Sobolev Norm Learning Rates for Conditional Mean Embeddings

Prem Talwai
MIT Operations Research Center

Joint work with Ali Shameli and David Simchi-Levi

Conditional Mean Embeddings

- **Mean embeddings** are a popular technique for representing distributions in Hilbert space [5]

Conditional Mean Embeddings

- **Mean embeddings** are a popular technique for representing distributions in Hilbert space [5]
- Let $\mathcal{H}_G$ be an reproducing kernel Hilbert space (RKHS) with p.d. kernel $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and nonnegative measure $\gamma \in \mathcal{M}^+(\mathcal{X})$. Then $\mu_\gamma \in \mathcal{H}_G$:

$$\mu_\gamma = \int_{\mathcal{X}} G(x, \cdot) d\mu(x)$$

Conditional Mean Embeddings

- **Mean embeddings** are a popular technique for representing distributions in Hilbert space [5]
- Let $\mathcal{H}_G$ be an reproducing kernel Hilbert space (RKHS) with p.d. kernel $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and nonnegative measure $\gamma \in \mathcal{M}^+(\mathcal{X})$. Then $\mu_\gamma \in \mathcal{H}_G$:

$$\mu_\gamma = \int_{\mathcal{X}} G(x, \cdot) d\mu(x)$$

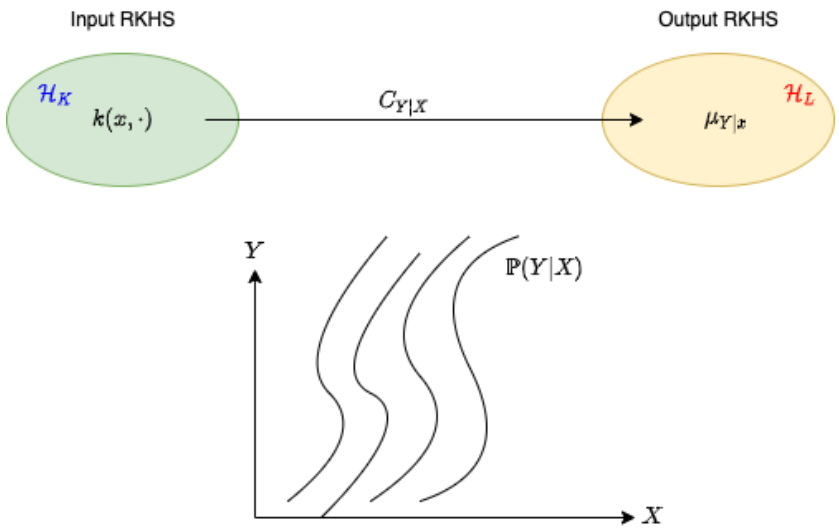
- **Conditional mean embeddings** encode *dependencies* between random variables X and Y:
  - Map input features $\phi(x)$ in one RKHS to the mean embedding of the corresponding *conditional* distribution $\mu_{Y|x}$ in another RKHS

Conditional Mean Embeddings

- Random variables X and Y over $\mathcal{X}$ and $\mathcal{Y}$, embedded in RKHS $\mathcal{H}_K$ and $\mathcal{H}_L$, respectively.

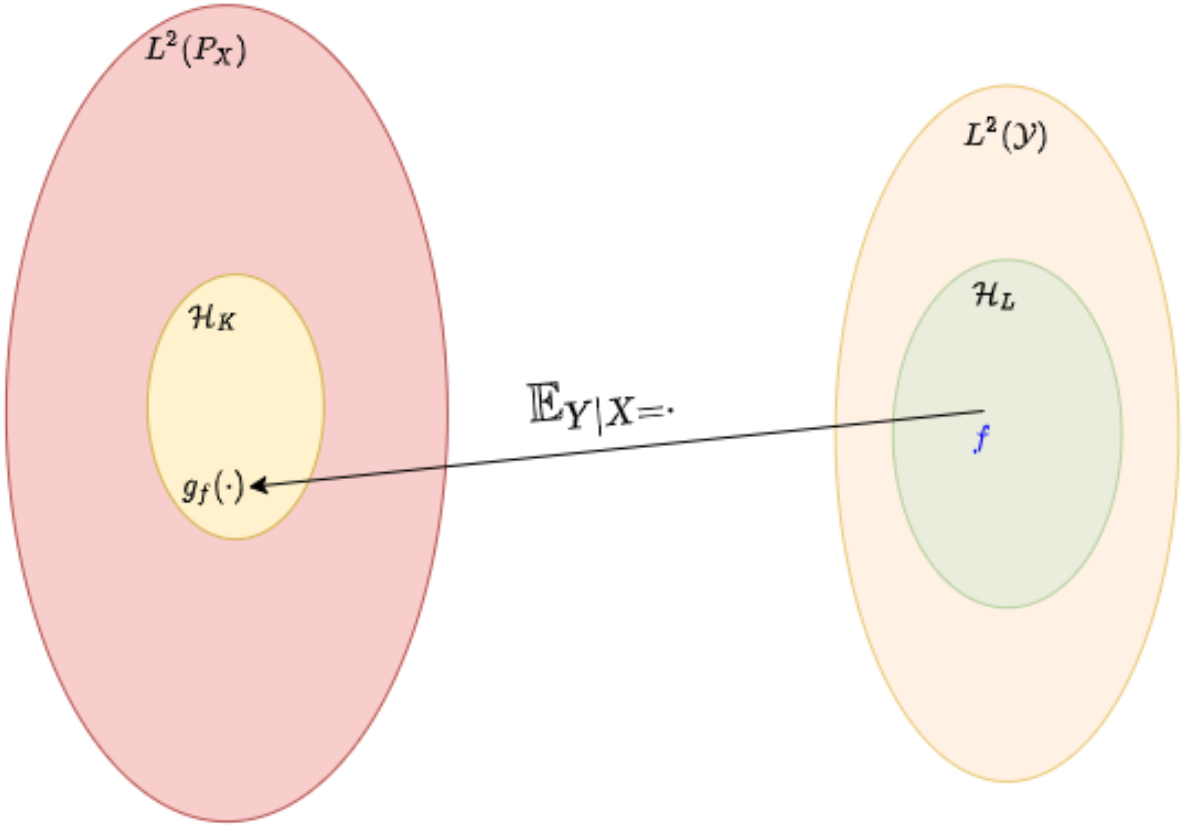
Conditional Mean Embeddings

- Random variables X and Y over $\mathcal{X}$ and $\mathcal{Y}$, embedded in RKHS $\mathcal{H}_K$ and $\mathcal{H}_L$, respectively.
- The conditional mean embedding (CME) $C_{Y|X} : \mathcal{H}_K \to \mathcal{H}_L$ is defined such that:
  - $\mu_{Y|x} \equiv \mathbb{E}_{Y|x}[l(Y, \cdot)] = C_{Y|X} k(x, \cdot)$
  - $\mathbb{E}_{Y|x}[g(\cdot)] = \langle g, \mu_{Y|x} \rangle_L$ for all $g \in \mathcal{H}_L$ and $x \in \mathcal{X}$



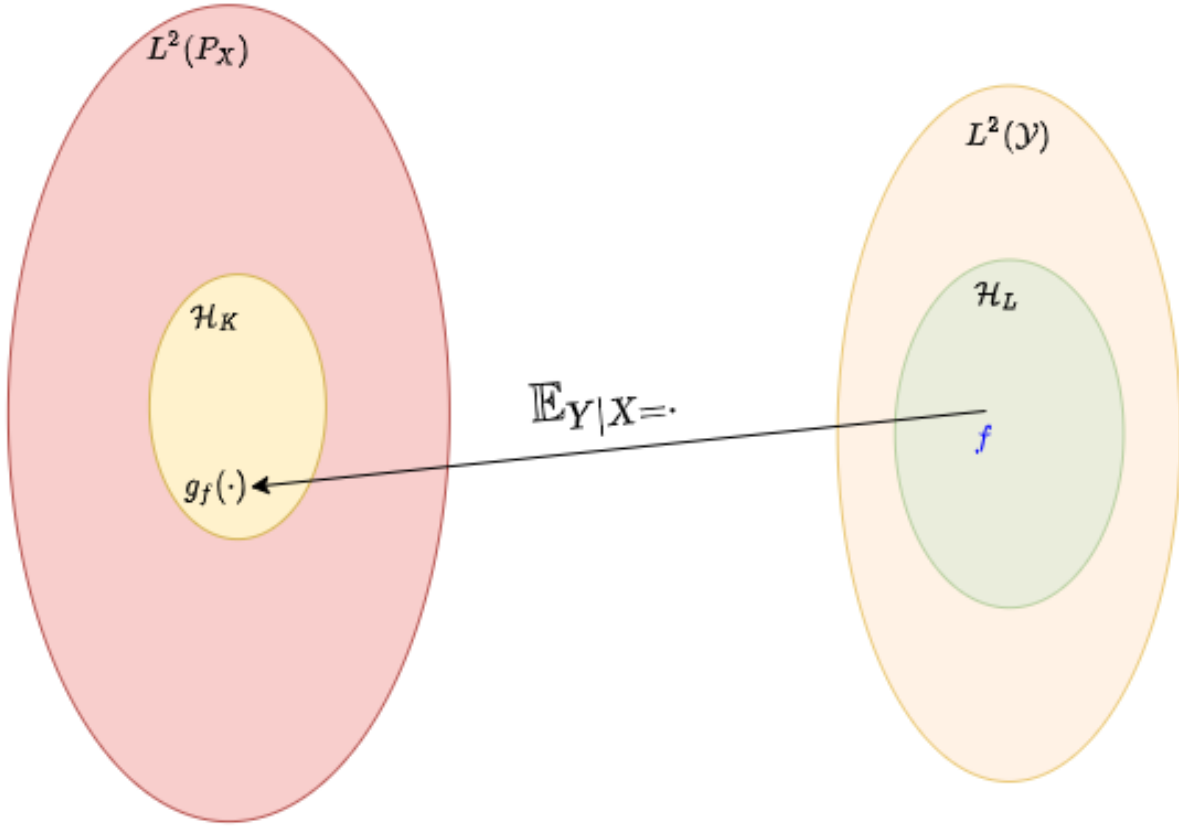
Existence of the CME

- Existence of the CME implicitly requires that for all $f \in \mathcal{H}_L$, $g_f(\cdot) \equiv \mathbb{E}[f(Y)|X = \cdot] \in \mathcal{H}_K$



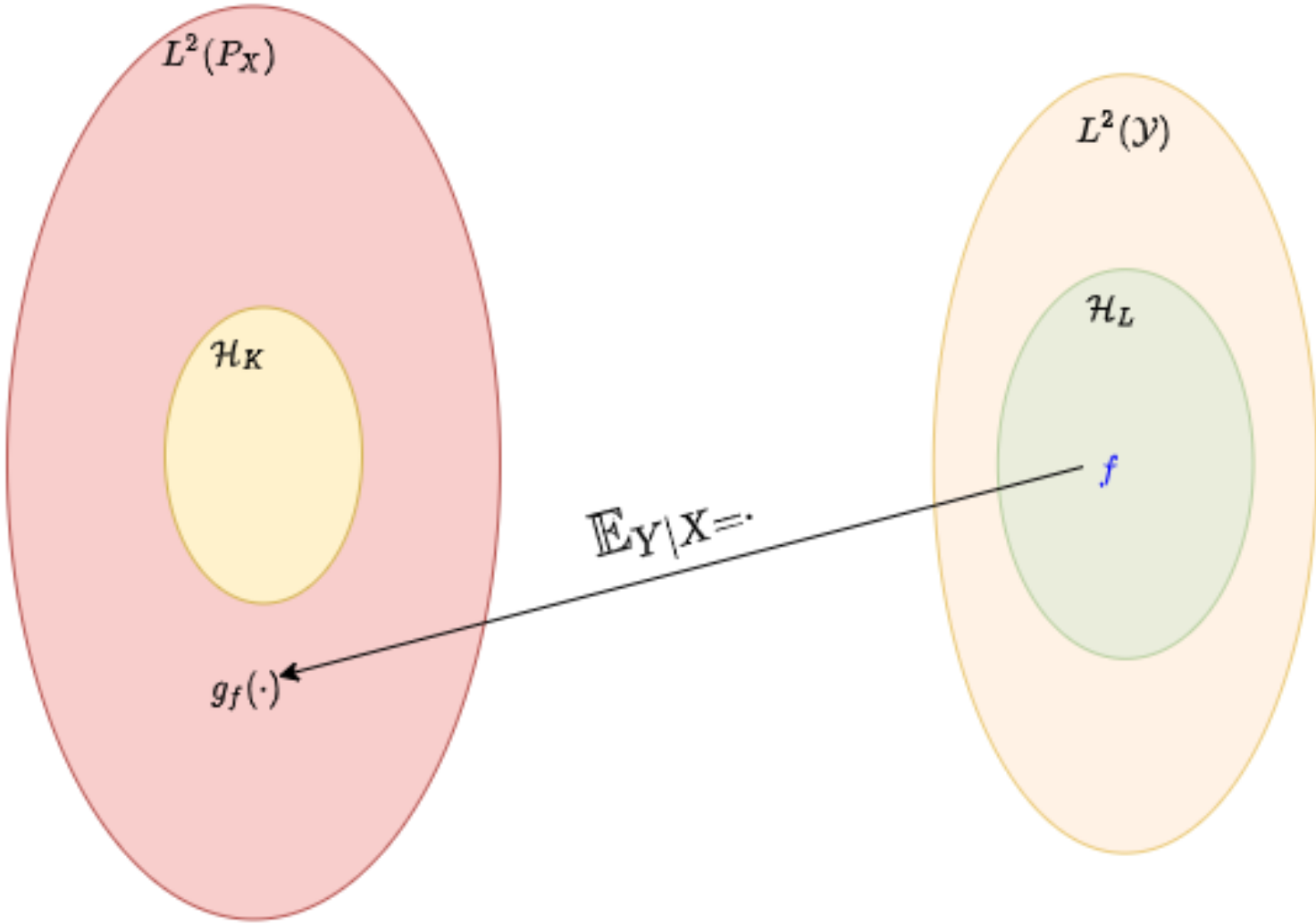
Existence of the CME

- Existence of the CME implicitly requires that for all $f \in \mathcal{H}_L$, $g_f(\cdot) \equiv \mathbb{E}[f(Y)|X = \cdot] \in \mathcal{H}_K$

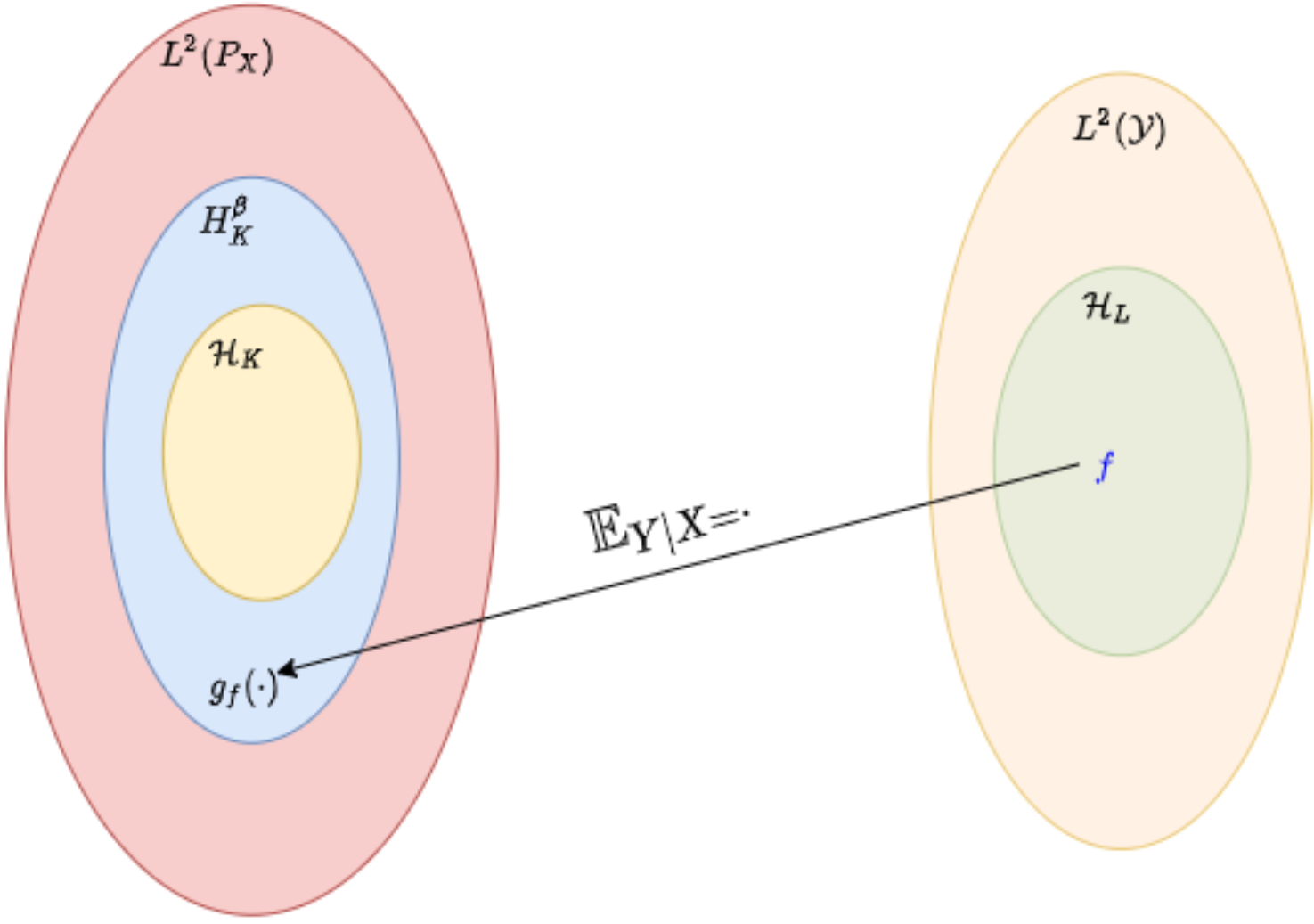


- This is a *very* strong assumption and violated in several common scenarios [4, 6]

Misspecification: $g_f \notin \mathcal{H}_K$



Misspecification: $g_f \notin \mathcal{H}_K$



$$\mathcal{H}_K^\beta \cong [L^2(P_X), \mathcal{H}_K]_{\beta,2}$$

Data-Driven Learning

- In practical settings, cannot compute population CME
- Interested in estimating CME from i.i.d data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ via solution to regularized LS problem [3]:

$$\hat{C}_{Y|X} \in \arg \min_{T:\mathcal{H}_K \to \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^{n} ||l(y_i, \cdot) - T[k(x_i, \cdot)]||_L^2 + \lambda ||T||_{\text{HS}}^2$$

Data-Driven Learning

- In practical settings, cannot compute population CME
- