{X|y}(dx), \qquad \tilde{Q}(dx,dy) := P_Y^0(dy)\,Q_{X|y}(dx).$$

*Then*

$$\mathrm{MMD}_k\big(\tilde{P},\tilde{Q}\big) \;\leqslant\; \sqrt{\kappa_Y}\,\mathbb{E}_{Y\sim P_Y^0}\Big[\mathrm{MMD}_{k_X}\big(P_{X|Y},Q_{X|Y}\big)\Big]. \tag{22}$$

*Proof.* Throughout the proof we use this identity of the MMD:

$$\mathrm{MMD}_k^2(\mu,\nu) = \mathbb{E}_{X,X'\sim\mu}k(X,X') + \mathbb{E}_{Y,Y'\sim\nu}k(Y,Y') - 2\,\mathbb{E}_{X\sim\mu,\,Y\sim\nu}k(X,Y). \tag{23}$$

Apply (23) with $\mu=\tilde{P}$ and $\nu=\tilde{Q}$:

$$\mathrm{MMD}_k^2(\tilde{P},\tilde{Q}) = \underbrace{\mathbb{E}_{(X,Y),(X',Y')\sim\tilde{P}}K}_{(A)} + \underbrace{\mathbb{E}_{(X,Y),(X',Y')\sim\tilde{Q}}K}_{(B)} - 2\,\underbrace{\mathbb{E}_{(X,Y)\sim\tilde{P},\,(X',Y')\sim\tilde{Q}}K}_{(C)}, \tag{24}$$

where we recall that $k\big((x,y),(x',y')\big) = k_X(x,x')\,k_Y(y,y')$ is the product kernel. Because $K\leqslant\kappa_X\kappa_Y<\infty$, all integrands are bounded, and Fubini's lemma allows us to change

integration order freely.

$$(A) = \iint\limits_{\mathcal{Y}^2} P_Y^0(dy)\, P_Y^0(dy') \iint\limits_{\mathcal{X}^2} P_{X|y}(dx)\, P_{X|y'}(dx')\; k_X(x,x')\, k_Y(y,y'). \tag{25}$$

$$(B) = \iint\limits_{\mathcal{Y}^2} P_Y^0(dy)\, P_Y^0(dy') \iint\limits_{\mathcal{X}^2} Q_{X|y}(dx)\, Q_{X|y'}(dx')\; k_X(x,x')\, k_Y(y,y'). \tag{26}$$

$$(C) = \iint\limits_{\mathcal{Y}^2} P_Y^0(dy)\, P_Y^0(dy') \iint\limits_{\mathcal{X}^2} P_{X|y}(dx)\, Q_{X|y'}(dx')\; k_X(x,x')\, k_Y(y,y'). \tag{27}$$

For each $(y, y') \in \mathcal{Y}^2$ define

$$\begin{aligned}\Delta(y,y') := \iint\limits_{\mathcal{X}^2} k_X(x,x') \Big\{ & P_{X|y}(dx)\, P_{X|y'}(dx') - P_{X|y}(dx)\, Q_{X|y'}(dx') \\ & - Q_{X|y}(dx)\, P_{X|y'}(dx') + Q_{X|y}(dx)\, Q_{X|y'}(dx') \Big\}.\end{aligned} \tag{28}$$

Substituting (25)–(27) into (24) yields (because the integrand and measure are symmetric in $(y, y')$):

$$\mathrm{MMD}_k^2(\tilde{P}, \tilde{Q}) = \iint\limits_{\mathcal{Y}^2} k_Y(y,y')\, \Delta(y,y')\, P_Y^0(dy)\, P_Y^0(dy'). \tag{29}$$

We have the identity:

$$\begin{aligned}\iint\limits_{\mathcal{X}^2} k_X(x,x')\, P_{X|y}(dx)\, P_{X|y'}(dx') &= \iint\limits_{\mathcal{X}^2} \big\langle \phi_X(x), \phi_X(x') \big\rangle P_{X|y}(dx)\, P_{X|y'}(dx') \\ &= \Big\langle \int_{\mathcal{X}} \phi_X(x)\, P_{X|y}(dx),\; \int_{\mathcal{X}} \phi_X(x')\, P_{X|y'}(dx') \Big\rangle \\ &= \big\langle m_P(y),\, m_P(y') \big\rangle,\end{aligned} \tag{30}$$

where $m_P(y)$ is the kernel mean embedding of $P_{X|y}$: $m_P(y) := m_{P_{X|y}}$, and $\phi$ is the feature map. The second equality uses Fubini's lemma (the integrand is bounded). Similarly, we define $m_Q(y) := m_{Q_{X|y}}$. Using similar calculations to that of (30) in all components of $\Delta(y, y')$ in (28), we have

$$\Delta(y,y') = \big\langle m_P(y) - m_Q(y),\; m_P(y') - m_Q(y') \big\rangle_{\mathcal{H}_{k_X}},$$

By Cauchy–Schwarz:

$$|\Delta(y,y')| \;\leqslant\; \|m_P(y) - m_Q(y)\|\, \|m_P(y') - m_Q(y')\|. \tag{31}$$

By definition,

$$\mathrm{MMD}_{k_X}\big(P_{X|y}, Q_{X|y}\big) = \|m_P(y) - m_Q(y)\| \geqslant 0. \tag{32}$$

Substituting (32) into (31) yields

$$|\Delta(y,y')| \;\leqslant\; \mathrm{MMD}_{k_X}\big(P_{X|y}, Q_{X|y}\big)\;\mathrm{MMD}_{k_X}\big(P_{X|y'}, Q_{X|y'}\big) \qquad \forall (y,y') \in \mathcal{Y}^2. \tag{33}$$

Insert (33) into (29) and use the fact that $k_Y \geqslant 0$:

$$\begin{aligned}
\mathrm{MMD}_k^2(\tilde{P},\tilde{Q}) &= \iint_{\mathcal{Y}^2} k_Y(y,y')\,\Delta(y,y')\,P_Y^0(dy)\,P_Y^0(dy') \\
&\leqslant \iint_{\mathcal{Y}^2} k_Y(y,y')\;|\Delta(y,y')|\;P_Y^0(dy)\,P_Y^0(dy') \\
&\leqslant \iint_{\mathcal{Y}^2} k_Y(y,y')\;\mathrm{MMD}_{k_X}\big(P_{X|y}, Q_{X|y}\big)\;\mathrm{MMD}_{k_X}\big(P_{X|y'}, Q_{X|y'}\big)\,P_Y^0(dy)\,P_Y^0(dy'). \\
&\leqslant \kappa_Y \iint_{\mathcal{Y}^2} \mathrm{MMD}_{k_X}\big(P_{X|y}, Q_{X|y}\big)\;\mathrm{MMD}_{k_X}\big(P_{X|y'}, Q_{X|y'}\big)\,P_Y^0(dy)\,P_Y^0(dy') \\
&= \kappa_Y \Big(\int_{\mathcal{Y}} \mathrm{MMD}_{k_X}\big(P_{X|y}, Q_{X|y}\big)\,P_Y^0(dy)\Big)^2 \\
&= \kappa_Y \left(\mathbb{E}_{Y\sim P_Y^0}\Big[\mathrm{MMD}_{k_X}\big(P_{X|Y}, Q_{X|Y}\big)\Big]\right)^2
\end{aligned}$$

Taking square roots on both sides completes the proof. □

**Lemma 5** (Jensen's inequality for the MMD)**.**

$$\mathrm{MMD}(P(X), Q(X)) \leqslant \int_{\mathcal{Y}} \mathrm{MMD}(P(X \mid y), Q(X \mid y))p(y)\,\mathrm{d}y \tag{34}$$

*We denote it Jensen's inequality because the LHS equals* $\mathrm{MMD}(\mathbb{E}_Y\,P(X \mid Y), \mathbb{E}_Y\,Q(X \mid Y))$*, and the RHS is* $\mathbb{E}_Y\,\mathrm{MMD}(P(X \mid Y), Q(X \mid Y))$

*Proof.* For any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leqslant 1$:

$$\begin{aligned}
\left|\int_{\mathcal{X}} f(x)P(x)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x)\,\mathrm{d}x\right| &= \left|\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x)P(x \mid y)p(y)\,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}}\int_{\mathcal{Y}} f(x)Q(x \mid y)p(y)\,\mathrm{d}y\,\mathrm{d}x\right| \\
&= \left|\int_{\mathcal{Y}}\left[\int_{\mathcal{X}} f(x)P(x \mid y)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x \mid y)\,\mathrm{d}x\right]p(y)\,\mathrm{d}y\right| \\
&\leqslant \int_{\mathcal{Y}}\left|\int_{\mathcal{X}} f(x)P(x \mid y)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x \mid y)\,\mathrm{d}x\right|p(y)\,\mathrm{d}y.
\end{aligned}$$

Because $|f(x)| \leqslant \|f\|_{\mathcal{H}}\sqrt{k(x,x)} \leqslant K$ with $K := \sup_{x\in\mathcal{X}}\sqrt{k(x,x)} < \infty$, the integrand is absolutely integrable. Hence, Fubini's theorem allows the change in the order of integration in the second equality. The last inequality follows from triangle inequality. Now by definition

of the MMD:

$$\begin{aligned}\mathrm{MMD}(P(X), Q(X)) &= \sup_{f:\|f\|_{\mathcal{H}}\leqslant 1} \left| \int_{\mathcal{X}} f(x)P(x)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x)\,\mathrm{d}x \right| \\ &\leqslant \sup_{f:\|f\|_{\mathcal{H}}\leqslant 1} \int_{\mathcal{Y}} \left| \int_{\mathcal{X}} f(x)P(x \mid y)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x \mid y)\,\mathrm{d}x \right| p(y)\,\mathrm{d}y\end{aligned}$$

By triangle inequality. For every $f : \|f\|_{\mathcal{H}} \leqslant 1$ and $y \in \mathcal{Y}$, we have

$$\begin{aligned}\left| \int_{\mathcal{X}} f(x)P(x \mid y)\,\mathrm{d}x - \int_{\mathcal{X}} f(x)Q(x \mid y)\mathrm{d}x \right| &\leqslant \sup_{g:\|g\|_{\mathcal{H}}\leqslant 1} \left| \int_{\mathcal{X}} g(x)P(x \mid y)\,\mathrm{d}x - \int_{\mathcal{X}} g(x)Q(x \mid y)\,\mathrm{d}x \right| \\ &= \mathrm{MMD}(p(X \mid y), Q(X|y)).\end{aligned}$$

Taking integral over $\mathcal{Y}$ on both sides and then taking sup over $f : \|f\|_{\mathcal{H}} \leqslant 1$ finishes the proof. □

We are now ready to prove Theorem 2.

*Proof of Theorem 2.* For clarity we collect the notation that governs the pseudo–sampling scheme. Given a pair $(w, y)$ and a data set of size $n$, the (misspecified) posterior predictive distribution of $X$ is

$$\Psi_n(\cdot \mid w, y) := \int_{\Theta} \Pi_\theta(\cdot \mid w, y)\Pi_n(\mathrm{d}\theta \mid \mathcal{D}_{1:n}),$$

and its marginal mixture with the true data law is

$$\Psi_n^{XY}(\mathrm{d}x, \mathrm{d}y) := \int_{\mathcal{W}} \Psi_n(\mathrm{d}x \mid w, y)p^0_{WY}(\mathrm{d}w, \mathrm{d}y).$$

Fix an integer $m \geqslant 1$. Conditional on the observed sample $\mathcal{D}_{1:n}$ we draw, for each $i = 1, \ldots, n$,

$$\tilde{X}_{ij} \overset{\text{iid}}{\sim} \Psi_n(\cdot \mid W_i, Y_i), \qquad j = 1, \ldots, m,$$

and collect all pseudo–draws in $S := \{\tilde{X}_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$. For later bounds we view each observation as generating two probability measures on $\mathcal{X} \times \mathcal{Y}$,

$$\mu_i := \Psi_n(\cdot \mid W_i, Y_i)\delta_{Y_i}, \qquad \widehat{\mu}_i := \frac{1}{m}\sum_{j=1}^{m} \delta_{(\tilde{X}_{ij}, Y_i)},$$

respectively, the exact posterior predictive and its empirical counterpart based on $m$ pseudo–samples. Averaging over $i$ produces the reference measure and the empirical ("pseudo") measure

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} \mu_i, \qquad \mathbb{P}^{\text{pseudo}}_{XY} := \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i.$$

Finally, we define the joint distribution of $(X, Y)$ in the posterior model implied by $\theta^*$ as

$$\Pi^{XY}_{\theta^*}(\mathrm{d}x, \mathrm{d}y) = \int_{\mathcal{W}\times\mathcal{Y}} \Pi_{\theta^*}(\mathrm{d}x \mid w, y)\delta_y(\mathrm{d}y)p^0_{WY}(\mathrm{d}w, \mathrm{d}y).$$

By the triangle inequality:

$$\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \mathbb{P}_{XY}^0) \leqslant \mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n) + \mathrm{MMD}_k(Q_n, \Pi_{\theta^*}^{XY}) + \mathrm{MMD}_k(\Pi_{\theta^*}^{XY}, \mathbb{P}_{XY}^0).$$

Take $\mathbb{E}_{\mathcal{D},S}$ on both sides, noting that the second term does not depend on $S$ and the third term does not depend on $\mathcal{D}$ or $S$, we have

$$\mathbb{E}_{\mathcal{D},S}\,\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \mathbb{P}_{XY}^0) \leqslant \underbrace{\mathbb{E}_{\mathcal{D},S}\,\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)}_{\text{term A}} + \underbrace{\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_k(Q_n, \Pi_{\theta^*}^{XY})}_{\text{term B}} + \underbrace{\mathrm{MMD}_k(\Pi_{\theta^*}^{XY}, \mathbb{P}_{XY}^0)}_{\text{term C}}.$$

*Step 1: Bounding term A.*

For any probability measure $\nu$ on $\mathcal{X} \times \mathcal{Y}$ we write

$$\varphi(\nu) := \mathbb{E}_{(X,Y)\sim\nu}\big[k\big((X,Y),\cdot\big)\big] \in \mathcal{H}_k, \qquad \mathrm{MMD}_k(\nu_1,\nu_2) := \big\|\varphi(\nu_1) - \varphi(\nu_2)\big\|_{\mathcal{H}_k}.$$

For a fixed observation pair $(W_i, Y_i)$ let

$$\mathbb{Q}_i := \Psi_n\big(\cdot \mid W_i, Y_i\big)\delta_{Y_i} \quad \Big(\text{probability on } \mathcal{X}\times\mathcal{Y}\Big),$$

and write its embedding $\varphi(\mathbb{Q}_i) \in \mathcal{H}_k$. Given $(\mathcal{D}, S) \equiv \big\{(W_i, Y_i); (\tilde{X}_{ij})_{j=1}^m\big\}$ we form the empirical measure based on the $m$ pseudo–samples

$$\widehat{\mathbb{Q}}_i := \frac{1}{m}\sum_{j=1}^m \delta_{(\tilde{X}_{ij}, Y_i)}, \qquad \varphi\big(\widehat{\mathbb{Q}}_i\big) = \frac{1}{m}\sum_{j=1}^m k\big((\tilde{X}_{ij}, Y_i),\cdot\big) \in \mathcal{H}_k.$$

Conditioned on the data set $\mathcal{D}_{1:n}$, the random vectors

$$\varphi_{ij} := k\big((\tilde{X}_{ij}, Y_i),\cdot\big) \in \mathcal{H}_k, \qquad j = 1,\dots,m,$$

are i.i.d. with mean $\mathbb{E}_{S|\mathcal{D}}[\varphi_{ij}] = \varphi(\mathbb{Q}_i)$, because each $\tilde{X}_{ij}$ is drawn from $\Psi_n(\cdot \mid W_i, Y_i)$. Moreover $\|\varphi_{ij}\|_{\mathcal{H}_k}^2 = k\big((\tilde{X}_{ij}, Y_i), (\tilde{X}_{ij}, Y_i)\big) \leqslant \kappa_X\kappa_Y$, so $\|\varphi_{ij}\|_{\mathcal{H}_k} \leqslant \sqrt{\kappa_X\kappa_Y}$.

We have $\mathbb{E}_{S|\mathcal{D}}[\varphi_{ij} - \varphi(\mathbb{Q}_i)] = 0$ and $\|\varphi_{ij} - \varphi(\mathbb{Q}_i)\|_{\mathcal{H}_k} \leqslant \|\varphi_{ij}\|_{\mathcal{H}_k} + \|\varphi(\mathbb{Q}_i)\|_{\mathcal{H}_k} \leqslant 2\sqrt{\kappa_X\kappa_Y}$.

