Deconditional Kernel Mean Embeddings and Gaussian Processes

Siu Lun Chau

College of Computing & Data Science,
Nanyang Technological University, Singapore

September 17, 2025

Outline

1. About my research
2. Background on Kernel Embeddings and Gaussian Process
3. Deconditional Downscaling with Gaussian Processes

From Probabilistic to Imprecise Probabilistic Machine Learning

Phase 1: Probabilistic Machine Learning

- DPhil Thesis: *Towards Trustworthy Machine Learning with Kernels*
- TL;DR: Methodological developments for kernel embedding of distributions and Gaussian process modelling, with applications to preference learning and explainability.

From Probabilistic to Imprecise Probabilistic Machine Learning

Phase 1: Probabilistic Machine Learning

- DPhil Thesis: *Towards Trustworthy Machine Learning with Kernels*
- TL;DR: Methodological developments for kernel embedding of distributions and Gaussian process modelling, with applications to preference learning and explainability.

Phase 2: Imprecise Probabilistic Machine Learning

- *"There is more to uncertainty than probability"* (SIPTA): credal sets, probability intervals, belief functions, possibility measures, Choquet capacities...
- TL;DR: How to integrate these mathematical models into machine learning pipelines to allow for more explicit appreciation of (epistemic) uncertainty?

Current research interests:

Foundations of Epistemic Uncertainty in

- Uncertainty representation and quantification [Singh et al., 2024]
- Measuring uncertainty discrepancy [Chau et al., 2025a]
- Validating uncertainty [Chau et al., 2025b, Singh et al., 2025]

Current research interests:

Foundations of Epistemic Uncertainty in

- Uncertainty representation and quantification [Singh et al., 2024]
- Measuring uncertainty discrepancy [Chau et al., 2025a]
- Validating uncertainty [Chau et al., 2025b, Singh et al., 2025]

Applications of Epistemic Uncertainty in

- Economic aspect of machine learning, such as credit allocation, mechanism design, strategic learning, causal inference, where epistemic uncertainty is not generally reducible. [Chau et al., 2021, Vo et al., 2024, 2025]
- Interpretability under uncertainty [Chau et al., 2023, Adachi et al., 2024]

Outline

1. About my research
2. Background on Kernel Embeddings and Gaussian Process
3. Deconditional Downscaling with Gaussian Processes

Kernels and Reproducing Kernel Hilbert Spaces

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation.
- Kernel function is as an *inner product of features*: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.

Kernels and Reproducing Kernel Hilbert Spaces

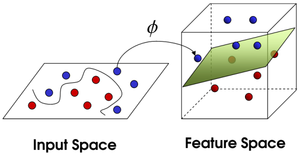
- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation.
- Kernel function is as an *inner product of features*: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.
- There exists a canonical feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS), with canonical feature map $\mapsto k(\cdot, x)$, where
  1. $\forall x \in \mathcal{X},\ k(\cdot, X) \in \mathcal{H}_k$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k,\ \langle f, k(\cdot, X) \rangle_{\mathcal{H}_k} = f(x)$.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation.
- Kernel function is as an *inner product of features*: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.
- There exists a canonical feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS), with canonical feature map $\mapsto k(\cdot, x)$, where
  1. $\forall x \in \mathcal{X},\ k(\cdot, X) \in \mathcal{H}_k$, and
  2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k,\ \langle f, k(\cdot, X) \rangle_{\mathcal{H}_k} = f(x)$.
- Moore-Aronszajin Theorem: every positive semidefinite $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel of a unique RKHS $\mathcal{H}_k$.

Kernel Trick and Mean Embedding

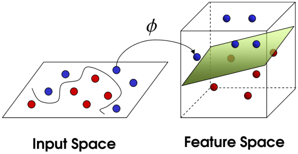
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes and Vapnik, 1995, Schölkopf et al., 1999]

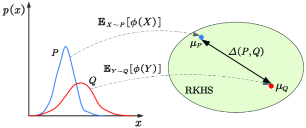
Kernel Trick and Mean Embedding

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes and Vapnik, 1995, Schölkopf et al., 1999]

- **RKHS embedding**: implicit feature mean [Sriperumbudur et al., 2011, Muandet et al., 2017]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X\sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q)\rangle_{\mathcal{H}_k} = \mathbb{E}_{X\sim P, Y\sim Q} k(X, Y)$
  inner products easy to estimate
  - nonparametric two-sample, independence, conditional independence, interaction testing, learning on distributions

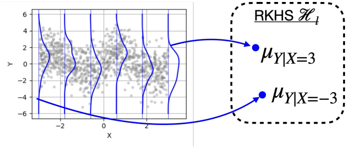


[Gretton et al., 2006, 2007, Muandet et al., 2012, Szabó et al., 2016]

Conditional Mean Embeddings

- Consider a joint distribution P_{XY} over random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. The conditional mean embedding (CME) of $P(Y \mid X = x)$ is defined as

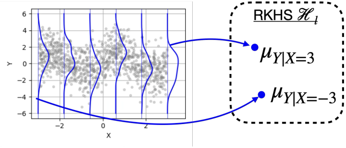
$$\mu_{Y \mid X=x} := \mathbb{E}_{Y \mid X=x}[k_y(\cdot, Y)] = \int_{\mathcal{Y}} k({}_y(\cdot, y) dP(y \mid X = x) \in \mathcal{H}_{k_y}$$



Conditional Mean Embeddings

- Consider a joint distribution P_{XY} over random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. The conditional mean embedding (CME) of $P(Y \mid X = x)$ is defined as

$$\mu_{Y|X=x} := \mathbb{E}_{Y|X=x}[k_y(\cdot, Y)] = \int_{\mathcal{Y}} k(_y(\cdot, y)dP(y \mid X = x) \in \mathcal{H}_{k_y}$$

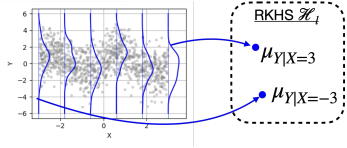


- To model CMEs as functions of x, we can either take an operator perspective, i.e. define a conditional mean operator (CMO) $C_{Y|X} : \mathcal{H}_{k_x} \to \mathcal{H}_{k_y}$ which satisfies $\mu_{Y|X=x} = C_{Y|X} k_x(\cdot, x)$,

Conditional Mean Embeddings

- Consider a joint distribution P_{XY} over random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. The conditional mean embedding (CME) of $P(Y \mid X = x)$ is defined as

$$\mu_{Y|X=x} := \mathbb{E}_{Y|X=x}[k_y(\cdot, Y)] = \int_{\mathcal{Y}} k(_y(\cdot, y) dP(y \mid X = x) \in \mathcal{H}_{k_y}$$



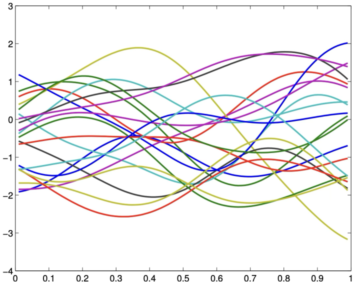
- To model CMEs as functions of x, we can either take an operator perspective, i.e. define a conditional mean operator (CMO) $C_{Y|X} : \mathcal{H}_{k_x} \to \mathcal{H}_{k_y}$ which satisfies $\mu_{Y|X=x} = C_{Y|X} k_x(\cdot, x)$,
- or take a vector-valued regression perspective, i.e. solve for

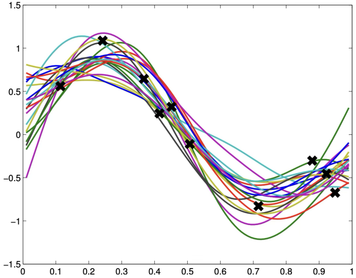
$$\mu_{Y|X} = \arg\min_{F \in \Gamma} \mathbb{E}_{XY} \|k_y(\cdot, Y) - F(X)\|^2_{\mathcal{H}_{k_y}}$$

Gaussian Processes

Consider function evaluations $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$ at a set of inputs, and observations $\mathbf{y} = (y_1, \ldots, y_n)$ with

$$\mathbf{f} \sim N(0, \mathbf{K})$$
$$\mathbf{y} \mid \mathbf{f} \sim \prod p(y_i \mid f(x_i))$$





GP priors on RKHS

- Can we formulate a GP model for marginal embeddings $\mu_{P_X}(\cdot) = \mathbb{E}[k(\cdot, X)]$? [Flaxman et al., 2016]

GP priors on RKHS

- Can we formulate a GP model for marginal embeddings $\mu_{P_X}(\cdot) = \mathbb{E}[k(\cdot, X)]$? [Flaxman et al., 2016]
- Note that the sample paths of a GP with kernel k lie outside $\mathcal{H}_k$ with probability 1 (Kallianpur's 0-1 law [Jain, 1971])

GP priors on RKHS

- Can we formulate a GP model for marginal embeddings $\mu_{P_X}(\cdot) = \mathbb{E}[k(\cdot, X)]$? [Flaxman et al., 2016]
- Note that the sample paths of a GP with kernel k lie outside $\mathcal{H}_k$ with probability 1 (Kallianpur's 0-1 law [Jain, 1971])
- A smoother kernel k can be used, e.g.

$$r(x, x') = \int k(x, u)k(u, x')\nu(dx),$$

then sample paths $f \in \mathcal{H}_k$ with probability 1 by nuclear dominance theory [Lukić and Beder, 2001], for any finite measure ν.

Outline

1. About my research
2. Background on Kernel Embeddings and Gaussian Process
3. Deconditional Downscaling with Gaussian Processes

This presentation is based on

Chau, SL*, *Shahine Bouabid*, and Dino Sejdinovic. "Deconditional downscaling with gaussian processes." Advances in Neural Information Processing Systems 34 (2021): 17813-17825.*

Deconditional Downscaling with Gaussian Processes

Siu Lun Chau[*†]
University of Oxford

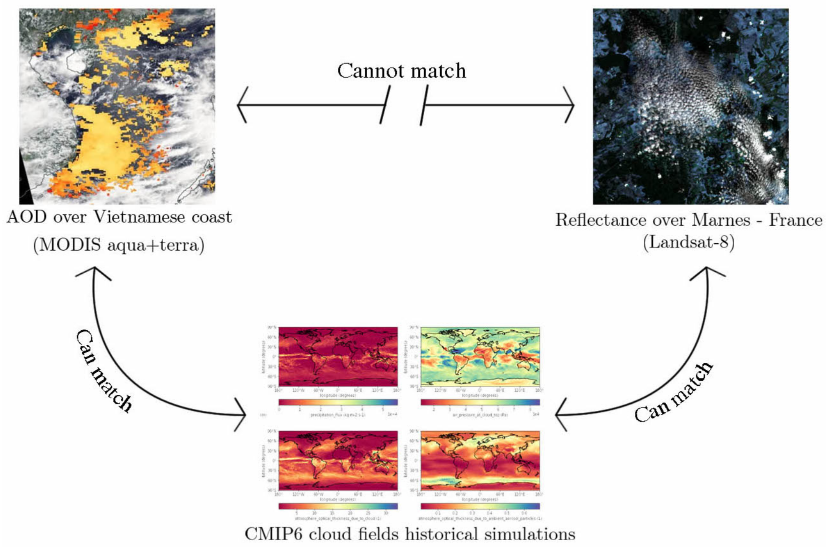
Shahine Bouabid[*†]
University of Oxford

Dino Sejdinovic[†]
University of Oxford

Abstract

Refining low-resolution (LR) spatial fields with high-resolution (HR) information, often known as *statistical downscaling*, is challenging as the diversity of spatial datasets often prevents direct matching of observations. Yet, when LR samples are modeled as aggregate conditional means of HR samples with respect to a mediating variable that is globally observed, the recovery of the underlying fine-grained field can be framed as taking an "inverse" of the conditional expectation, namely a *deconditioning problem*. In this work, we propose a Bayesian formulation of deconditioning which naturally recovers the initial reproducing kernel Hilbert space formulation from Hsu and Ramos [1]. We extend deconditioning to a downscaling setup and devise efficient conditional mean embedding estimator for multiresolution data. By treating conditional expectations as inter-domain features of the underlying field, a posterior for the latent field can be established as

Motivation



Cannot match
AOD over Vietnamese coast
(MODIS aqua+terra)
Reflectance over Marnes - France
(Landsat-8)
Can match
Can match
CMIP6 cloud fields historical simulations

Problem Setup

Data

- We have a dataset of N bags of high-resolution (HR) covariates ${}^{b}\boldsymbol{x}_j := \{x_j^{(1)}, \ldots, x_j^{(n_j)}\}$ each paired with a mediating low-resolution (LR) variable y_j

$$\mathcal{D}_1 = \{{}^{b}\boldsymbol{x}_j, y_j\}_{j=1}^{N}.$$

- We have a separate dataset of M mediating LR variables $\tilde{y}_j$ paired with a LR response of interest $\tilde{z}_j$.

$$\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^{M}.$$

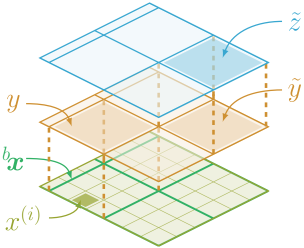


Figure: Illustration of HR and LR observations – indirect pairing

Problem Setup

Objective

- Downscale response z to the HR granularity level of $x_j^{(i)}$ covariates
  i.e. find a function $f : \mathcal{X} \to \mathbb{R}$ which maps between HR covariates and HR responses.

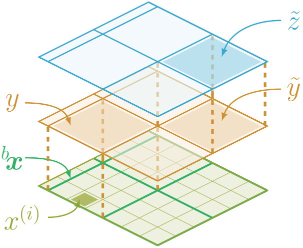


Figure: We wish to learn a map from HR covariates to an HR estimate of the response

Deconditional Formulation

Observation Model

- We assume that the HR responses $f(x)$ aggregate into the LR response $\tilde{z}_j$ as

$$\tilde{z}_j = \mathbb{E}_X[f(X)|Y = \tilde{y}_j] + \varepsilon_j$$

with noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

Deconditional Formulation

Observation Model

- We assume that the HR responses $f(x)$ aggregate into the LR response $\tilde{z}_j$ as

$$\tilde{z}_j = \mathbb{E}_X[f(X)|Y = \tilde{y}_j] + \varepsilon_j$$

  with noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

This is similar to the *deconditioning* problem studied by Hsu & Ramos (2019):

- Given an RKHS function $g : \mathcal{Y} \to \mathbb{R}$, infer an RKHS function $f : \mathcal{X} \to \mathbb{R}$ such that

$$g(y) = \mathbb{E}_X[f(X)|Y = y].$$

  f is called the *deconditional mean of* g w.r.t. $\mathbb{P}_{X|Y}$.

Hsu and Ramos [2019] develop a deconditioning procedure based on estimating so called deconditional mean operators and complex chained inference derivations.

Bayesian formulation for f and g

- By placing a GP prior on $f \sim \mathcal{GP}(m, k)$, we can represent the LR field of responses as

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x)\mathbb{P}_{X|Y=y}(x) \sim \mathcal{GP}(\nu, q).$$

Bayesian formulation for f and g

- By placing a GP prior on $f \sim \mathcal{GP}(m, k)$, we can represent the LR field of responses as

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x)\mathbb{P}_{X|Y=y}(x) \sim \mathcal{GP}(\nu, q).$$

By linearity of expectation, g is also a GP where

$$\nu(y) = \mathbb{E}_X[m(X)|Y = y]$$

$$q(y, y') = \mathbb{E}_{X,X'}[k(X, X')|Y = y, Y' = y'] = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle$$

we call g the conditional mean process.

Bayesian formulation for f and g

- By placing a GP prior on $f \sim \mathcal{GP}(m, k)$, we can represent the LR field of responses as

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x)\mathbb{P}_{X|Y=y}(x) \sim \mathcal{GP}(\nu, q).$$

  By linearity of expectation, g is also a GP where

$$\nu(y) = \mathbb{E}_X[m(X)|Y = y]$$

$$q(y, y') = \mathbb{E}_{X,X'}[k(X, X')|Y = y, Y' = y'] = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle$$

  we call g the conditional mean process.
- Estimation of ν and q via conditional mean embeddings using $\mathcal{D}_1$.

Bayesian formulation for f and g

- By placing a GP prior on $f \sim \mathcal{GP}(m, k)$, we can represent the LR field of responses as

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x)\mathbb{P}_{X|Y=y}(x) \sim \mathcal{GP}(\nu, q).$$

  By linearity of expectation, g is also a GP where

$$\nu(y) = \mathbb{E}_X[m(X)|Y = y]$$

$$q(y, y') = \mathbb{E}_{X,X'}[k(X, X')|Y = y, Y' = y'] = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle$$

  we call g the conditional mean process.

- Estimation of ν and q via conditional mean embeddings using $\mathcal{D}_1$.
- By joint normality between LR and HR fields, recover a posterior for HR field f using $\mathcal{D}_2$.

Side track: Application of CMP to Interpretability

In feature attribution problems, we often quantify the importance of a feature subset $S \subseteq [d]$ at instance x by

$$\omega(S, f, x) = \mathbb{E}[f(X) \mid X_S = x_S] - \mathbb{E}[f(X)]$$

Side track: Application of CMP to Interpretability

In feature attribution problems, we often quantify the importance of a feature subset $S \subseteq [d]$ at instance x by

$$\omega(S, f, x) = \mathbb{E}[f(X) \mid X_S = x_S] - \mathbb{E}[f(X)]$$

1. *How to explain Kernel methods with CMEs?* [Chau et al., 2022, Mohammadi et al., 2025a]
2. *How to explain Gaussian processes through the (stochastic) Shapley value formulation?* [Chau et al., 2023]
3. *How to incorporate GPSHAP for an explainable Bayesian optimisation?* [Adachi et al., 2024]
4. *How to turn exact computation of Stochastic Shapley values from exponential to quadratic?* [Mohammadi et al., 2025b]

Deconditional Posterior

Joint normality between LR and HR field:

The latent HR field $f(x)$ and the observed noisy LR field $\tilde{z} = g(\tilde{y}) + \epsilon$ are jointly normal:

$$\begin{bmatrix} f(x) \\ \tilde{z} \end{bmatrix} \mid \tilde{y} \sim \mathcal{N}\left(\begin{bmatrix} m(x) \\ \nu(\tilde{y}) \end{bmatrix}, \begin{bmatrix} k(x,x) & \langle k(x,\cdot), C_{X|Y}\ell(\tilde{y},\cdot)\rangle_{\mathcal{H}_k} \\ \langle C_{X|Y}\ell(\tilde{y},\cdot), k(x,\cdot)\rangle_{\mathcal{H}_k} & q(y,y) + \sigma^2 \end{bmatrix} \right)$$

Deconditional Posterior

Joint normality between LR and HR field:

The latent HR field $f(x)$ and the observed noisy LR field $\tilde{z} = g(\tilde{y}) + \epsilon$ are jointly normal:

$$\begin{bmatrix} f(x) \\ \tilde{z} \end{bmatrix} \mid \tilde{y} \sim \mathcal{N}\left(\begin{bmatrix} m(x) \\ \nu(\tilde{y}) \end{bmatrix}, \begin{bmatrix} k(x,x) & \langle k(x,\cdot), C_{X|Y}\ell(\tilde{y},\cdot)\rangle_{\mathcal{H}_k} \\ \langle C_{X|Y}\ell(\tilde{y},\cdot), k(x,\cdot)\rangle_{\mathcal{H}_k} & q(y,y)+\sigma^2 \end{bmatrix}\right)$$

- Allows to directly obtain *deconditional posterior* $f|\tilde{z} \sim \mathcal{GP}(m_d, k_d)$ from $\mathcal{D}_2$ with:

$$\hat{m}_{\mathrm{d}}(x) = m(x) + k(x,\mathbf{x})\mathbf{A}(\hat{\mathbf{Q}} + \sigma^2{}_M)^{-1}(\tilde{\mathbf{z}} - \nu(\tilde{\mathbf{y}}))$$
$$\hat{k}_{\mathrm{d}}(x,x') = k(x,x') - k(x,\mathbf{x})(\hat{\mathbf{Q}} + \sigma^2{}_M)^{-1\top}k(\mathbf{x},x')$$

where $\mathbf{A} := (\ell(\mathbf{y},\mathbf{y}) + N\lambda_N)^{-1}\ell(\mathbf{y},\tilde{\mathbf{y}})$ with $\lambda > 0$, $\hat{\mathbf{Q}} := \hat{q}(\tilde{\mathbf{y}},\tilde{\mathbf{y}})$.

Posterior mean has a form essentially identical to the estimator by Hsu and Ramos [2019]

Additional Contributions: Convergence rate for DMO

- Deconditioning can be formulated as the vector-valued regression of the operator $D_{X|Y} : \mathcal{H}_k \to \mathcal{H}_\ell$ such that

$$D_{X|Y}^\top C_{X|Y}^\top f = f \quad \forall f \in \mathcal{H}_k$$

Additional Contributions: Convergence rate for DMO

- Deconditioning can be formulated as the vector-valued regression of the operator $D_{X|Y} : \mathcal{H}_k \to \mathcal{H}_\ell$ such that

$$D_{X|Y}^\top C_{X|Y}^\top f = f \quad \forall f \in \mathcal{H}_k$$

Convergence Rate

Assume $\mathcal{H}_\ell$ is finite dimensional and place mild assumptions on original spaces, kernels, RKHSs and probability distributions, which are characterized by parameters $b > 1$, $c, c' \in]1, 2]$ and $\iota \in]0, 1[$. Let

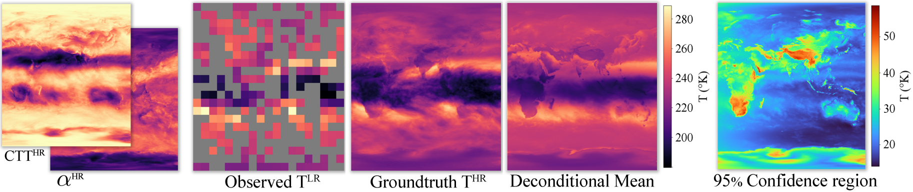
$$\mathcal{E}_\mathrm{d}(D) = \mathbb{E}[\|\ell(Y, \cdot) - DC_{X|Y}\ell(Y, \cdot)\|^2_{\mathcal{H}_\ell}]$$

the exact regression objective and $D^\star = \arg\min_{\mathrm{HS}(\mathcal{H}_k, \mathcal{H}_\ell)} \mathcal{E}_\mathrm{d}$.

Then if we choose $\lambda = N^{-\frac{1}{c'+1}}$ and $N = M^{\frac{a(c'+1)}{\iota(c'-1)}}$ with $a > 0$, we have

- If $a \le \frac{b(c+1)}{bc+1}$, then $\mathcal{E}_\mathrm{d}(\hat{D}_{X|Y}) - \mathcal{E}_\mathrm{d}(D^\star) = \mathcal{O}(M^{\frac{-ac}{c+1}})$ with $\epsilon = M^{\frac{-a}{c+1}}$
- If $a \ge \frac{b(c+1)}{bc+1}$, then $\mathcal{E}_\mathrm{d}(\hat{D}_{X|Y}) - \mathcal{E}_\mathrm{d}(D^\star) = \mathcal{O}(M^{\frac{-bc}{bc+1}})$ with $\epsilon = M^{\frac{-b}{bc+1}}$

Mediated Downscaling of Atmospheric Temperature



Model	RMSE ↓	MAE ↓	Corr. ↑	SSIM ↑
Kriging	8.02±0.28	5.55±0.17	0.831±0.012	**0.212**±0.011
VBAgg	8.25±0.15	5.82±0.11	0.821±0.006	0.182±0.004
Our method	**7.40**±0.25	**5.34**±0.22	**0.848**±0.011	**0.212**±0.013

Table: Downscaling similarity scores of posterior mean against HR groundtruth; reports 1 s.d. VBAgg approach from Law et al (2018) also operates on aggregate likelihoods but cannot handle unmatched data and thus requires to first estimate LR response for each bag of HR covariates. It can be thought of as a special case of the proposed method where mediating LR covariate is simply one-hot encoding of the bag.

Summary

- A scalable Bayesian solution to the mediated statistical downscaling problem, which handles unmatched multi-resolution data.
- Combines Gaussian Processes with the framework of deconditioning using RKHSs and recovers previous approaches as its special cases.
- Future challenges: can we integrate this framework to instrumental and proximal variables problems in causal inference?

EurIPS Workshop: Epistemic Intelligence in Machine Learning

EIML EPISTEMIC INTELLIGENCE IN MACHINE LEARNING

Still think Epistemic Uncertainty is just error bars?

Join us at the EIML workshop @ EurIPS 2025; where we bring together researchers to explore ***foundational***, ***methodological***, *and* ***practical*** *questions around Epistemic Uncertainty in machine learning!*

brought to you by:

Michele Caprio (Manchester UK) | Siu Lun Chau (NTU Singapore) | Ruobin Gong (Rutgers US) | Shireen Manchingal (Oxford Brookes UK) | Krikamol Muandet (CISPA Germany) | Bob Williamson (Tubingen Germany)

References I

Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. *arXiv preprint arXiv:2404.04669*, 2024.

Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. *arXiv preprint arXiv:2505.16156*, 2025a.

Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2025b.

Anurag Singh, Siu Lun Chau, and Krikamol Muandet. Truthful elicitation of imprecise forecasts. *arXiv preprint arXiv:2503.16395*, 2025.

Siu Lun Chau, Jean-Francois Ton, Javier González, Yee Teh, and Dino Sejdinovic. Bayesimp: Uncertainty quantification for causal data fusion. *Advances in Neural Information Processing Systems*, 34:3466–3477, 2021.

Kiet QH Vo, Muneeb Aadil, Siu Lun Chau, and Krikamol Muandet. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15411–15419, 2024.

References II

Kiet QH Vo, Siu Lun Chau, Masahiro Kato, Yixin Wang, and Krikamol Muandet. Explanation design in strategic learning: Sufficient explanations that induce non-harmful responses. *arXiv preprint arXiv:2502.04058*, 2025.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36:50769–50795, 2023.

Masaki Adachi, Brady Planden, David Howey, Michael A Osborne, Sebastian Orbell, Natalia Ares, Krikamol Muandet, and Siu Lun Chau. Looping in the human: Collaborative and explainable bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 505–513. PMLR, 2024.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support Vector Method for Novelty Detection. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL `https://proceedings.neurips.cc/paper_files/paper/1999/hash/8725fb777f25776ffa9076e44fcfd776-Abstract.html`.

References III

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19:513–520, 2006.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25*, pages 10–18. 2012.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

References IV

Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Association for Computing Machinery, 2016.

Naresh C Jain. A zero-one law for gaussian processes. *Proceedings of the American Mathematical Society*, 29(3):585–587, 1971.

Milan Lukić and Jay Beder. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

Kelvin Hsu and Fabio Ramos. Bayesian deconditional kernel mean embeddings. In *International Conference on Machine Learning*, pages 2830–2838. PMLR, 2019.

Siu Lun Chau, Robert Hu, Javier Gonzalez, and Dino Sejdinovic. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 35:13050–13063, 2022.

Majid Mohammadi, Siu Lun Chau, and Krikamol Muandet. Computing exact shapley values in polynomial time for product-kernel methods. *arXiv preprint arXiv:2505.16516*, 2025a.

References V

Majid Mohammadi, Krikamol Muandet, Ilaria Tiddi, Annette Ten Teije, and Siu Lun Chau. Exact shapley attributions in quadratic-time for fanova gaussian processes. *arXiv preprint arXiv:2508.14499*, 2025b.