The difference of embeddings admits the representation

$$\varphi\big(\widehat{\mathbb{Q}}_i\big) - \varphi(\mathbb{Q}_i) = \frac{1}{m}\sum_{j=1}^m \left[\varphi_{ij} - \varphi(\mathbb{Q}_i)\right].$$

Conditional on $\mathcal{D}$ the $\varphi_{ij} - \varphi(\mathbb{Q}_i)$ are independent and centred, so

$$\begin{aligned}\mathbb{E}_{S|\mathcal{D}}\left[\left\|\varphi(\widehat{\mathbb{Q}}_i) - \varphi(\mathbb{Q}_i)\right\|_{\mathcal{H}_k}^2\right] &= \mathbb{E}_{S|\mathcal{D}}\left[\frac{1}{m^2}\left\|\sum_{j=1}^m \varphi_{ij} - \varphi(\mathbb{Q}_i)\right\|_{\mathcal{H}_k}^2\right] \\ &= \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}_{S|\mathcal{D}}\left[\|\varphi_{ij} - \varphi(\mathbb{Q}_i)\|_{\mathcal{H}_k}^2\right] \\ &\leqslant \frac{4\kappa_X\kappa_Y}{m}.\end{aligned}$$

Observe that $\mathbb{E}_{S|\mathcal{D}}[\varphi(\widehat{\mathbb{Q}}_i) - \varphi(\mathbb{Q}_i)] = \mathbb{E}_{S|\mathcal{D}}[\frac{1}{m}\sum_{j=1}^m \varphi_{ij} - \varphi(\mathbb{Q}_i)] = 0$, since $\varphi(\widehat{\mathbb{Q}}_i)$ are independent accross $i$ given $\mathcal{D}$, we have

$$\begin{aligned}\mathbb{E}_{S|\mathcal{D}}\big[\mathrm{MMD}_k^2(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)\big] &= \mathbb{E}_{S|\mathcal{D}}\left[\left\|\varphi\left(\frac{1}{n}\sum_{i=1}^n \widehat{\mathbb{Q}}_i\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^n \mathbb{Q}_i\right)\right\|_{\mathcal{H}_k}^2\right] \\ &= \mathbb{E}_{S|\mathcal{D}}\left[\left\|\frac{1}{n}\sum_{i=1}^n\Big(\varphi(\widehat{\mathbb{Q}}_i) - \varphi(\mathbb{Q}_i)\Big)\right\|_{\mathcal{H}_k}^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{S|\mathcal{D}}\left[\left\|\varphi(\widehat{\mathbb{Q}}_i) - \varphi(\mathbb{Q}_i)\right\|_{\mathcal{H}_k}^2\right] \\ &\leqslant \frac{4\kappa_X\kappa_Y}{nm}.\end{aligned} \tag{35}$$

Jensen's inequality implies

$$\mathbb{E}_{S|\mathcal{D}}\Big[\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)\Big] \leqslant \sqrt{\mathbb{E}_{S|\mathcal{D}}\left[\mathrm{MMD}_k^2(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)\right]} \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{nm}}.$$

The RHS is deterministic. Taking expectation with respect to $\mathcal{D}$ yields

$$\mathbb{E}_{\mathcal{D},S}\big[\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)\big] = \mathbb{E}_{\mathcal{D}}\,\mathbb{E}_{S|\mathcal{D}}\big[\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n)\big] \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{nm}} \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}}. \tag{36}$$

*Step 2: Bounding term B.*

By triangle inequality:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k(Q_n, \Pi_{\theta*}^{XY})\right] \leqslant\ & \mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k\left(Q_n, \frac{1}{n}\sum_{i=1}^n \Psi_{\theta*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right)\right] && \text{term B.1} \\ &+ \mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^n \Psi_{\theta*}(\cdot \mid W_i, Y_i)\delta_{Y_i}, \Pi_{\theta*}^{XY}\right)\right] && \text{term B.2.}\end{aligned}$$

We will first bound term B.1 by the misspecified Bernstein von-Mises theorem established

by Kleijn and Van der Vaart [2012]. For each $(W_i, Y_i)$

$$\begin{aligned}
&\mathrm{MMD}_{k_X}\left(\Psi_n(\cdot \mid W_i, Y_i), \Psi_{\theta^*}(\cdot \mid W_i, Y_i)\right) \\
&= \mathrm{MMD}_{k_X}\left(\int_\Theta \Pi_\theta(\cdot \mid W_i, Y_i)\Pi_n(\theta)\mathrm{d}\theta, \int_\Theta \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\Pi_n(\theta)\mathrm{d}\theta\right) \\
&\leqslant \int_\Theta \mathrm{MMD}_{k_X}\left(\Pi_\theta(\cdot \mid W_i, Y_i), \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\right)\Pi_n(\theta)\mathrm{d}\theta \qquad \text{(Lemma 5)} \\
&= \int_{B_n + B_n^C} \mathrm{MMD}_{k_X}\left(\Pi_\theta(\cdot \mid W_i, Y_i), \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\right)\Pi_n(\theta)\mathrm{d}\theta,
\end{aligned}$$

where $B_n := \{\theta : \|\theta - \theta^*\| \leqslant M_n/\sqrt{n}\}$, and $M_n$ is any sequence such that $M_n \to \infty$. We first choose $M_n$ such that $M_n \leqslant \sqrt{n}\sup_{\Theta_\rho}\|\theta - \theta^*\|$ for all $n \geqslant 1$. Then $B_n \subset \Theta_\rho$ for all $n$. The TVD Lipschitz condition A2 and Lemma 3 gives the MMD Lipschitz property:

$$\begin{aligned}
\mathrm{MMD}_{k_X}\left(\Pi_\theta(\cdot \mid W_i, Y_i), \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\right) &\leqslant L(W_i, Y_i)\,\|\theta - \theta^*\| \\
&\leqslant \frac{M_n}{\sqrt{n}} L(W_i, Y_i), \qquad \forall \theta \in B_n.
\end{aligned} \tag{37}$$

Therefore,

$$\int_{B_n} \mathrm{MMD}_{k_X}\left(\Pi_\theta(\cdot \mid W_i, Y_i), \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\right)\Pi_n(\theta)\mathrm{d}\theta \leqslant \frac{M_n}{\sqrt{n}} L(W_i, Y_i)\Pi_n(B_n) \leqslant \frac{M_n}{\sqrt{n}} L(W_i, Y_i). \tag{38}$$

We denote $\tilde{r}_n := \Pi_n(B_n^C) \leqslant 1$, then by assumption A1, $\Pi_n$ satisfies posterior contraction [Kleijn and Van der Vaart, 2012, Theorem 3.1]:

$$\int_{B_n^C} \mathrm{MMD}_{k_X}\left(\Pi_\theta(\cdot \mid W_i, Y_i), \Pi_{\theta^*}(\cdot \mid W_i, Y_i)\right)\Pi_n(\theta)\mathrm{d}\theta \leqslant \sqrt{\kappa_X}\Pi_n(B_n^C) = \sqrt{\kappa_X}\tilde{r}_n. \tag{39}$$

Combining (38) and (39), we have

$$\mathrm{MMD}_{k_X}\left(\Psi_n(\cdot \mid W_i, Y_i), \Psi_{\theta^*}(\cdot \mid W_i, Y_i)\right) \leqslant \frac{M_n}{\sqrt{n}} L(W_i, Y_i) + \sqrt{\kappa_X}\tilde{r}_n.$$

By Lemma 4:

$$\mathrm{MMD}_k\left(\Psi_n(\cdot \mid W_i, Y_i)\delta_{Y_i}, \Psi_{\theta^*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right) \leqslant \frac{\sqrt{\kappa_Y}M_n}{\sqrt{n}} L(W_i, Y_i) + \sqrt{\kappa_X\kappa_Y}\tilde{r}_n.$$

By the triangle inequality in $\mathcal{H}_k$ and the linearity of kernel mean embedding $\varphi(\cdot)$:

$$\begin{aligned}
\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^n P_i, \frac{1}{n}\sum_{i=1}^n Q_i\right) &= \left\|\varphi\left(\frac{1}{n}\sum_{i=1}^n P_i\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^n Q_i\right)\right\|_{\mathcal{H}_k} \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \|P_i - Q_i\|_{\mathcal{H}_k} \\
&= \frac{1}{n}\sum_{i=1}^n \mathrm{MMD}_k(P_i, Q_i).
\end{aligned}$$

Therefore,

$$\mathrm{MMD}_k\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\Psi_n(\cdot \mid W_i, Y_i)\delta_{Y_i}}_{=Q_n}, \frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta^*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right) \leqslant \frac{\sqrt{\kappa_Y}M_n}{\sqrt{n}}\bar{L}(W,Y) + \sqrt{\kappa_X\kappa_Y}\tilde{r}_n,$$

where $\bar{L}(W,Y) = \frac{1}{n}\sum_{i=1}^{n} L(W_i, Y_i)$. Taking expectation over $\mathcal{D} = \{(W_i, Y_i)\}$ gives

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_k\left(Q_n, \frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta^*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right) \leqslant \frac{\sqrt{\kappa_Y}M_n}{\sqrt{n}}\underbrace{\mathbb{E}_{W,Y\sim P^0}[L(W,Y)]}_{:=C_L<\infty \text{ by } A2} + \sqrt{\kappa_Y}\,\mathbb{E}_{\mathcal{D}}[\tilde{r}_n]. \quad (40)$$

Note that (40) holds for any any $M_n$ such that $M_n \leqslant \sqrt{n}\sup_{\Theta_\rho}\|\theta - \theta^*\|$ for all $n$. We can scale with $M_n' := M_n/\max\{\sqrt{\kappa_Y}C_L, 1\}$, then $M_n'$ is divergent, and $M_n' \leqslant M_n \leqslant \sqrt{n}\sup_{\Theta_\rho}\|\theta - \theta^*\|$. Substituting $M_n$ with $M_n'$ in the arguments above yields

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_k\left(Q_n, \frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta^*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right) &\leqslant \frac{\sqrt{\kappa_Y}C_L M_n'}{\sqrt{n}} + \sqrt{\kappa_X\kappa_Y}\,\mathbb{E}_{\mathcal{D}}[\tilde{r}_n'] \\ &\leqslant \frac{M_n}{\sqrt{n}} + \sqrt{\kappa_X\kappa_Y}\,\mathbb{E}_{\mathcal{D}}[\tilde{r}_n']\end{aligned}$$

By Assumption A1, $r_n := \mathbb{E}_{\mathcal{D}}[\tilde{r}_n'] \leqslant 1$ satisfies [Kleijn and Van der Vaart, 2012, Theorem 3.1]:

$$r_n = \mathbb{E}_{\mathcal{D}}[\tilde{r}_n'] = \mathbb{E}_{\mathcal{D}}\left[\Pi_n(\|\theta - \theta^*\| \geqslant M_n'/\sqrt{n})\right] \to 0.$$

Therefore,

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_k\left(Q_n, \frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta^*}(\cdot \mid W_i, Y_i)\delta_{Y_i}\right) \leqslant \frac{M_n}{\sqrt{n}} + \sqrt{\kappa_X\kappa_Y}r_n. \quad (41)$$

The condition $M_n \leqslant \sqrt{n}\sup_{\Theta_\rho}\|\theta - \theta^*\|$ can be dropped. For any divergent $M_n$, we let $M_n'' := \min\{M_n, \sqrt{n}\sup_{\Theta_\rho}\|\theta - \theta^*\|\}$ for each $n$, then $M_n''$ is also divergent, and $M_n'' \leqslant M_n$ for all $n$. Applying (41) with $M_n''$ shows that (41) holds with any divergence $M_n$.

Next, we bound term B.2 via finite sample convergence of iid observations. For each observation $(W_i, Y_i)$ we denote

$$\mu_i := \Psi_{\theta^*}\big(\cdot \mid W_i, Y_i\big)\delta_{Y_i},$$

Then

$$\mathbb{E}_{(W_i,Y_i)\sim p^0_{WY}}[\mu_i(\mathrm{d}x, \mathrm{d}y)] = \int_{\mathcal{W}\times\mathcal{Y}}\Psi_{\theta^*}\big(\mathrm{d}x \mid w, y\big)\delta_y(\mathrm{d}y)p^0_{WY}(\mathrm{d}w, \mathrm{d}y) = \Pi^{XY}_{\theta^*}(\mathrm{d}x, \mathrm{d}y)$$

Define the RKHS embedding

$$\xi_i := \int_{\mathcal{X}\times\mathcal{Y}} k((x,y),\cdot)\mu_i(\mathrm{d}x, \mathrm{d}y) \in \mathcal{H}_k.$$

Because the pairs $(W_i, Y_i)$ are i.i.d. under the data-generating law $p^0_{WY}$, the vectors $\xi_1, \ldots, \xi_n$

are i.i.d. in $\mathcal{H}_k$ with a common mean

$$\xi_* := \int_{\mathcal{X}\times\mathcal{Y}} k((x,y),\cdot)\Pi_{\theta^*}^{XY}(\mathrm{d}x,\mathrm{d}y) = \int_{\mathcal{X}\times\mathcal{Y}} k((x,y),\cdot)\,\mathbb{E}_{W_i,Y_i\sim p^0_{WY}}\big[\mu_i(\mathrm{d}x,\mathrm{d}y)\big] = \mathbb{E}_{W_i,Y_i\sim p^0_{WY}}\big[\xi_i\big].$$

The last equality follows by Fubini's theorem since $k$ is bounded. By the independence of $\xi_1,\ldots,\xi_n$, we have

$$\mathbb{E}_{\mathcal{D}}\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \xi_*\right\|_{\mathcal{H}_k}^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{D}}\,\|\xi_i-\xi_*\|_{\mathcal{H}_k}^2 \leqslant \frac{4\kappa_X\kappa_Y}{n}.$$

By Jensen's inequality:

$$\mathbb{E}_{\mathcal{D}}\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \xi_*\right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}}.$$

The LHS is exactly the definition of the expectation of the MMD:

$$\mathbb{E}_{\mathcal{D}}\Big[\mathrm{MMD}_k\Big(\frac{1}{n}\sum_{i=1}^n \Psi_{\theta*}(\cdot\mid W_i,Y_i)\delta_{Y_i},\Pi_{\theta*}^{XY}\Big)\Big] \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}}. \tag{42}$$

Combining term B.1 (41) and term B.2 (42) gives

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k(Q_n,\Pi_{\theta^*}^{XY})\right] \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + \sqrt{\kappa_X\kappa_Y}r_n. \tag{43}$$

*Step 3: Bounding term C.*

By the chain rule of Kullback-Leibler divergence:

$$\mathrm{KL}\big(p_0\|p_{\theta*}\big) = \mathrm{KL}\big(p^0_{W,Y}\|p^{\theta*}_{W,Y}\big) + \mathbb{E}_{W,Y\sim p^0_{W,Y}}\,\mathrm{KL}\big(\Pi_0(\cdot\mid W,Y)\|\Pi_{\theta*}(\cdot\mid W,Y)\big).$$

Since the first term is non-negative, we have

$$\mathbb{E}_{W,Y\sim p^0_{W,Y}}\,\mathrm{KL}\left(\Pi_0(\cdot\mid W,Y)\|\Pi_{\theta*}(\cdot\mid W,Y)\right) \leqslant \mathrm{KL}\big(p_0\|p_{\theta*}\big).$$

For any $(W,Y)$, we have

$$\mathrm{MMD}_{k_X}(\Pi_0,\Pi_{\theta*}) \leqslant 2\sqrt{\kappa_X}\,\|\Pi_0-\Pi_{\theta*}\|_{\mathrm{TV}} \leqslant 2\sqrt{\kappa_X}\sqrt{1-\exp(-\mathrm{KL}(\Pi_0\|\Pi_{\theta*}))},$$

by the Bretagnolle–Huber inequality [Bretagnolle and Huber, 1979]. Since $\sqrt{1-\exp(-x)}$ is concave, we have, by (reversed) Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{W,Y\sim p^0_{WY}}\,\mathrm{MMD}_{k_X}\left(\Pi_0(X\mid W,Y),\Pi_{\theta*}(X\mid W,Y)\right) &\leqslant 2\sqrt{\kappa_X}\,\mathbb{E}_{W,Y\sim p^0_{WY}}\left[\sqrt{1-\exp(-\mathrm{KL}(\Pi_0\|\Pi_{\theta*}))}\right]\\
&\leqslant 2\sqrt{\kappa_X}\sqrt{1-\exp\left(-\mathbb{E}_{W,Y\sim p^0_{WY}}\,\mathrm{KL}(\Pi_0\|\Pi_{\theta*})\right)}\\
&\leqslant 2\sqrt{\kappa_X}\sqrt{1-\exp\left(-\mathrm{KL}(p_0\|p_{\theta*})\right)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathrm{MMD}_k\big(\mathbb{P}^0_{XY}, \Pi^{XY}_{\theta*}\big)\\
&= \mathrm{MMD}_k\Big(\int_{\mathcal{W}\times\mathcal{Y}} \Pi_0(\mathrm{d}x \mid w,y)\delta_y(\mathrm{d}y)p^0_{WY}(\mathrm{d}w,\mathrm{d}y), \int_{\mathcal{W}\times\mathcal{Y}} \Pi_{\theta*}(\mathrm{d}x \mid w,y)\delta_y(\mathrm{d}y)p^0_{WY}(\mathrm{d}w,\mathrm{d}y)\Big)\\
&\leqslant \mathbb{E}_{(W,Y)\sim p^0_{WY}} \mathrm{MMD}_k\big(\Pi_0(\cdot \mid W,Y)\delta_Y, \Pi_{\theta*}(\cdot \mid W,Y)\delta_Y\big) \quad \text{(Lemma 5)}\\
&\leqslant \sqrt{\kappa_Y}\, \mathbb{E}_{(W,Y)\sim p^0_{WY}} \mathrm{MMD}_{k_X}\big(\Pi_0(\cdot \mid W,Y), \Pi_{\theta*}(\cdot \mid W,Y)\big) \quad \text{(Lemma 4)}\\
&\leqslant 2\sqrt{\kappa_X\kappa_Y}\sqrt{1-\exp\big(-\mathrm{KL}(p_0\|p_{\theta*})\big)}.
\end{aligned}$$

Finally, we have the identity

$$\mathrm{KL}(p_0\|p_{\theta*}) = \mathrm{KL}\left(p^0_X(X)f^0_N(W-X)f^0_E(Y-g_0(X))\|p_X(X)f_N(W-X)f_E(Y-g(\theta^*,X))\right).$$

Since $W - X = N \perp X$, and $Y \mid X \perp W$, we have, by the chain rule

$$\mathrm{KL}(p_0\|p_{\theta*}) = \mathrm{KL}(p^0_X\|p_X) + \mathrm{KL}(f^0_N\|f_N) + \mathbb{E}_{X\sim p^0_X}\mathrm{KL}\Big(p^0_{Y|X}\Big\|p^{\theta*}_{Y|X}\Big) = \mathrm{KL}_X + \mathrm{KL}_N + \mathrm{KL}_E := \mathrm{KL}_*,$$

which gives

$$\mathrm{MMD}_k\big(\mathbb{P}^0_{XY}, \Pi^{XY}_{\theta*}\big) \leqslant 2\sqrt{\kappa_X\kappa_Y}\sqrt{1-\exp\big(-\mathrm{KL}_*\big)}. \tag{44}$$

Combining term A (36), term B (43), and term C (44) and recalling that $\kappa := \kappa_X\kappa_Y$ proves Theorem 2. □

### *B.3. Proof of Proposition 1*

*Proof of Proposition 1.* The proof mirrors that of Theorem 2, with the joint kernel $k = k_X k_Y$ replaced everywhere by $k^2_X$ and every step using Lemma 4 omitted; for brevity we will only record the changes.

*Term A (Monte-Carlo error).* For each $(W_i, Y_i)$ let $\mu_i := \Psi_n(\cdot \mid W_i, Y_i)$ and $\hat{\mu}_i := m^{-1}\sum_j \delta_{\tilde{X}_{ij}}$. For $k^2_X$ the kernel mean embedding is $\varphi_{k^2_X}(\mu) = \mathbb{E}_{X,X'\sim\mu}[k_X(X,\cdot)k_X(X',\cdot)]$, whose norm is bounded by $\kappa_X$. The exact same Hoeffding inequality in the RKHS $\mathcal{H}_{k_X}$ yields

$$\mathbb{E}_{\mathcal{D},S}\,\mathrm{MMD}_{k_X}\big(\mathbb{P}^{\mathrm{pseudo}}_X, Q^X_n\big) \leqslant \frac{2\kappa_X}{\sqrt{n}}, \qquad Q^X_n := \frac{1}{n}\sum_i \mu_i.$$

*Term B (posterior contraction).* The Lipschitz envelope for $\mathrm{MMD}_{k^2_X}$ is $2\kappa_X L(W,Y)$; every appearance of $\sqrt{\kappa_X}$ in the $\mathrm{MMD}_{k_X}$ bound is replaced by $\kappa_X$. Since no $Y$-kernel is used, all factors $\kappa_Y$ disappear, which gives

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{MMD}_{k^2_X}(Q^X_n, \Pi_{\theta*})\big] \leqslant \frac{2\kappa_X}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + \kappa_X r_n.$$

*Term C (model misspecification).* Applying the same Bretagnolle–Huber inequality [Bretagnolle and Huber, 1979] directly to $\Pi_0(X \mid W,Y)$ and $\Pi_{\theta*}(X \mid W,Y)$ yields

$$\mathrm{MMD}_{k^2_X}\big(\mathbb{P}^0_X, \Pi_{\theta*}\big) \leqslant 2\kappa_X\sqrt{1-\exp(-\mathrm{KL}_*)}.$$

Combining the three bounds with the triangle inequality finishes the proof. □

### B.4. Proof of Proposition 2

*Proof of Proposition 2.* We first prove the joint bound. This proof follows the structure of Theorem 2, highlighting the points at which the Berkson design requires additional justification.

*Term A (Monte-Carlo error).* Since $(W_i, Y_i)$ are still i.i.d, $\mu_i$ and $\widehat{\mu}_i$ are defined exactly as in the classical proof. The bound (36) is unchanged.

*Term B (posterior contraction).* **B.1** Contracting the misspecified posterior still relies on Kleijn and Van der Vaart [2012]'s theorem, so Inequality (41) holds.**B.2** The same variance calculation gives (42). Combining **B.1** and **B.2** recovers (43) verbatim.

*Term C (model misspecification).* Since $X - W = N \perp W$ and $Y \mid X \perp W$, by the chain rule we have

$$\begin{aligned}
\mathrm{KL}(p_0 \| p_{\theta^*}) &= \mathrm{KL}\left(p_W^0(W) f_N^0(X-W) f_E^0(Y - g^0(X) \| p_W^0(W) f_N(X-W) f_E(Y - g(\theta^*, X))\right) \\
&= \mathrm{KL}(f_N^0 \| f_N) + \mathbb{E}_{W \sim p_W^0} \, \mathbb{E}_{X \sim f_N^0(X|W)} \left[\mathrm{KL}(p_{Y|X}^0 \| p_{Y|X}^{\theta^*})\right] \\
&= \mathrm{KL}_N + \mathrm{KL}_E \\
&= \mathrm{KL}_* \, .
\end{aligned}$$

Applying the total-variation–to–KL bound as in (44) therefore gives

$$\mathrm{MMD}_k\left(\mathbb{P}_{XY}^0, \Pi_{\theta^*}^{XY}\right) \leqslant 2\kappa \sqrt{1 - \exp(-\mathrm{KL}_*)}.$$

Inserting the bounds from Steps 1–4 into the triangle inequality used at the start of the proof of Theorem 2 yields the stated inequality, with $\mathrm{KL}_*$ now equal to $\mathrm{KL}_N + \mathrm{KL}_E$. All other constants and residual terms are identical to those in the classical case. The marginal-X bound version follows exactly from that of 1. □

### B.5. Proof of Lemma 1

We will split this proof in two parts: proof for the marginal version (13) and for the joint version (12).

*Proof of the Joint MMD bound* (12). We first prove the theorem in the Berkson ME model, where

$$X = W + N, \quad Y = g_0(X) + E, \quad N \sim F_N^0, \quad E \sim F_E^0, \quad N, E \perp\!\!\!\perp W, \; N \perp\!\!\!\perp E. \tag{45}$$

By triangle inequality:

$$\mathrm{MMD}_k(\mathbb{P}_{XY}^0, \hat{\mathbb{P}}_{WY}^n) \leqslant \mathrm{MMD}_k(\mathbb{P}_{XY}^0, \mathbb{P}_{WY}^0) + \mathrm{MMD}_k(\mathbb{P}_{WY}^0, \hat{\mathbb{P}}_{WY}^n).$$

For the first term, we write

$$\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^0_{WY}) = \left\|\varphi(\mathbb{P}^0_{XY}) - \varphi(\mathbb{P}^0_{WY})\right\|_{\mathcal{H}_k}.$$

Let $\Phi(x,y) := k((x,y),\cdot) = k_X(x,\cdot)k_Y(y,\cdot) \in \mathcal{H}_k$ denote the feature map of the product kernel $k((x,y),(x',y')) = k_X(x,x')k_Y(y,y')$, and let $\Phi_X := k_X(x,\cdot)$ and $\Phi_Y := k_Y(y,\cdot)$ be the respective feature maps of $k_X$ and $k_Y$. Then we can write the kernel mean embeddings $\varphi(\cdot)$ of $\mathbb{P}^0_{XY}$ and $\mathbb{P}^0_{WY}$ as

$$\begin{aligned}\varphi(\mathbb{P}^0_{XY}) &= \mathbb{E}_{W,N,E}\left[\Phi(W+N,Y)\right] = \mathbb{E}_{W,N,E}[\Phi_X(W+N)\Phi_Y(Y))]\,,\\ \varphi(\mathbb{P}^0_{WY}) &= \mathbb{E}_{W,N,E}[\Phi(W,Y)] = \mathbb{E}_{W,N,E}\left[\Phi_X(W)\Phi_Y(Y))\right],\end{aligned} \tag{46}$$

where $Y = g^0(W+N) + E$. Taking their difference and applying the reproducing property:

$$\begin{aligned}\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^0_{WY}) &= \left\|\mathbb{E}_{W,N,E}\Big[\left(\Phi_X(W+N) - \Phi_X(W)\right)\Phi_Y(Y)\Big]\right\|_{\mathcal{H}_k}\\ &\leqslant \mathbb{E}_{W,N,E}\left[\left\|\left(\Phi_X(W+N) - \Phi_X(W)\right)\Phi_Y(Y)\right\|_{\mathcal{H}_k}\right]\\ &= \mathbb{E}_{W,N,E}\left[\left\|\Phi_X(W+N) - \Phi_X(W)\right\|_{\mathcal{H}_{k_X}}\left\|\Phi_Y(Y)\right\|_{\mathcal{H}_{k_Y}}\right]\\ &\leqslant \sqrt{\mathbb{E}_{W,N}\left[\left\|\Phi_X(W+N) - \Phi_X(W)\right\|^2_{\mathcal{H}_{k_X}}\right]}\sqrt{\mathbb{E}_{W,N,E}\left[\left\|\Phi_Y(Y)\right\|^2_{\mathcal{H}_{k_Y}}\right]}\\ &\leqslant \sqrt{\kappa_Y}\sqrt{\mathbb{E}_{W,N}\left[\left\|\Phi_X(W+N) - \Phi_X(W)\right\|^2_{\mathcal{H}_{k_X}}\right]}\end{aligned} \tag{47}$$

The first inequality follows from Jensen's inequality; the second equality holds because $\Phi_X(\cdot)\Phi_Y(\cdot)$ is a pure tensor in the RKHS $\mathcal{H}_X \otimes \mathcal{H}_Y$ with product kernel; the second inequality is Cauchy-Schwarz, and the final inequality follows by $k_Y(y,y') \leqslant \kappa_Y \quad \forall y, y'$. Recall that $k_X$ is translation-invariant, i.e., $k_X(x,x') = \psi(x-x')$ with $\psi(0) = \kappa_X$, therefore

$$\begin{aligned}\mathbb{E}_{W,N}\left[\left\|\Phi_X(W+N) - \Phi_X(W)\right\|^2_{\mathcal{H}_{k_X}}\right] &= \mathbb{E}_{W,N}\left[k_X(W,W) + k_X(W+N,W+N) - 2k_X(W+N,W)\right]\\ &= 2\kappa_X - 2\,\mathbb{E}_N[\psi(N)].\end{aligned} \tag{48}$$

Now

$$\mathrm{MMD}^2_{k_X}(\mathbb{F}^0_N, \delta_0) = \mathbb{E}_{N,N'}[\psi(N-N')] + \kappa_X - 2\,\mathbb{E}_N[\psi(N)]. \tag{49}$$

We note the following identities:

$$\begin{aligned}\mathbb{E}_{N,N'}[\psi(N-N')] &= \mathbb{E}_{N,N'}[k(N,N')] = \mathbb{E}_{N,N'}[\langle\Phi_X(N),\Phi_X(N')\rangle]\\ &= \langle\mathbb{E}_N[\Phi_X(N)],\mathbb{E}_{N'}[\Phi(N')]\rangle = \left\|\mathbb{E}_N[\Phi_X(N)]\right\|^2_{\mathcal{H}_{k_X}},\end{aligned}$$

and

$$\mathbb{E}_N[\psi(N)] = \mathbb{E}_N[k_X(N,0)] = \mathbb{E}_N[\langle\Phi_X(N),\Phi_X(0)\rangle] = \langle\mathbb{E}_N[\Phi_X(N)],\Phi_X(0)\rangle.$$

By Cauchy-Schwarz:

$$\mathbb{E}_N[\psi(N)]^2 \leqslant \left\|\mathbb{E}_N[\Phi_X(N)]\right\|^2_{\mathcal{H}_{k_X}}\left\|\Phi_X(0)\right\|^2_{\mathcal{H}_{k_X}} = \kappa_X\,\mathbb{E}_{N,N'}[\psi(N-N')]$$

Hence

$$\begin{aligned}(\kappa_X - \mathbb{E}_N[\psi(N)])^2 &= \kappa_X^2 + \mathbb{E}_N[\psi(N)]^2 - 2\kappa_X \,\mathbb{E}_N[\psi(N)] \\ &\leqslant \kappa_X^2 + \kappa_X \,\mathbb{E}_{N,N'}[\psi(N-N')] - 2\kappa_X \,\mathbb{E}_N[\psi(N)] \\ &= \kappa_X(\kappa_X + \mathbb{E}_{N,N'}[\psi(N-N')] - 2\,\mathbb{E}_N[\psi(N)]) \qquad \text{by (49)} \\ &= \kappa_X \,\mathrm{MMD}^2_{k_X}(\mathbb{F}^0_N, \delta_0)\end{aligned}$$

Taking square-root and substituting into (48):

$$\mathbb{E}_{W,N}\left[\|\Phi_X(W+N) - \Phi_X(W)\|^2_{\mathcal{H}_{k_X}}\right] \leqslant 2\sqrt{\kappa_X}\,\mathrm{MMD}_{k_X}(\mathbb{F}^0_N, \delta_0)$$

By (47):

$$\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^0_{WY}) \leqslant \sqrt{2}\kappa_Y^{1/2}\kappa_X^{1/4}\sqrt{\mathrm{MMD}_{k_X}(\mathbb{F}^0_N, \delta_0)} \tag{50}$$

For the second term, since $\mathcal{D} = (W_i, Y_i)_{i=1}^n$ consistes of iid samples $\{(W_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}^0_{WY}$, we have, by [Alquier and Gerber, 2024, Lemma S6]:

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_k(\hat{\mathbb{P}}^n_{WY}, \mathbb{P}^0_{WY}) \leqslant \frac{\sqrt{\kappa}}{\sqrt{n}}. \tag{51}$$

Here we generalised their last inequality in the proof where Alquier and Gerber [2024] assumed $k(\cdot,\cdot) \leqslant 1$ but we have $k(\cdot,\cdot) \leqslant \kappa$. Combining (50) and (51) finishes the proof for the Berkson case.

For the classical ME case, recall that

$$W = X + N, \quad Y = g_0(X) + E, \quad X \sim P^0_X, \quad N \sim F^0_N, \quad E \sim F^0_E, \quad N, E \perp\!\!\!\perp X,\ N \perp\!\!\!\perp E.$$

We only need to substitute every $W$ with $X$ in (46), (47), and (48), which gives

$$\begin{aligned}\mathrm{MMD}_k(\mathbb{P}^0_{XY}, \mathbb{P}^0_{WY}) &\leqslant \sqrt{\kappa_Y}\sqrt{\mathbb{E}_{X,N}\left[\|\Phi_X(X+N) - \Phi_X(X)\|^2_{\mathcal{H}_{k_X}}\right]} \\ &= \sqrt{\kappa_Y}\,(2\kappa_X - 2\,\mathbb{E}_N[\psi(N)]) \\ &\leqslant \sqrt{2}\kappa_Y^{1/2}\kappa_X^{1/4}\sqrt{\mathrm{MMD}_{k_X}(\mathbb{F}^0_N, \delta_0)}.\end{aligned}$$

Combining with (51), which only relies on $\{(W_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}^0_{WY}$, completes with the proof. □

*Proof of the Marginal MMD bound* (13). By the triangle inequality and then taking expectation over $\mathcal{D}$,

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_{k_X^2}(\mathbb{P}^0_X, \hat{\mathbb{P}}^n_W) \leqslant \mathrm{MMD}_{k_X^2}(\mathbb{P}^0_X, \mathbb{P}^0_W) + \mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_{k_X^2}(\mathbb{P}^0_W, \hat{\mathbb{P}}^n_W).$$

For the second term, since $W_{1:n} \overset{iid}{\sim} \mathbb{P}^0_W$ and $\sup_x k_X^2(x,x) = \kappa_X^2$, by [Alquier and Gerber, 2024, Lemma S6])

$$\mathbb{E}_{\mathcal{D}}\,\mathrm{MMD}_{k_X^2}(\mathbb{P}^0_W, \hat{\mathbb{P}}^n_W) \leqslant \frac{\kappa_X}{\sqrt{n}}. \tag{52}$$

It remains to bound the population term $\mathrm{MMD}_{k_X^2}(\mathbb{P}^0_X, \mathbb{P}^0_W)$. Let $\Phi^{(2)}_X$ be the feature map of

$k_X^2(\cdot,\cdot) = k_X \otimes k_X$. As in (47) (with $k$ replaced by $k_X^2$ and no $Y$ factor),

$$\mathrm{MMD}_{k_X^2}(\mathbb{P}_X^0, \mathbb{P}_W^0) = \left\| \mathbb{E}_{X,W}[\Phi_X^{(2)}(X) - \Phi_X^{(2)}(W)] \right\|_{\mathcal{H}_{k_X^2}} \leqslant \sqrt{\left\| \mathbb{E}_{X,W} \|\Phi_X^{(2)}(X) - \Phi_X^{(2)}(W)\|^2 \right\|_{\mathcal{H}_{k_X^2}}}.$$

By the same expansion as (48) and bounded translation–invariance of $k_X$,

$$\begin{aligned} \mathbb{E}_{X,W} \left\| \Phi_X^{(2)}(X) - \Phi_X^{(2)}(W) \right\|_{\mathcal{H}_{k_X^2}}^2 &= \mathbb{E}_{X,W}[k_X^2(X,X) + k_X^2(W,W) - 2k_X^2(X,W)] \\ &= 2\kappa_X^2 - 2\,\mathbb{E}_{X,W}[k_X(X,W)^2]. \end{aligned} \tag{53}$$

Under either Berkson ($X = W+N$) or classical ($W = X+N$) error, $k_X(X,W) = \psi(W-X) = \psi(\pm N)$, hence $k_X^2(X,W) = \psi(N)^2$ and

$$\mathbb{E}_{X,W} \|\Phi_X^{(2)}(X) - \Phi_X^{(2)}(W)\|^2 = 2\kappa_X^2 - 2\,\mathbb{E}_N[\psi(N)^2].$$

Next, exactly as in (49)–(50) but with $k_X$ replaced by $k_X^2$, we have

$$\mathrm{MMD}_{k_X^2}^2(\mathbb{F}_N^0, \delta_0) = \mathbb{E}_{N,N'}[\psi(N-N')^2] + \kappa_X^2 - 2\,\mathbb{E}_N[\psi(N)^2],$$

Using the same Cauchy-Schwarz argument as before, we have

$$\mathbb{E} \|\Phi_X^{(2)}(X) - \Phi_X^{(2)}(W)\|^2 \leqslant 2\kappa_X \,\mathrm{MMD}_{k_X^2}(\mathbb{F}_N^0, \delta_0),$$

and hence

$$\mathrm{MMD}_{k_X^2}(\mathbb{P}_X^0, \mathbb{P}_W^0) \leqslant \sqrt{\mathbb{E} \|\Phi_X^{(2)}(X) - \Phi_X^{(2)}(W)\|^2} \leqslant \sqrt{2\kappa_X \,\mathrm{MMD}_{k_X^2}(\mathbb{F}_N^0, \delta_0)}. \tag{54}$$

Finally, combining (54) and (52) gives completes the proof for (13). □

### *B.6. Proof of Theorem 3*

Before proving Theorem 3, we first state and prove the following lemma, which is an application of the classical concentration theory for martingales in Banach spaces by Pinelis [1994].

**Lemma 6** (Conditional & unconditional Bernstein bound in Hilbert space)**.** *Let $\mathcal{H}$ be a real separable Hilbert space. Let $V_1, \ldots, V_n : \Omega \to \mathcal{H}$ satisfy*

*(i) $V_1, \ldots, V_n$ are* independent *under* $\Pr_{\mathrm{DP}}(\cdot \mid \mathcal{D}, S)$*;*

*(ii) $\mathbb{E}_{\mathrm{DP}}[V_i \mid \mathcal{D}, S] = 0$ for every $i$;*

*(iii) $\|V_i\|_{\mathcal{H}} \leqslant B$ for a deterministic constant $B < \infty$.*

*Then for every $\varepsilon > 0$*

$$\Pr_{\mathrm{DP}}\Big( \Big\| \frac{1}{n} \sum_{i=1}^n V_i \Big\|_{\mathcal{H}} > \varepsilon \;\Big|\; \mathcal{D}, S \Big) \leqslant 2 \exp\Big( -\frac{n\varepsilon^2}{4B^2} \Big). \tag{55}$$

*A simple consequence is that the same exponential bound holds under the* full *law:*

$$\Pr\Big(\Big\|\frac{1}{n}\sum_{i=1}^{n} V_i\Big\|_{\mathcal{H}} > \varepsilon\Big) \leqslant 2\exp\Big(-\frac{n\varepsilon^2}{4B^2}\Big), \tag{56}$$

*which means that*

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} V_i\Big\|_{\mathcal{H}} \xrightarrow{\Pr} 0.$$

*Proof.* Conditioning on $(\mathcal{D}, S)$, we define a $\mathcal{H}$-valued martingale by

$$f_j = \begin{cases} 0, & j = 0 \\ \frac{1}{n}\sum_{i=1}^{j} V_i, & 1 \leqslant j \leqslant n \\ \frac{1}{n}\sum_{i=1}^{n} V_i, & j > n. \end{cases} \tag{57}$$

Let $\mathcal{F}_j = \sigma(V_1, \ldots, V_j, \mathcal{D}, S)$, then the sequence $\{f_j, \mathcal{F}_j\}_{j=1}^{\infty}$ forms a martingale under $\Pr_{\mathrm{DP}}(\cdot \mid \mathcal{D}, S)$ by the independence of $V_i$. Its differences are $d_j := f_j - f_{j-1} = n^{-1}V_j$, which satisfy $\|d_j\|_{\mathcal{H}} \leqslant B/n$ for $j \leqslant n$ and $\|d_j\|_{\mathcal{H}} \equiv 0$ for $j > n$. Let $f^* := \sup_{j\geqslant 0}\|f_j\|_{\mathcal{H}}$, we note that $f^* \geqslant \|f_n\|_{\mathcal{H}} = \big\|\frac{1}{n}\sum_{i=1}^{n} V_i\big\|_{\mathcal{H}}$ almost surely. Because $\mathcal{H}$ is a Hilbert space, it is a $(2,1)$–smooth Banach space. Here $(2,1)$–smooth means that the parallelogram identity holds: for every $x, y \in \mathcal{H}$, we have $\|x+y\|_{\mathcal{H}}^2 + \|x-y\|_{\mathcal{H}}^2 \leqslant 2\|x\|_{\mathcal{H}}^2 + 2\|y\|_{\mathcal{H}}^2$. Applying Pinelis [1994, Theorem 3.1] gives

$$\begin{aligned} \Pr_{\mathrm{DP}}\big(f^* \geqslant \varepsilon \mid \mathcal{D}, S\big) &\leqslant 2\exp\left\{-\lambda\varepsilon + \left\|\sum_{j=1}^{\infty}\mathbb{E}_{j-1}\left(\exp\left(\lambda\|d_j\|_{\mathcal{H}}\right) - 1 - \lambda\|d_j\|_{\mathcal{H}}\right)\right\|_{\infty}\right\} \\ &= 2\exp\left\{-\lambda\varepsilon + \left\|\sum_{j=1}^{n}\mathbb{E}_{j-1}\left(\exp\left(\lambda\|d_j\|_{\mathcal{H}}\right) - 1 - \lambda\|d_j\|_{\mathcal{H}}\right)\right\|_{\infty}\right\}, \end{aligned} \tag{58}$$

where $\|\cdot\|_{\infty}$ represents the essential supremum of the enclosed random variable, and it is required that $\mathbb{E}\big[e^{\lambda\|d_j\|_{\mathcal{H}}}\big] < \infty$.

Fix $\varepsilon > 0$ with $\varepsilon < 2B$. We set

$$\lambda := \frac{n\varepsilon}{2B^2} \quad \Longrightarrow \quad \lambda\|d_j\|_{\mathcal{H}} \leqslant \frac{\varepsilon}{2B} < 1 \quad \text{for all } j,$$

so the exponential moments $\mathbb{E}\big[e^{\lambda\|d_j\|_{\mathcal{H}}}\big] < \infty$.

Using $\exp(u) - 1 - u \leqslant u^2$ for $0 \leqslant u < 1$ and $\|d_j\|_{\mathcal{H}} \leqslant B/n$, we have

$$\sum_{j=1}^{n}\mathbb{E}_{j-1}\big[\exp(\lambda\|d_j\|_{\mathcal{H}}) - 1 - \lambda\|d_j\|_{\mathcal{H}}\big] \leqslant \lambda^2\sum_{j=1}^{n}\mathbb{E}_{j-1}\|d_j\|_{\mathcal{H}}^2 \leqslant \lambda^2 n\Big(\frac{B}{n}\Big)^2 = \frac{\lambda^2 B^2}{n}.$$

The RHS is deterministic, so taking the essential supremum $\|\cdot\|_{\infty}$ does not change the bound. Since

$$\exp\left\{-\lambda\varepsilon + \frac{\lambda^2 B^2}{n}\right\} = \exp\left\{-\frac{n\varepsilon^2}{2B^2} + \frac{1}{n}\Big(\frac{n\varepsilon}{2B^2}\Big)^2 B^2\right\} = \exp\left(-\frac{n\varepsilon^2}{4B^2}\right),$$

we have

$$\Pr_{\mathrm{DP}}\Big(\Big\|\frac{1}{n}\sum_{i=1}^{n} V_i\Big\|_{\mathcal{H}} > \varepsilon \;\Big|\; \mathcal{D}, S\Big) \leqslant \Pr_{\mathrm{DP}}\big(f^* \geqslant \varepsilon \mid \mathcal{D}, S\big) \leqslant 2\exp\Big(-\frac{n\varepsilon^2}{4B^2}\Big).$$

This is exactly (55). Since the RHS is deterministic once $\varepsilon$ is chosen, taking the expectation with respect to $\Pr_{\mathcal{D},S}$ preserves the right–hand side, yielding (56). □

We are now ready to prove Theorem 3.

*Proof of Theorem 3. Step (a)*

We first recall the following notations:

$$\mathcal{P}_{XY}^{\mathrm{DP}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{X,Y|W_i}^{\mathrm{DP}}, \qquad \mathcal{P}_{X}^{\mathrm{DP}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{X|W_i}^{\mathrm{DP}};$$

$$\mathbb{P}_{XY,i}^{\mathrm{base}} = \frac{c}{c+m}\,\mathbb{Q}_{XY,i} + \frac{1}{c+m}\sum_{k=1}^{m} \delta_{\left(\tilde{X}_{ij}, Y_i\right)}, \qquad \mathbb{P}_{X,i}^{\mathrm{base}} = \frac{c}{c+m}\,\mathbb{Q}_{X,i} + \frac{1}{c+m}\sum_{k=1}^{m} \delta_{\tilde{X}_{ij}},$$

Then $\mathbb{P}_{XY,i}^{\mathrm{base}} = \mathbb{E}_{\mathrm{DP},i}\left[\mathbb{P}_{X,Y|W_i}^{\mathrm{DP}} \mid \mathcal{D}, S\right]$, and $\mathbb{P}_{X,i}^{\mathrm{base}} = \mathbb{E}_{\mathrm{DP},i}\left[\mathbb{P}_{X|W_i}^{\mathrm{DP}} \mid \mathcal{D}, S\right]$. Furthermore, we define

$$\mathbb{P}_{XY}^{\mathrm{base}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{XY,i}^{\mathrm{base}}, \qquad \mathbb{P}_{X}^{\mathrm{base}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{X,i}^{\mathrm{base}}\,.$$

and the MMD quantities

$$M_n(\theta) := \mathrm{MMD}_k\big(\mathcal{P}_{XY}^{\mathrm{DP}}, \mathcal{P}_{X}^{\mathrm{DP}}\,\mathbb{P}_{g(X,\theta)}\big), \tag{59}$$

$$\tilde{M}_n(\theta) := \mathrm{MMD}_k\big(\mathcal{P}_{XY}^{\mathrm{DP}}, \mathbb{P}_{X}^{\mathrm{base}}\,\mathbb{P}_{g(X,\theta)}\big), \tag{60}$$

$$M_n^*(\theta) := \mathrm{MMD}_k\big(\mathbb{P}_{XY}^{\mathrm{base}}, \mathbb{P}_{X}^{\mathrm{base}}\,\mathbb{P}_{g(X,\theta)}\big). \tag{61}$$

By triangle inequality and [Alquier and Gerber, 2024, Lemma 2], we have, for all $\theta \in \Theta$:

$$\left|M_n(\theta) - \tilde{M}_n(\theta)\right| \leqslant \mathrm{MMD}_k(\mathcal{P}_{X}^{\mathrm{DP}}\,\mathbb{P}_{g(X,\theta)}, \mathbb{P}_{X}^{\mathrm{base}}\,\mathbb{P}_{g(X,\theta)}) \leqslant \Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathcal{P}_{X}^{\mathrm{DP}}, \mathbb{P}_{X}^{\mathrm{base}}\right)$$

The RHS doesn't depend on $\theta$, so

$$\sup_{\theta\in\Theta}\left|M_n(\theta) - \tilde{M}_n(\theta)\right| \leqslant \Lambda\,\mathrm{MMD}_{k_X^2}\left(\mathcal{P}_{X}^{\mathrm{DP}}, \mathbb{P}_{X}^{\mathrm{base}}\right)$$

Let $V_i := \varphi_{k_X^2}(\mathbb{P}_{X|W_i}^{\mathrm{DP}}) - \varphi_{k_X^2}(\mathbb{P}_{X,i}^{\mathrm{base}})$, then

$$\mathrm{MMD}_{k_X^2}\left(\mathcal{P}_{X}^{\mathrm{DP}}, \mathbb{P}_{X}^{\mathrm{base}}\right) = \left\|\varphi\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{X|W_i}^{\mathrm{DP}}\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{X,i}^{\mathrm{base}}\right)\right\|_{\mathcal{H}_{k_X^2}} = \left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\|_{\mathcal{H}_{k_X^2}}$$

Since $V_1, \ldots, V_n$ are independent conditional on $(\mathcal{D}, S)$, $\mathbb{E}_{\mathrm{DP}}[V_i \mid \mathcal{D}, S] = 0$, and $\|V_i\|_{\mathcal{H}_{k_X^2}} \leqslant 2\kappa_X$, we get, by Lemma 6:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\|_{\mathcal{H}} \xrightarrow{\Pr} 0.$$

Therefore,

$$\sup_{\theta\in\Theta}\left|M_n(\theta)-\tilde{M}_n(\theta)\right|\xrightarrow{\Pr}0. \tag{62}$$

Again, by triangle inequality, we have

$$\sup_{\theta\in\Theta}\left|\tilde{M}_n(\theta)-M_n^*(\theta)\right|\leqslant \mathrm{MMD}_k(\mathcal{P}_{XY}^{\mathrm{DP}},\mathbb{P}_{XY}^{\mathrm{base}})$$

Similarly, we let $U_i := \varphi_k(\mathbb{P}^{\mathrm{DP}}_{X,Y|W_i}) - \varphi_k(\mathbb{P}^{\mathrm{base}}_{XY,i})$ and apply Lemma 6 with $\mathbb{E}_{\mathrm{DP}}[U_i \mid \mathcal{D}, S] = 0$ and $\|U_i\|_{\mathcal{H}_k} \leqslant 2\sqrt{\kappa}$, we have

$$\sup_{\theta\in\Theta}\left|\tilde{M}_n(\theta)-M_n^*(\theta)\right|\xrightarrow{\Pr}0. \tag{63}$$

Since

$$\sup_{\theta\in\Theta}|M_n(\theta)-M_n^*(\theta)|\leqslant\sup_{\theta\in\Theta}\left|M_n(\theta)-\tilde{M}_n(\theta)\right|+\sup_{\theta\in\Theta}\left|\tilde{M}_n(\theta)-M_n^*(\theta)\right|,$$

We have

$$\sup_{\theta\in\Theta}|M_n(\theta)-M_n^*(\theta)|\xrightarrow{\Pr}0. \tag{64}$$

*Step (b)* By the same argument as that in step (a), we have

$$\sup_{\theta\in\Theta}|M_n^*(\theta)-M(\theta)|\leqslant\Lambda\,\mathrm{MMD}_{k_X^2}(\mathbb{P}_X^{\mathrm{base}},\mathbb{P}_X^{\infty})+\mathrm{MMD}_k(\mathbb{P}_{X,Y}^{\mathrm{base}},\mathbb{P}_{X,Y}^{\infty}) \tag{65}$$

By condition 14, the RHS converges to zero in $\Pr_{\mathcal{D},S}$-probability, so the LHS also converges to zero in $\Pr_{\mathcal{D},S}$-probability (note that neither $M_n^*$ and $M$ depend on the random DP realisations under $\Pr_{\mathrm{DP}}$). Combining (63) and (65) and using the triangle inequality, we have

$$\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{\Pr}0. \tag{66}$$

by [Newey and McFadden, 1994, Theorem 2.1], we have consistency $\hat{\theta}_n \xrightarrow{\Pr} \theta^\dagger$. □

### *B.7. Proof of Proposition 3*

*Proof of Proposition 3.* We first state the following inequality on the MMD. For any $0 \leqslant \alpha, \beta \leqslant 1$ with $\alpha+\beta=1$ and probability measures $P_1, P_2, Q_1, Q_2$:

$$\begin{aligned}\mathrm{MMD}_k(\alpha P_1+\beta P_2,\alpha Q_1+\beta Q_2) &= \|\varphi(\alpha P_1+\beta P_2)-\varphi(\alpha Q_1+\beta Q_2)\|_{\mathcal{H}_k}\\ &= \|\alpha\left[\varphi(P_1)-\varphi(Q_1)\right]+\beta\left[\varphi(P_2)-\varphi(Q_2)\right]\|_{\mathcal{H}_k}\\ &\leqslant \alpha\|\varphi(P_1)-\varphi(Q_1)\|_{\mathcal{H}_k}+\beta\|\varphi(P_2)-\varphi(Q_2)\|_{\mathcal{H}_k}\\ &= \alpha\,\mathrm{MMD}_k(P_1,Q_1)+\beta\,\mathrm{MMD}_k(P_2,Q_2)\end{aligned} \tag{67}$$

Now we show (3.a).

For equation (16) with non-vanishing prior effect, we recall

$$\mathbb{P}_{XY}^{\infty}=\frac{c}{c+m}\,\mathbb{Q}_{XY}^{\infty}+\frac{m}{c+m}\Pi_{\theta^*}^{XY},\qquad \mathbb{P}_{X}^{\infty}=\frac{c}{c+m}\,\mathbb{Q}_{X}^{\infty}+\frac{m}{c+m}\Pi_{\theta^*}^{X}.$$

By (67) we can decompose

$$\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{base}}, \mathbb{P}_{XY}^{\infty}) \leqslant \frac{c}{c+m}\,\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{Q}_{XY,i}, \mathbb{Q}_{XY}^{\infty}\right) + \frac{m}{c+m}\,\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right) \tag{68}$$

The first term converges to zero in probability by condition (D). For the second term, we recall in the proof of Theorem 2 for the classical ME model (or Proposition 2 for Berkson ME) that, by term A (36) and term B (43), we have, for any $M_n \to \infty$,

$$\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right)\right] \leqslant \frac{4\kappa}{\sqrt{n}} + \frac{M_n}{\sqrt{n}} + 2\kappa r_n \tag{69}$$

where $r_n \to 0$ as $n \to \infty$. Taking $M_n = \log n$, we have $\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right)\right] \to 0$ as $n \to \infty$. By Markov's inequality, we have

$$\Pr\left(\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right) > \varepsilon\right) \leqslant \frac{\mathbb{E}_{\mathcal{D},S}\left[\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right)\right]}{\varepsilon}. \tag{70}$$

The RHS converges to zero as $n \to \infty$, which proves $\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right) \xrightarrow{\Pr} 0$. Substituting this into (68) gives

$$\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{base}}, \mathbb{P}_{XY}^{\infty}\right) \xrightarrow{\Pr} 0 \qquad \text{as } n \to \infty.$$

The same argument applied to $\mathbb{P}_{\mathrm{base}}^{X}$ and $\Pi_{\theta*}^{X}$ combined with Proposition 1 (Classical) or Proposition 2 (Berkson) shows

$$\mathrm{MMD}_{k_X^2}\left(\mathbb{P}_{X}^{\mathrm{base}}, \mathbb{P}_{X}^{\infty}\right) \xrightarrow{\Pr} 0.$$

When $c/m \to 0$ as $n \to \infty$, we can decompose

$$\begin{aligned}\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{base}}, \Pi_{\theta*}^{XY}) &\leqslant \frac{c}{c+m}\,\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{Q}_{XY,i}, \Pi_{\theta*}^{XY}\right) + \frac{m}{c+m}\,\mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right) \\ &\leqslant \underbrace{\frac{2\kappa c}{c+m}}_{\to 0 \text{ as } m/c \to \infty} + \mathrm{MMD}_k\left(\mathbb{P}_{XY}^{\mathrm{pseudo}}, \Pi_{\theta*}^{XY}\right)\end{aligned} \tag{71}$$

The convergence of the second term gives $\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{base}}, \Pi_{\theta*}^{XY}) \xrightarrow{\Pr} 0$. Similarly, we have $\mathrm{MMD}_k(\mathbb{P}_{X}^{\mathrm{base}}, \Pi_{\theta*}) \xrightarrow{\Pr} 0$.

Next, we show (3.b). Since $(W_i, Y_i) \overset{i.i.d.}{\sim} \mathbb{P}_{WY}^{0}$, we have, by [Alquier and Gerber, 2024, Lemma S6],

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{(W_i,Y_i)}, \mathbb{P}_{WY}^{0}\right)\right] \leqslant \frac{\sqrt{\kappa}}{\sqrt{n}}. \tag{72}$$

Again, by Markov's inequality,

$$\mathrm{MMD}_k\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{(W_i,Y_i)}, \mathbb{P}_{WY}^{0}\right) \xrightarrow{\Pr} 0 \tag{73}$$

By the same decompositions as the ones in the proof of (3.a), we get that (17) follows from (D) and (73); The final statement follows from $c \to 0$ and (73).

□

## C. Sufficient conditions for Assumptions A1- A2 and verifying them in two scenarios

Throughout, $\|\cdot\|$ denotes the Euclidean norm, and for a $p \times p$ matrix $A$ we write $\|A\|$ for the operator norm.

### *C.1. Sufficient conditions for Assumptions A1- A2*

Recall that the true data-generating process (DGP) under classical ME is

$$W = X + N, \quad Y = g_0(X) + E, \quad X \sim P_X^0, \quad N \sim F_N^0, \quad E \sim F_E^0, \quad N, E \perp\!\!\!\perp X,\ N \perp\!\!\!\perp E. \tag{74}$$

The joint density of the observed pair $(W, Y)$ is $p_{WY}^0(w, y) = \int_{\mathcal{X}} p_X^0(x) f_N^0(w - x) f_E^0\big(y - g_0(x)\big)\mathrm{d}x$.

We work under a misspecified model

$$W = X + N, \quad Y = g(X, \theta) + E, \quad X \sim P_X, \quad N \sim F_N, \quad E \sim F_E, \quad N, E \perp\!\!\!\perp X,\ N \perp\!\!\!\perp E,$$

and $g_0(\cdot) \notin \{g(\cdot, \theta) : \theta \in \Theta\}$; here $\Theta \subset \mathbb{R}^d$ is compact. For each $\theta$ the induced density of $(W, Y)$ is $p_{WY}^\theta(w, y) = \int_{\mathcal{X}} p_X(x) f_N(w - x) f_E\big(y - g(x, \theta)\big)\mathrm{d}x$.

Define the pseudo-true parameter by $\theta^* := \arg\min_{\theta \in \Theta} \mathrm{KL}\big(p_{WY}^0 \| p_{WY}^\theta\big)$.

Write $\ell_\theta(W, Y) := \log p_{WY}^\theta(W, Y)$. For i.i.d. $(W_i, Y_i) \sim p_{WY}^0$ we collect and list sufficient conditions for assumptions A1-A2:

**Con.1** *Likelihood–ratio integrability:* $\mathbb{E}_{p_{WY}^0}\left[p_{WY}^\theta / p_{WY}^{\theta^*}\right] < \infty$ for all $\theta \in \Theta$.

**Con.2** *Differentiability in probability:* $\theta \mapsto \ell_\theta(W_1, Y_1)$ differentiable at $\theta^*$ in $p_{WY}^0$–probability.

**Con.3** *Local Lipschitz envelope:* there exists $m_{\theta^*} \in L^2(p_{WY}^0)$ such that for $\theta_1, \theta_2$ near $\theta^*$, $\left|\ell_{\theta_1} - \ell_{\theta_2}\right| \leqslant m_{\theta^*}\|\theta_1 - \theta_2\|$.

**Con.4** *Quadratic KL expansion:* $-\mathbb{E}_{p_{WY}^0}\left[\log(p_{WY}^\theta / p_{WY}^{\theta^*})\right] = \frac{1}{2}(\theta - \theta^*)^\top V_{\theta^*}(\theta - \theta^*) + o(\|\theta - \theta^*\|^2)$ with $V_{\theta^*} \succ 0$.

**Con.5** $L^1(p_{WY}^0)$ *continuity:* for every fixed $\theta_0 \in \Theta$, the map $\theta \mapsto p_{WY}^\theta / p_{WY}^{\theta_0}$ is $L^1(p_{WY}^0)$–continuous at every $\theta \in \Theta$.

**Con.6** *Exponential moment of the envelope:* $\exists s > 0$ with $\mathbb{E}_{p_{WY}^0}[e^{s m_{\theta^*}}] < \infty$.

**Con.7** *Non–singular score covariance:* $S_{\theta^*} := \mathbb{E}_{p_{WY}^0}\left[\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^\top\right]$ is invertible.

**Con.8** *Compact* $\Theta$ *and uniqueness:* $\Theta$ compact and $\theta^* \in \mathrm{int}(\Theta)$ the unique minimiser of $\theta \mapsto -\mathbb{E}_{p_{WY}^0} \log p_{WY}^\theta$.

**Con.9** The prior $\Pi$ on $\Theta$ admits a density $\pi$ with respect to Lebesgue measure that is continuous and strictly positive on a neighbourhood of $\theta^*$.

**Con.10** *Posterior Lipschitz in MMD:*

$$\mathrm{MMD}_k\left(\Pi_{\theta_1}(\cdot \mid w,y) - \Pi_{\theta_2}(\cdot \mid w,y)\right) \leqslant L(w,y)\,\|\theta_1 - \theta_2\|, \qquad L(W,Y) \in L^2(p^0_{WY}),$$

for $\theta_1, \theta_2$ in a neighbourhood $\Theta_\rho$ of $\theta^*$.

Assumption A1 invokes the LAN, smoothness, integrability and regularity requirements of [Kleijn and Van der Vaart, 2012, Theorem 3.1] for $\{p^\theta_{WY} : \theta \in \Theta\}$ around $\theta^*$. We now explain why the conditions verified above are sufficient:

(a) *LAN via Lemma 2.1.* Our **Con.2** (differentiability in probability), **Con.3** (local Lipschitz envelope) and **Con.4** (quadratic KL expansion with $V_{\theta^*} \succ 0$) yield LAN with $\delta_n = n^{-1/2}$ and central sequence $V^{-1}_{\theta^*}\mathbb{G}_n\dot{\ell}_{\theta^*}$.

(b) *Existence of tests via Theorem 3.2.* Compactness and uniqueness (**Con.8**) together with the $L^1$–continuity of likelihood ratios (**Con.5**) satisfy the sufficient conditions in [Kleijn and Van der Vaart, 2012, Theorem 3.2], guaranteeing tests $(\phi_n)$ with the properties required in [Kleijn and Van der Vaart, 2012, Theorem 3.1].

(c) *Moment and prior conditions for Theorem 3.1.* Our **Con.1** ensures $\mathbb{E}_{p^0_{WY}}[p_\theta/p_{\theta^*}] < \infty$ for all $\theta \in \Theta$. Condition **Con.6** provides $\mathbb{E}_{p^0_{WY}}[e^{s m_{\theta^*}}] < \infty$ for some $s > 0$. Condition **Con.9** supplies a prior density continuous and strictly positive near $\theta^*$. Invertibility of $\mathbb{E}_{p^0_{WY}}[\dot{\ell}_{\theta^*}\dot{\ell}^\top_{\theta^*}]$ is **Con.7**.

Consequently, Theorem 3.1 of Kleijn and Van der Vaart [2012] applies under conditions **Con.1**-**Con.9**. Condition **Con.10** implies assumption A2. It is worth noting that **Con.1**-**Con.9** are a compilation of *sufficient* conditions for A1, and many of them can be relaxed depending on the specific model and true distribution. We refer the reader to Kleijn and Van der Vaart [2012] for a detailed discussion on weaker forms of these conditions and circumstances where they can be relaxed.

### *C.2. Two scenarios where Assumptions A1- A2 hold*

Next, we demonstrate two scenarios under Gaussian noise models where conditions **Con.1**-**Con.10** conditions are satisfied. In this section, ME and outcome noise are modelled as independent centred Gaussians $N \sim \mathcal{N}_d(0, \Sigma_N), E \sim \mathcal{N}(0, \sigma^2_E)$, with known $\Sigma_N \succ 0$, $\sigma^2_E > 0$. A working prior density for $X$ is $p_X$. Write

$$\varphi_\Sigma(z) := (2\pi)^{-d/2}(\det\Sigma)^{-1/2}e^{-\frac{1}{2}z^\top\Sigma^{-1}z}, \qquad \varphi_\sigma(t) := (2\pi\sigma^2)^{-1/2}e^{-t^2/(2\sigma^2)}.$$

Then

$$p^\theta_{WY}(w,y) = \int_{\mathbb{R}^d} p_X(x)\varphi_{\Sigma_N}(w-x)\varphi_{\sigma_E}\big(y - g(x,\theta)\big)dx. \tag{75}$$

*Global assumptions.* We collect here the standing assumptions used throughout. They are invoked in both scenarios below.

(M1) *Local smoothness and curvature near $\theta^*$:* there exists a neighbourhood $U$ of $\theta^*$ such that $g(\cdot,\theta)$ is $C^2$ in $\theta$ on $U$; for $\theta \in U$, $\dot{\ell}_\theta$ and $\ddot{\ell}_\theta$ admit an $L^1(p^0_{WY})$ envelope uniform in $\theta \in U$; and $H(\theta) := -\mathbb{E}_{p^0_{WY}}\left[\partial^2_\theta \ell_\theta(W,Y)\right]$ exists for $\theta \in U$, is continuous at $\theta^*$, and $H^* := H(\theta^*) \succ 0$.

(M2) *Outcome tail:* $\mathbb{E}\left[e^{\tau_0 Y^2}\right] < \infty$ for some $\tau_0 > 0$.

(M3) *Information non–singularity at $\theta^*$:* $S_{\theta*} := \mathbb{E}_{p^0_{WY}}\left[\dot{\ell}_{\theta*}\dot{\ell}_{\theta*}^\top\right]$ is invertible. This is **Con.7**.

(M4) *Compactness and uniqueness:* $\Theta$ is compact and $\theta^* \in \mathrm{int}(\Theta)$ is the unique minimiser of $R(\theta) = -\mathbb{E}_{p^0_{WY}}[\ell_\theta(W,Y)]$. This is **Con.8**.

(M5) *Continuous and positive prior:* The prior $\Pi$ on $\Theta$ admits a density $\pi$ with respect to Lebesgue measure that is continuous and strictly positive on a neighbourhood of $\theta^*$: this is **Con.9**.

For smooth finite-dimensional nonlinear regressions, these requirements are mild: (M1) is a local $C^2$ and integrability condition that justifies differentiation under the integral and yields positive-definite local curvature $H^*$.(M2) is used for algebraic convenience to control exponential envelopes; it can be weakened to a sub-exponential tail (e.g. $\mathbb{E}[e^{\tau|Y|}] < \infty$) while adjusting the envelope arguments. (M3) is standard: finiteness of $S_{\theta*}$ follows in both scenarios below from the score envelopes and tail conditions, so it effectively asks only for non–degeneracy of the score covariance at $\theta^*$. (M4) is the standard well-separated pseudo-true parameter assumption. The prior condition (M5) is the usual prior thickness requirement.

We consider two concrete scenarios that ensure all conditions **Con.1**-**Con.10** are satisfied. In both cases we assume (M1)–(M5).

*Scenario 1 (bounded regression).* This covers models where the regression surface and its first two derivatives are uniformly bounded over $\Theta \times \mathbb{R}^d$.

(S1.1) $C_g := \sup_{\theta\in\Theta, x\in\mathbb{R}^d} |g(x,\theta)| < \infty$.

(S1.2) There exist finite constants $C_{\partial g}, C_{\partial^2 g}$ such that for all $(\theta, x) \in \Theta \times \mathcal{X}$, $\|\partial_\theta g(x,\theta)\| \leqslant C_{\partial g}, \|\partial^2_{\theta\theta} g(x,\theta)\| \leqslant C_{\partial^2 g}$.

*Scenario 2 (compact latent support and working prior).* This covers settings where $X$ is confined to a compact region and the working prior respects that support; boundedness of $g$ and its derivatives is then only needed on that region.

(S2.1) $\mathrm{supp}\, p_X \subseteq B_M := \{x : \|x\| \leqslant M\}$.

(S2.2) $\|X\| \leqslant M$ $P^0_X$–a.s.

(S2.3) $\widetilde{C}_{\partial g} := \sup_{\theta\in\Theta, x\in B_M} \|\partial_\theta g(x,\theta)\| < \infty$ and $\widetilde{C}_{\partial^2 g} := \sup_{\theta\in\Theta, x\in B_M} \|\partial^2_{\theta\theta} g(x,\theta)\| < \infty$.

(S2.4) $g$ is continuous on $\Theta \times B_M$, hence $C_{g,M} := \sup_{\theta\in\Theta, \|x\|\leqslant M} |g(x,\theta)| < \infty$.

### C.3. Verification of **Con.1–Con.10**

We verify **Con.1**–**Con.10** for both scenarios. Since **Con.7**-**Con.9** are already assumed by (M3)-(M5), we only need to verify **Con.1**-**Con.6**, and **Con.10**.

#### C.3.1. **Con.1** Likelihood–ratio integrability

*Goal:* $\mathbb{E}_{p^0_{WY}}\left[p^{\theta}_{WY}/p^{\theta^*}_{WY}\right] < \infty$ for all $\theta \in \Theta$.

For $R_\theta(W,Y) := p^{\theta}_{WY}(W,Y)/p^{\theta^*}_{WY}(W,Y)$ write $a(x) := p_X(x)\varphi_{\Sigma_N}(W-x)$ and $b_\theta(x) := \varphi_{\sigma_E}(Y-g(x,\theta))$. Then

$$R_\theta = \frac{\int a(x)b_\theta(x)dx}{\int a(x)b_{\theta*}(x)dx} \leqslant \sup_x \frac{b_\theta(x)}{b_{\theta*}(x)}.$$

Since $b_\theta(x) = c\exp\{-(Y-g(x,\theta))^2/(2\sigma_E^2)\}$,

$$\log\frac{b_\theta(x)}{b_{\theta*}(x)} = -\frac{(Y-g(x,\theta))^2-(Y-g(\theta^*,x))^2}{2\sigma_E^2} = \frac{(g(x,\theta)-g(\theta^*,x))}{\sigma_E^2}Y + \frac{g(\theta^*,x)^2-g(x,\theta)^2}{2\sigma_E^2}.$$

*Scenario 1.* Using $|g(x,\theta)-g(\theta^*,x)| \leqslant 2C_g$ and $|g(\theta^*,x)^2-g(x,\theta)^2| \leqslant 4C_g^2$,

$$R_\theta \leqslant \exp\Big\{\frac{2C_g}{\sigma_E^2}|Y| + \frac{2C_g^2}{\sigma_E^2}\Big\}.$$

By the bound $e^{a|Y|} \leqslant e^{a^2/(4\varepsilon)}e^{\varepsilon Y^2}$ (valid for all $a,\varepsilon>0$) and (M2) (with any $\varepsilon < \tau_0$), $\mathbb{E}\,R_\theta < \infty$.

*Scenario 2.* Because $\operatorname{supp} p_X \subseteq B_M$ and $\|X\| \leqslant M$ a.s., the integrals in (75) are over $B_M$. With $\Delta_\theta(M) := \sup_{\|x\|\leqslant M}|g(x,\theta)-g(\theta^*,x)|$,

$$R_\theta \leqslant \exp\Big\{\frac{\Delta_\theta(M)}{\sigma_E^2}|Y| + \frac{C_{g,M}\Delta_\theta(M)}{\sigma_E^2}\Big\}.$$

Since $\Delta_\theta(M) \leqslant 2C_{g,M}$ by (S2.4), (M2) implies $\mathbb{E}\,R_\theta < \infty$.

#### C.3.2. **Con.2** Differentiability

*Goal:* $\theta \mapsto \ell_\theta(W,Y)$ differentiable at $\theta^*$ in $p^0_{WY}$–probability.

For $(w,y) \in \mathbb{R}^d \times \mathbb{R}$,

$$p_\theta(w,y) = \int p_X(x)\varphi_{\Sigma_N}(w-x)\varphi_{\sigma_E}\big(y-g(x,\theta)\big)dx.$$

Under (S1.2) or (S2.3) there exists an integrable envelope $\Gamma$ (constant in Scenario 1, bounded on $B_M$ in Scenario 2) such that $\|\partial_\theta g(x,\theta)\| \leqslant \Gamma(x)$ and $\int p_X(x)\Gamma(x)dx < \infty$. Differentiation under the integral (Leibniz rule) gives

$$\nabla_\theta p_\theta(w,y) = \int p_X(x)\varphi_{\Sigma_N}(w-x)\nabla_\theta\varphi_{\sigma_E}\big(y-g(x,\theta)\big)dx,$$

and

$$\nabla_\theta \varphi_{\sigma_E}(y - g(x,\theta)) = \varphi_{\sigma_E}(y - g(x,\theta)) \frac{y - g(x,\theta)}{\sigma_E^2} \nabla_\theta g(x,\theta).$$

Hence

$$\nabla_\theta p_\theta(w,y) = \frac{1}{\sigma_E^2} \int p_X(x) \varphi_{\Sigma_N}(w-x) \varphi_{\sigma_E}(y - g(x,\theta))(y - g(x,\theta)) \nabla_\theta g(x,\theta) dx.$$

Introduce the $\theta$–posterior density

$$q_\theta(x \mid w,y) := \frac{p_X(x)\varphi_{\Sigma_N}(w-x)\varphi_{\sigma_E}(y - g(x,\theta))}{p_\theta(w,y)}.$$

Dividing by $p_\theta(w,y)$ yields the score

$$\dot{\ell}_\theta(W,Y) = \frac{1}{\sigma_E^2} \mathbb{E}_{q_\theta(\cdot|W,Y)} \Big[ (Y - g(X,\theta)) \partial_\theta g(X,\theta) \mid W, Y \Big]. \tag{76}$$

Let $f_\theta(x) := p_X(x)\varphi_{\Sigma_N}(W-x)\varphi_{\sigma_E}(Y - g(x,\theta))$, so $\ell_\theta(W,Y) = \log \int f_\theta(x) dx$. For $h \in \mathbb{R}^p$ and $\theta_t := \theta + th$,

$$\ell_{\theta+h} - \ell_\theta = \int_0^1 \frac{\int \langle \partial_\theta f_{\theta_t}(x), h \rangle \mathrm{d}x}{\int f_{\theta_t}(x) \mathrm{d}x} dt = \int_0^1 \langle \dot{\ell}_{\theta_t}, h \rangle \mathrm{d}t.$$

A second differentiation gives

$$\begin{aligned} \ddot{\ell}_\theta(W,Y) &:= \partial^2_{\theta\theta} \ell_\theta(W,Y) \\ &= \mathbb{E}_{q_\theta(\cdot|W,Y)} \Big[ -\tfrac{1}{\sigma_E^2} \partial_\theta g \partial_\theta g^\top + \tfrac{Y-g}{\sigma_E^2} \partial^2_{\theta\theta} g \;\Big|\; W, Y \Big] + \mathrm{Var}_{q_\theta(\cdot|W,Y)} \Big( \tfrac{Y-g}{\sigma_E^2} \partial_\theta g \;\Big|\; W, Y \Big). \end{aligned} \tag{77}$$

Therefore

$$\ell_{\theta+h} - \ell_\theta = \langle \dot{\ell}_\theta, h \rangle + \int_0^1 (1-t) \langle \ddot{\ell}_{\theta_t} h, h \rangle \mathrm{d}t,$$

and hence

$$\frac{|\ell_{\theta+h} - \ell_\theta - \langle \dot{\ell}_\theta, h \rangle|}{\|h\|} \leqslant \frac{\|h\|}{2} \sup_{t \in [0,1]} \|\ddot{\ell}_{\theta_t}(W,Y)\|. \tag{78}$$

From (77) and (S1.2) (Scenario 1) or (S2.3)–(S2.4) (Scenario 2), there exist finite $B_0, B_1, B_2$ (scenario–dependent, $\theta$–uniform) such that

$$\|\ddot{\ell}_\theta(W,Y)\| \leqslant \frac{B_0 + B_1|Y|}{\sigma_E^2} + \frac{B_2(|Y| + C)^2}{\sigma_E^4}, \qquad C = C_g \text{ or } C_{g,M}.$$

Thus $\|\ddot{\ell}_\theta(W,Y)\| \leqslant a_0 + a_1|Y| + a_2 Y^2$ for some finite $(a_0, a_1, a_2)$, which is integrable by (M2). Denote $M(W,Y) := a_0 + a_1|Y| + a_2 Y^2 \in L^1(p^0_{WY})$.

Fix $\theta = \theta^*$. For every $\varepsilon > 0$,

$$p^0_{WY} \Big( \frac{|\ell_{\theta^*+h} - \ell_{\theta^*} - \langle \dot{\ell}_{\theta^*}, h \rangle|}{\|h\|} > \varepsilon \Big) \leqslant p^0_{WY} \big( \tfrac{\|h\|}{2} M(W,Y) > \varepsilon \big) \xrightarrow[h \to 0]{} 0.$$

Hence **Con.2** holds. The bounds also give $\dot{\ell}_{\theta^*} \in L^2(p^0_{WY})$, used below. The same domination shows differentiability in $p^0_{WY}$–probability holds uniformly along compact line segments in $\Theta$, as used in Subsec. C.3.5.

### *C.3.3.* ***Con.3*** *Local Lipschitz envelope*

By the fundamental theorem of calculus,

$$\ell_{\theta_1} - \ell_{\theta_2} = \int_0^1 \big\langle \dot{\ell}_{\theta_t}, \theta_1 - \theta_2 \big\rangle \mathrm{d}t$$

and hence

$$|\ell_{\theta_1} - \ell_{\theta_2}| \leqslant \|\theta_1 - \theta_2\| \sup_{t\in[0,1]} \|\dot{\ell}_{\theta_t}\|.$$

From (76) and (S1.2) (Scenario 1),

$$\|\dot{\ell}_\theta\| \leqslant \frac{C_{\partial g}}{\sigma_E^2}(|Y| + C_g) =: m_{\theta*}^{(1)}(W,Y),$$

and from (S2.3)–(S2.4) (Scenario 2),

$$\|\dot{\ell}_\theta\| \leqslant \frac{\widetilde{C}_{\partial g}}{\sigma_E^2}(|Y| + C_{g,M}) =: m_{\theta*}^{(2)}(W,Y).$$

By (M2), $m_{\theta*}^{(j)} \in L^2(p^0_{WY})$ $(j = 1, 2)$, proving **Con.3**.

### *C.3.4.* ***Con.4*** *Curvature of the KL risk*

Let $H^* := -\mathbb{E}_{p^0_{WY}}[\ddot{\ell}_{\theta*}]$, which is finite by the moment bounds used in C.3.2. Differentiation under the integral (per (M1)) yields

$$\nabla R(\theta) = -\mathbb{E}_{p^0_{WY}}[\dot{\ell}_\theta(W,Y)].$$

Since $\theta^*$ minimises $R$ and $\theta^* \in \mathrm{int}(\Theta)$, the first-order condition gives $\mathbb{E}_{p^0_{WY}}[\dot{\ell}_{\theta*}] = 0$. A Taylor expansion of $R$ at $\theta^*$ then implies

$$-\mathbb{E}_{p^0_{WY}}\big[\ell_\theta - \ell_{\theta*}\big] = \tfrac{1}{2}(\theta - \theta^*)^\top H^*(\theta - \theta^*) + o(\|\theta - \theta^*\|^2),$$

so $V_{\theta*} = H^*$. By (M1), $H^* \succ 0$, proving **Con.4**.

### *C.3.5.* ***Con.5*** *$L^1$–continuity of the likelihood ratio*

Fix $\theta_0 \in \Theta$ and let $\theta, \theta_1 \in \Theta$ be arbitrary. set $h := \theta - \theta_1$, $\theta_t := \theta_1 + th$. Define

$$r_t(W,Y) := \frac{p^{\theta_t}_{WY}(W,Y)}{p^{\theta_0}_{WY}(W,Y)} = \exp\big\{\ell_{\theta_t}(W,Y) - \ell_{\theta_0}(W,Y)\big\}.$$

Since $t \mapsto r_t$ is absolutely continuous and $\partial_t r_t = r_t \langle h, \dot{\ell}_{\theta_t}\rangle$, we obtain the identity

$$r_{\theta,\theta_0}(W,Y) - r_{\theta_1,\theta_0}(W,Y) = \int_0^1 \langle h, \dot{\ell}_{\theta_t}(W,Y)\rangle\, r_t(W,Y)\mathrm{d}t. \tag{79}$$

Taking absolute values and expectations, by the triangle inequality, Tonelli and Cauchy–Schwarz,

$$\begin{aligned}\|r_{\theta,\theta_0} - r_{\theta_1,\theta_0}\|_{L^1(p^0_{WY})} &\leqslant \|h\| \int_0^1 \mathbb{E}_{p^0_{WY}} \left[ \left\| \dot{\ell}_{\theta_t} \right\| r_t \right] \mathrm{d}t \\ &\leqslant \|h\| \int_0^1 \left( \mathbb{E}_{p^0_{WY}} \left\| \dot{\ell}_{\theta_t} \right\|^2 \right)^{1/2} \left( \mathbb{E}_{p^0_{WY}} r_t^2 \right)^{1/2} \mathrm{d}t. \end{aligned} \tag{80}$$

Using (S1.2) or (S2.3)–(S2.4) together with (M2),

$$M_1 := \sup_{\vartheta \in \Theta} \mathbb{E}_{p^0_{WY}} \|\dot{\ell}_\vartheta\|^2 < \infty.$$

From C.3.1 we have the envelopes

$$r_{\vartheta,\theta_0} \leqslant \exp\Big\{ \frac{2C_g}{\sigma_E^2}|Y| + \frac{2C_g^2}{\sigma_E^2} \Big\} \quad \text{(Scenario 1)}, \; r_{\vartheta,\theta_0} \leqslant \exp\Big\{ \frac{\Delta_{\vartheta,\theta_0}(M)}{\sigma_E^2}|Y| + \frac{C_{g,M}\Delta_{\vartheta,\theta_0}(M)}{\sigma_E^2} \Big\} \quad \text{(Scenario 2)},$$

where $\Delta_{\vartheta,\theta_0}(M) := \sup_{\|x\| \leqslant M} |g(\vartheta, x) - g(\theta_0, x)| \leqslant 2C_{g,M}$. Hence by (M2),

$$M_2 := \sup_{\vartheta \in \Theta} \mathbb{E}\, r^2_{\vartheta,\theta_0} < \infty.$$

Therefore (80) gives

$$\big\| r_{\theta,\theta_0} - r_{\theta_1,\theta_0} \big\|_{L^1(P^0_{WY})} \leqslant \|h\| \sqrt{M_1 M_2} \xrightarrow[\theta \to \theta_1]{} 0,$$

proving **Con.5**.

### C.3.6. ***Con.6*** *Exponential moment*

Recall

$$m_{\theta *} = \begin{cases} \dfrac{C_{\partial g}}{\sigma_E^2}(|Y| + C_g), & \text{Scenario 1}, \\ \dfrac{\widetilde{C}_{\partial g}}{\sigma_E^2}(|Y| + C_{g,M}), & \text{Scenario 2}. \end{cases}$$

In Scenario 1, for any $s > 0$,

$$\mathbb{E}\, e^{s m_{\theta *}} = e^{s C_{\partial g} C_g / \sigma_E^2}\, \mathbb{E} \exp\Big\{ \frac{s C_{\partial g}}{\sigma_E^2}|Y| \Big\} \leqslant e^{s C_{\partial g} C_g / \sigma_E^2} e^{\frac{(s C_{\partial g}/\sigma_E^2)^2}{4\varepsilon}}\, \mathbb{E}\, e^{\varepsilon Y^2} < \infty$$

for any $\varepsilon < \tau_0$ by (M2). Thus **Con.6** holds for all $s > 0$. The same argument applies in Scenario 2:

$$\mathbb{E}\, e^{s m_{\theta *}} \leqslant \exp\Big\{ \frac{s \widetilde{C}_{\partial g} C_{g,M}}{\sigma_E^2} + \frac{(s \widetilde{C}_{\partial g}/\sigma_E^2)^2}{4\varepsilon} \Big\}\, \mathbb{E}\, e^{\varepsilon Y^2} < \infty,$$

for any $s > 0$ and $\varepsilon < \tau_0$. Hence **Con.6** holds.

### C.3.7. ***Con.10*** *Posterior Lipschitz in total variation*

**Lemma 7** (Pathwise total-variation bound)**.** *Fix $(w, y) \in \mathbb{R}^d \times \mathbb{R}$ and define*

$$f_\theta(x) := p_X(x) \varphi_{\Sigma_N}(w - x) \varphi_{\sigma_E}\big(y - g(x, \theta)\big), \quad Z_\theta := \int f_\theta(u) du, \quad \pi_\theta(x) := \frac{f_\theta(x)}{Z_\theta}.$$

*For* $\theta_t := \theta_2 + t(\theta_1 - \theta_2)$,

$$\big\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\big\|_{\mathrm{TV}} \leqslant \|\theta_1 - \theta_2\| \int_0^1 \mathbb{E}_{\Pi_{\theta_t}^{w,y}}\big[|y - g(\theta_t, X)|\|\partial_\theta g(\theta_t, X)\|\big] \frac{dt}{\sigma_E^2}. \tag{81}$$

*Proof.* Total variation is $\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |p - q| dx$. With $\pi_t := \pi_{\theta_t}$,

$$\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\|_{\mathrm{TV}} = \tfrac{1}{2}\int \Big|\int_0^1 \partial_t \pi_t(x) dt\Big| dx \leqslant \tfrac{1}{2}\int_0^1 \int |\partial_t \pi_t(x)| dx dt,$$

by the triangle inequality and Tonelli. Since $\partial_t \pi_t = (\partial_\theta \pi_\theta)|_{\theta=\theta_t}(\theta_1 - \theta_2)$,

$$|\partial_t \pi_t(x)| \leqslant \|\theta_1 - \theta_2\| \|\partial_\theta \pi_{\theta_t}(x)\|.$$

Using $\pi_\theta = f_\theta / Z_\theta$,

$$\partial_\theta \pi_\theta(x) = \pi_\theta(x)\Big(\partial_\theta \log f_\theta(x) - \mathbb{E}_{\Pi_\theta(\cdot|w,y)}[\partial_\theta \log f_\theta(X)]\Big).$$

Hence

$$\int \|\partial_\theta \pi_\theta(x)\| dx \leqslant 2\, \mathbb{E}_{\Pi_\theta(\cdot|w,y)}\big[\|\partial_\theta \log f_\theta(X)\|\big].$$

Cancelling the prefactor $\frac{1}{2}$ gives

$$\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\|_{\mathrm{TV}} \leqslant \|\theta_1 - \theta_2\| \int_0^1 \mathbb{E}_{\Pi_{\theta_t}^{w,y}}\big[\|\partial_\theta \log f_{\theta_t}(X)\|\big] dt.$$

Finally, $\partial_\theta \log f_\theta(x) = \frac{y - g(x,\theta)}{\sigma_E^2} \partial_\theta g(x, \theta)$, which gives (81). □

Let $A_1 := C_{\partial g}/\sigma_E^2$ and $A_2 := \widetilde{C}_{\partial g}/\sigma_E^2$. In Scenario 1,

$$\mathbb{E}_{\Pi_\theta(\cdot|w,y)}\big[|y - g(X, \theta)|\|\partial_\theta g(X, \theta)\|\big] \leqslant C_{\partial g}(|y| + C_g),$$

so

$$\big\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\big\|_{\mathrm{TV}} \leqslant A_1(|y| + C_g)\|\theta_1 - \theta_2\|.$$

In Scenario 2 (on $B_M$),

$$\mathbb{E}_{\Pi_\theta(\cdot|w,y)}\big[|y - g(X, \theta)|\|\partial_\theta g(X, \theta)\|\big] \leqslant \widetilde{C}_{\partial g}(|y| + C_{g,M}),$$

hence

$$\big\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\big\|_{\mathrm{TV}} \leqslant A_2(|y| + C_{g,M})\|\theta_1 - \theta_2\|.$$

By (M2), $L(W, Y) \in L^2(p^0_{WY})$ for

$$L(w, y) := \begin{cases} A_1(|y| + C_g), & \text{Scenario 1,} \\ A_2(|y| + C_{g,M}), & \text{Scenario 2,} \end{cases}$$

By Lemma 3, we have

$$\mathrm{MMD}_k\left(\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\right) \leqslant 2\sqrt{\kappa}\|\Pi_{\theta_1}(\cdot \mid w, y) - \Pi_{\theta_2}(\cdot \mid w, y)\big\|_{\mathrm{TV}} \leqslant 2\sqrt{\kappa} L(w, y),$$

which proves **Con.10**.

## D. Implementation and additional experiment set-up details

We optimise all MMD objectives using automatic differentiation with the Adam optimiser [Kingma and Ba, 2014] as implemented in `JAX`. The squared MMD between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with kernel $k$ is approximated by the unbiased U–statistic [Gretton et al., 2012] using independent samples $\{x_i\}_{i=1}^n \sim \mathbb{P}$ and $\{y_j\}_{j=1}^s \sim \mathbb{Q}$:

$$\widehat{\mathrm{MMD}}_k^2(\mathbb{P},\mathbb{Q}) = \frac{1}{n(n-1)}\sum_{i\neq i'} k(x_i,x_{i'}) + \frac{1}{s(s-1)}\sum_{j\neq j'} k(y_j,y_{j'}) - \frac{2}{ns}\sum_{i=1}^{n}\sum_{j=1}^{s} k(x_i,y_j).$$

In all experiments we set $s = n$ equal to the number of observations in the corresponding datasets.

For the Berkson ME experiments, the observed covariates $W$ are organised into 100 distinct groups, each repeated 3 times (group size = 3). The 100 distinct group values are drawn i.i.d. from $\mathcal{N}(0,2)$. This design reflects common Berkson settings where $W$ are pre–specified targets, categories, or group averages.

For the Classical ME experiments, the latent covariates $X$ are i.i.d. $\mathcal{N}(0,3)$. We choose the classical–error variance of $X$ to be larger than the variance of $W$ in the Berkson setting so that the marginal scales of $X$ are comparable across the two regimes (recall that in Berkson error $\sigma_X^2 = \sigma_W^2 + \sigma_N^2$).

We use $B_{\mathrm{boot}} = 200$ posterior bootstrap realisations for both R–MEM and NPL–HMC in synthetic experiments, and $B_{\mathrm{boot}} = 100$ for real-world experiments. Since we do not assume a strong prior, the DP concentration parameter is set to $c = 10^{-4}$ for both methods. In all experiments, we use Gaussian (RBF) kernels for $k_X$ and $k_Y$, with bandwidth selected by the median heuristic in every MMD computation [Gretton et al., 2012].

All HMC sampling is performed using `cmdstanpy` (the Python interface to `CmdStan`). Code to reproduce all results in this paper is available at `https://github.com/MengqiChenMC/tot_robust_code`.

## E. HMC diagnostics and sensitivity

### *E.1. HMC mixing diagnostics*

We report diagnostics for $\theta$ in the HMC runs used to produce pseudo samples under the setting with ME scale 1.5, 10% Huber contamination (contaminated points having 9× the clean noise scale), and a working ME scale equal to 0.7× the true scale. Four chains were run with $(T, B) = (10000, 5000)$ and 20,000 post–warm-up draws were retained in total for both the Classical and Berkson ME models. Table 6 summarises the scalar diagnostics for the components of $\theta$ together with sampler-level checks. All $\hat{R}$ values are 1.0, bulk and tail effective sample sizes are large, and Monte-Carlo standard errors are small relative to posterior standard deviations. There were no divergent transitions, no iterations reached the configured maximum tree depth (10), and the per-chain BFMI values are high in both

| **Model** | | mean | sd | hdi 3% | hdi 97% | mcse mean | mcse sd | ess bulk | ess tail | $\hat{R}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_1$ | 4.965 | 0.158 | 4.676 | 5.268 | 0.002 | 0.001 | 7089 | 11685 | 1.0 |
| **Classical ME** | $\theta_2$ | 1.322 | 0.212 | 0.943 | 1.716 | 0.003 | 0.002 | 6199 | 7580 | 1.0 |
| | $\theta_3$ | 0.052 | 0.152 | -0.233 | 0.338 | 0.002 | 0.001 | 5286 | 8461 | 1.0 |
| *Sampler-level:* draws = 20000; divergences = 0; max tree depth = 10 with 0 hits. | | | | | | | | | | |
| | $\theta_1$ | 4.907 | 0.164 | 4.608 | 5.223 | 0.002 | 0.001 | 6512 | 9658 | 1.0 |
| **Berkson ME** | $\theta_2$ | 1.814 | 0.278 | 1.319 | 2.337 | 0.004 | 0.002 | 5901 | 9086 | 1.0 |
| | $\theta_3$ | -0.087 | 0.124 | -0.326 | 0.139 | 0.002 | 0.001 | 5332 | 8806 | 1.0 |
| *Sampler-level:* draws = 20000; divergences = 0; max tree depth = 10 with 0 hits. | | | | | | | | | | |

Table 6: HMC summary diagnostics for $\theta$ (Classical and Berkson).

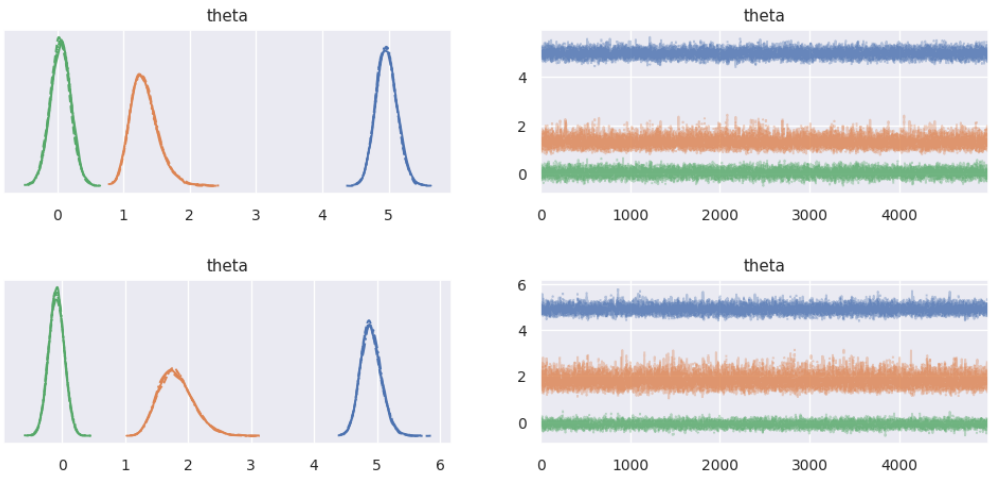



Fig. 6: Trace and marginal density for $\theta$ ($\theta_1$: blue, $\theta_2$: orange, $\theta_3$: green) across four chains: Classical (top) and Berkson (bottom).

models ($\geqslant 0.92$ for all chains), indicating good exploration of energy levels. Figure 6 shows well-mixed traces with stable marginal densities, consistent with sampling from the stationary distribution and low autocorrelation in the retained states. Figure 7 overlays the marginal and transition energy densities and reports the BFMI per chain; the close overlap supports the absence of pathologies. These checks justify using the HMC draws of $\theta$ to implement the independent posterior predictive scheme described in the paper for both ME models.

### *E.2. Sensitivity analysis for the HMC posterior bootstrap*

We compare three ways to draw the latent replicates used by the NPL update.

(a) *Regime A* runs $n \times m$ *independent* HMC chains and retains one post–warm-up draw $\theta_{ij} \sim \Pi_n(\theta \mid W_{1:n}, Y_{1:n})$ from each, followed by one latent draw $X_{ij} \sim \Pi(X_i \mid \theta_{ij}, W_i, Y_i)$. This aligns exactly with the theoretical construct in Section 2.2.3 but is computationally expensive.

(b) *Regime B* fits a single multi-chain HMC run with four parallel chains and takes every $L = 50$-th state for $\theta$, pairing the corresponding latent draws $X_i$ from the same iterations. This assesses sensitivity to thinning.

(c) *Regime C* (default) uses the same multi-chain HMC (four parallel chains) but, instead

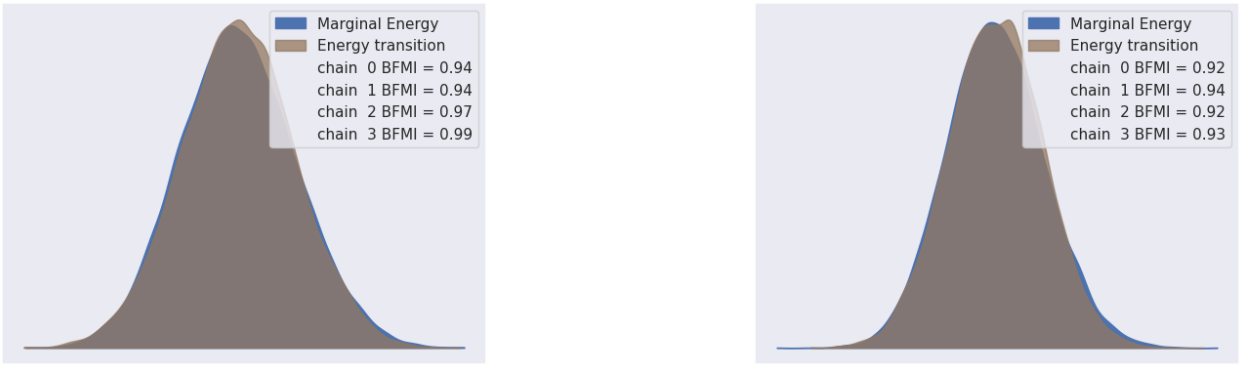



Fig. 7: Marginal energy and energy transitions with per-chain BFMI: Classical (left) and Berkson (right).

| Comparison | joint $\widehat{\mathrm{MMD}}^2$ | Bootstrap 95% CI | Permutation $p$–value |
|---|---|---|---|
| | | $n = 50,\ m = 3$ | |
| A vs B | $-5.51 \times 10^{-3}$ | $[-4.71 \times 10^{-3},\ 1.41 \times 10^{-2}]$ | 1.000 |
| A vs C | $-4.32 \times 10^{-3}$ | $[-4.44 \times 10^{-3},\ 1.49 \times 10^{-2}]$ | 0.919 |
| B vs C | $-3.82 \times 10^{-3}$ | $[-4.18 \times 10^{-3},\ 1.52 \times 10^{-2}]$ | 0.846 |
| | | $n = 100,\ m = 3$ | |
| A vs B | $-2.65 \times 10^{-3}$ | $[-2.28 \times 10^{-3},\ 6.69 \times 10^{-3}]$ | 0.997 |
| A vs C | $-2.21 \times 10^{-3}$ | $[-2.25 \times 10^{-3},\ 7.51 \times 10^{-3}]$ | 0.925 |
| B vs C | $-2.57 \times 10^{-3}$ | $[-2.21 \times 10^{-3},\ 6.72 \times 10^{-3}]$ | 0.990 |
| | | $n = 500,\ m = 3$ | |
| A vs B | $-5.56 \times 10^{-4}$ | $[-4.88 \times 10^{-4},\ 1.20 \times 10^{-3}]$ | 1.000 |
| A vs C | $-5.56 \times 10^{-4}$ | $[-4.68 \times 10^{-4},\ 1.24 \times 10^{-3}]$ | 1.000 |
| B vs C | $-5.46 \times 10^{-4}$ | $[-4.73 \times 10^{-4},\ 1.25 \times 10^{-3}]$ | 1.000 |

Table 7: Berkson: Sensitivity of pseudo–sampling schemes across sample sizes.

of systematic thinning, selects $m$ spaced-out states to define $\theta_{ij}$ and the corresponding $X_{ij}$ for each $i$. No additional thinning is applied.

We use the classical or Berkson ME model with the same ME, model misspecification, and HMC settings as in E.1. We consider $n \in \{50, 100, 500\}$ with $m = 3$ (so $N = nm$ latent replicates per regime).

We compare regimes using the unbiased squared maximum mean discrepancy ($\mathrm{MMD}^2$) with a Gaussian kernel and bandwidth fixed by the median heuristic, applied to the *joint* empirical law of $(X, Y)$ ($\mathrm{MMD}^2$). We report (i) the point estimate $\widehat{\mathrm{MMD}}^2$, (ii) a bootstrap 95% confidence interval (resampling within each regime), and (iii) a permutation $p$–value based on 2,000 randomisations. We present results for $n \in \{50, 100, 500\}$ and $m = 3$ under both Berkson and classical ME models: see Tables 7 and 8. Note that the unbiased $\mathrm{MMD}^2$ estimator can be slightly negative in finite samples. Under equality of distributions, it is $O_p(1/N)$ with $N = nm$.

Across $n \in \{50, 100, 500\}$ the three pairwise *joint* $\mathrm{MMD}^2$ estimates are close to zero and decrease in magnitude as $N = nm$ increases, consistent with the $O_p(1/N)$ scale under equality.

| Comparison | joint $\widehat{\mathrm{MMD}}^2$ | Bootstrap 95% CI | Permutation $p$–value |
|---|---|---|---|
| | | $n = 50,\ m = 3$ | |
| A vs B | $-4.72 \times 10^{-3}$ | $[-4.65 \times 10^{-3},\ 1.64 \times 10^{-2}]$ | 0.960 |
| A vs C | $-3.94 \times 10^{-3}$ | $[-3.94 \times 10^{-3},\ 1.72 \times 10^{-2}]$ | 0.859 |
| B vs C | $-4.93 \times 10^{-3}$ | $[-4.47 \times 10^{-3},\ 1.62 \times 10^{-2}]$ | 0.977 |
| | | $n = 100,\ m = 3$ | |
| A vs B | $-2.67 \times 10^{-3}$ | $[-2.27 \times 10^{-3},\ 7.35 \times 10^{-3}]$ | 1.000 |
| A vs C | $-2.86 \times 10^{-3}$ | $[-2.32 \times 10^{-3},\ 6.72 \times 10^{-3}]$ | 1.000 |
| B vs C | $-2.72 \times 10^{-3}$ | $[-2.41 \times 10^{-3},\ 6.31 \times 10^{-3}]$ | 0.999 |
| | | $n = 500,\ m = 3$ | |
| A vs B | $-5.56 \times 10^{-4}$ | $[-4.73 \times 10^{-4},\ 1.17 \times 10^{-3}]$ | 0.9995 |
| A vs C | $-5.70 \times 10^{-4}$ | $[-4.72 \times 10^{-4},\ 1.18 \times 10^{-3}]$ | 1.000 |
| B vs C | $-5.65 \times 10^{-4}$ | $[-4.76 \times 10^{-4},\ 1.35 \times 10^{-3}]$ | 1.000 |

Table 8: Classical: Sensitivity of pseudo–sampling schemes across sample sizes.

The bootstrap intervals contain 0 and the permutation $p$–values are large (Tables 7–8) for all $n$. There is no evidence that the joint distribution of the pseudo samples $\{(X_{ij}, Y_i)\}$ differs across regimes. In particular, using a few well–mixed chains with sparse retention and paired latent draws yields pseudo–samples that are empirically indistinguishable from those obtained by launching $n \times m$ independent chains. Hence our practical implementation gives the same pseudo–sample distribution, within Monte Carlo uncertainty, as the theoretical construction.

### *E.3. Discussion: independence requirement of Theorem 2*

Our theoretical construct in Section 2.2.3 imposes independence of the pseudo-samples $\{X_{ij}\}$ given $\mathcal{D}$ by drawing $\theta_{ij} \overset{\text{iid}}{\sim} \Pi_n(\theta \mid \mathcal{D})$ and then $X_{ij} \sim \Pi(\cdot \mid W_i, Y_i, \theta_{ij})$. In this section we relax that requirement across $(i, j)$: we draw $m$ parameter states $\theta_j \sim \Pi_n(\theta \mid \mathcal{D})$ that may be dependent (e.g. states from one or a few HMC chains), and for each fixed $j$ we set $X_{ij} \sim \Pi(\cdot \mid W_i, Y_i, \theta_j)$ for $i = 1, \ldots, n$. Conditional on $(\mathcal{D}, \theta_j)$, the collection $\{X_{ij}\}_{i=1}^n$ is independent across $i$ by the i.i.d. nature of $\{(W_i, Y_i)\}$.

We show below that, in finite samples with small $n$, enforcing independence of $\theta_{ij}$ can reduce the sampling error bound (term A of proof B.2) from $2\sqrt{\kappa}/\sqrt{n} + 2\,M_n/\sqrt{n} + 2\sqrt{\kappa}\,r_n$ to $2\sqrt{\kappa}/\sqrt{nm}$. However, due to posterior contraction, the parameter–mixture contribution associated with the $\theta$–mixture in (6) can be bouned of order $M_n/\sqrt{n}+r_n$ regardless of whether or not the $\theta_{ij}$ are independent given $\mathcal{D}$. As $n$ grows (with $m$ fixed), the gain from enforcing independence across $\theta_{ij}$ becomes negligible.

This yields the following practical implementation guide:

(a) *Small $n$:* the posterior $\Pi_n(\theta \mid \mathcal{D})$ may be dispersed; independent (or well-spaced) posterior draws of $\theta$ can improve exploration of the posterior and reduce Monte Carlo error in the pseudo-samples, at modest cost when $m$ is small.

(b) *Large $n$:* as $\Pi_n$ concentrates, the bound is driven by $\frac{M_n}{\sqrt{n}} + r_n$ and near-independence of $\theta$ is unnecessary. Our sensitivity analysis above (Appendix E.2) empirically confirms

that both implementations deliver indistinguishable pseudo-sample distributions as $n$ increases in the regimes considered.

We now derive an alternative bound for term A in the proof of Theorem 2(B.2) without the independence condition on $\{\tilde{X}_{ij}\}_{i=1,j=1}^{n,\ m}$. All expectations over $\theta_{1:m}$ are taken with respect to their joint law induced by the sampler. For each fixed $j$, conditional on $(\mathcal{D}, \theta_j)$ the latent coordinates factorise as

$$\Pi(X_{1:n,j} \mid \mathcal{D}, \theta_j) = \prod_{i=1}^{n} \Pi(X_{ij} \mid W_i, Y_i, \theta_j),$$

because the observations $\{(W_i, Y_i)\}_{i=1}^n$ are iid. Therefore, the model implies conditional independence of $X_i$ given $(W_i, Y_i, \theta_j)$. The proof below first exploits this conditional independence to obtain the $1/\sqrt{n}$ bound, then averages over $j$, and finally integrates over the (possibly dependent) vector $\theta_{1:m}$ using Jensen's inequality and applies posterior contraction.

Recall that term A is

$$\mathbb{E}_{\mathcal{D},S}\, \mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n),$$

where $Q_i = \Psi_n\big(\cdot \mid W_i, Y_i\big)\delta_{Y_i}$ and its embedding is $\varphi(Q_i) \in \mathcal{H}_k$.

Given $(\mathcal{D}, S) \equiv \big\{(W_i, Y_i); (\tilde{X}_{ij})\big\}_{i=1,j=1}^{n,\ m}$ we form the empirical measures for each $i$

$$\widehat{\mathbb{Q}}_i := \frac{1}{m}\sum_{j=1}^{m} \delta_{(\tilde{X}_{ij}, Y_i)}, \qquad \varphi\big(\widehat{\mathbb{Q}}_i\big) = \frac{1}{m}\sum_{j=1}^{m} k\big((\tilde{X}_{ij}, Y_i), \cdot\big) \in \mathcal{H}_k.$$

Rewrite the MMD by re-indexing the double sum:

$$\begin{aligned} \mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n) &= \varphi\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbb{Q}}_i\right) - \varphi(Q_n) \\ &= \frac{1}{m}\sum_{j=1}^{m}\left\{\varphi\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{(\tilde{X}_{ij}, Y_i)}\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbb{Q}}_{ij}\right)\right\} \\ &\quad + \frac{1}{m}\sum_{j=1}^{m}\left\{\varphi\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n)\right\}, \end{aligned} \tag{82}$$

where, for each $i, j$, we set $\tilde{\mathbb{Q}}_{ij} := \Pi(\cdot \mid W_i, Y_i, \theta_j)\delta_{Y_i}$. Fix $j$. Conditional on $(\mathcal{D}, \theta_j)$, the vectors

$$\phi_{ij} := k\big((\tilde{X}_{ij}, Y_i), \cdot\big) \in \mathcal{H}_k, \qquad i = 1, \ldots, n,$$

are independent with mean $\mathbb{E}_{S|\mathcal{D},\theta_j}[\phi_{ij}] = \varphi(\tilde{\mathbb{Q}}_{ij})$ and $\|\phi_{ij}\|^2_{\mathcal{H}_k} \leqslant \kappa_X\kappa_Y$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{S|\mathcal{D},\theta_j}\left\|\varphi\left(\frac{1}{n}\sum_{i=1}^n \delta_{(\tilde{X}_{ij},Y_i)}\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbb{Q}}_{ij}\right)\right\|^2_{\mathcal{H}_k} \\
&\quad= \mathbb{E}_{S|\mathcal{D},\theta_j}\left\|\frac{1}{n}\sum_{i=1}^n\big(\phi_{ij} - \mathbb{E}_{S|\mathcal{D},\theta_j}\,\phi_{ij}\big)\right\|^2_{\mathcal{H}_k} = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{S|\mathcal{D},\theta_j}\left\|\phi_{ij} - \mathbb{E}_{S|\mathcal{D},\theta_j}\,\phi_{ij}\right\|^2_{\mathcal{H}_k} \\
&\quad\leqslant \frac{1}{n^2}\sum_{i=1}^n\big(2\sqrt{\kappa_X\kappa_Y}\big)^2 = \frac{4\kappa_X\kappa_Y}{n}.
\end{aligned}$$

By Jensen's inequality,

$$\mathbb{E}_{S|\mathcal{D},\theta_j}\left\|\varphi\left(\frac{1}{n}\sum_{i=1}^n \delta_{(\tilde{X}_{ij},Y_i)}\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbb{Q}}_{ij}\right)\right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}}.$$

Averaging over $j = 1,\dots,m$ and then over $\theta_{1:m}$ and $\mathcal{D}$, we obtain the bound

$$\mathbb{E}_{\mathcal{D}}\,\mathbb{E}_{\theta_{1:m},S|\mathcal{D}}\left\|\frac{1}{m}\sum_{j=1}^m\left[\varphi\left(\frac{1}{n}\sum_{i=1}^n \delta_{(\tilde{X}_{ij},Y_i)}\right) - \varphi\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbb{Q}}_{ij}\right)\right]\right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_X\kappa_Y}}{\sqrt{n}}. \tag{83}$$

The extra (parameter–mixture) term can be bounded by posterior contraction. By the triangle inequality in $\mathcal{H}_k$ and linearity of $\varphi(\cdot)$,

$$\left\|\varphi\left(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n)\right\|_{\mathcal{H}_k} \leqslant \frac{1}{n}\sum_{i=1}^n \mathrm{MMD}_k\left(\tilde{\mathbb{Q}}_{ij}, \mathbb{Q}_i\right), \qquad \mathbb{Q}_i := \Psi_n(\cdot \mid W_i, Y_i)\delta_{Y_i}.$$

By Lemma 5 and Lemma 4,

$$\begin{aligned}
\mathrm{MMD}_k\left(\tilde{\mathbb{Q}}_{ij}, \mathbb{Q}_i\right) &= \mathrm{MMD}_k\left(\Pi_{\theta_j}(\cdot \mid W_i, Y_i)\delta_{Y_i}, \int \Pi_\vartheta(\cdot \mid W_i, Y_i)\delta_{Y_i}\Pi_n(d\vartheta)\right) \\
&\leqslant \int \mathrm{MMD}_k\left(\Pi_{\theta_j}(\cdot \mid W_i, Y_i)\delta_{Y_i}, \Pi_\vartheta(\cdot \mid W_i, Y_i)\delta_{Y_i}\right)\Pi_n(d\vartheta) \\
&\leqslant \sqrt{\kappa_Y}\int \mathrm{MMD}_{k_X}\left(\Pi_{\theta_j}(\cdot \mid W_i, Y_i), \Pi_\vartheta(\cdot \mid W_i, Y_i)\right)\Pi_n(d\vartheta).
\end{aligned}$$

Split the $\vartheta$–integral over $B_n \cup B_n^c$. On $B_n$, Assumption A2 gives

$$\mathrm{MMD}_{k_X}\left(\Pi_{\theta_j}(\cdot \mid W_i, Y_i), \Pi_\vartheta(\cdot \mid W_i, Y_i)\right) \leqslant L(W_i, Y_i)\|\theta_j - \vartheta\|.$$

Hence, for any fixed $\theta_j$,

$$\int_{B_n} \mathrm{MMD}_{k_X}\big(\Pi_{\theta_j}, \Pi_\vartheta\big)\Pi_n(d\vartheta) \leqslant \begin{cases} L(W_i, Y_i)\int_{B_n}\|\theta_j - \vartheta\|\Pi_n(d\vartheta) \leqslant \frac{2M_n}{n}L(W_i, Y_i), & \theta_j \in B_n, \\ \sqrt{\kappa_X}\Pi_n(B_n), & \theta_j \in B_n^c, \end{cases}$$

where we used $\|\theta_j - \vartheta\| \leqslant \|\theta_j - \theta^*\| + \|\vartheta - \theta^*\| \leqslant 2M_n/\sqrt{n}$ in the first case and $\mathrm{MMD}_{k_X} \leqslant \sqrt{\kappa_X}$ in the second. On $B_n^c$, we have

$$\int_{B_n^c} \mathrm{MMD}_{k_X}\big(\Pi_{\theta_j}, \Pi_\vartheta\big)\Pi_n(d\vartheta) \leqslant \sqrt{\kappa_X}\Pi_n(B_n^c).$$

Combining the pieces and averaging over $i$ gives, for each fixed $j$ and $\theta_j$,

$$\left\| \varphi\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n) \right\|_{\mathcal{H}_k} \leqslant \sqrt{\kappa_Y}\left[ \mathbb{1}_{\{\theta_j\in B_n\}}\cdot\frac{2M_n}{\sqrt{n}}\bar{L}(W,Y) + \mathbb{1}_{\{\theta_j\in B_n^c\}}\cdot\sqrt{\kappa_X}\Pi_n(B_n) + \sqrt{\kappa_X}\Pi_n(B_n^c)\right],$$

where $\bar{L}(W,Y) := \frac{1}{n}\sum_{i=1}^{n} L(W_i,Y_i)$. Taking expectation over $\theta_j \sim \Pi_n(\cdot \mid \mathcal{D})$ and using

$$\mathbb{E}_{\theta_j|\mathcal{D}}\big[\mathbb{1}_{\{\theta_j\in B_n\}}\big] = \Pi_n(B_n), \qquad \mathbb{E}_{\theta_j|\mathcal{D}}\big[\mathbb{1}_{\{\theta_j\in B_n^c\}}\big] = \Pi_n(B_n^c),$$

together with $\int_{B_n}\|\vartheta-\theta^*\|\Pi_n(d\vartheta) \leqslant M_n/\sqrt{n}$, we obtain

$$\mathbb{E}_{\theta_j|\mathcal{D}}\left\| \varphi\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n) \right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_Y}M_n}{\sqrt{n}}\bar{L}(W,Y) + 2\sqrt{\kappa_X\kappa_Y}\Pi_n(B_n^c).$$

By convexity of the norm,

$$\mathbb{E}_{\theta_{1:m}|\mathcal{D}}\left\| \frac{1}{m}\sum_{j=1}^{m}\left\{ \varphi\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n)\right\} \right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_Y}M_n}{\sqrt{n}}\bar{L}(W,Y) + 2\sqrt{\kappa_X\kappa_Y}\Pi_n(B_n^c).$$

Finally, taking expectation over $\mathcal{D}$ and using $C_L := \mathbb{E}[\bar{L}(W,Y)] < \infty$ and $r_n := \mathbb{E}_{\mathcal{D}}[\Pi_n(B_n^c)] \to 0$, we obtain

$$\mathbb{E}_{\mathcal{D}}\,\mathbb{E}_{\theta_{1:m}|\mathcal{D}}\left\| \frac{1}{m}\sum_{j=1}^{m}\left\{ \varphi\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbb{Q}}_{ij}\right) - \varphi(Q_n)\right\} \right\|_{\mathcal{H}_k} \leqslant \frac{2\sqrt{\kappa_Y}C_L M_n}{\sqrt{n}} + 2\sqrt{\kappa_X\kappa_Y}r_n. \tag{84}$$

As in (40), rescale $M_n$ by a fixed constant if desired (replace $M_n$ with $M_n/\max\{2\sqrt{\kappa_Y}C_L, 1\}$) to write the right–hand side as $\frac{M_n}{\sqrt{n}} + \sqrt{\kappa_X\kappa_Y}r_n$. Combining (83) and (84) gives

$$\mathbb{E}_{\mathcal{D},S}\,\mathrm{MMD}_k(\mathbb{P}_{XY}^{\mathrm{pseudo}}, Q_n) \leqslant \frac{2\sqrt{\kappa}}{\sqrt{n}} + \frac{2M_n}{\sqrt{n}} + 2\sqrt{\kappa}r_n. \tag{85}$$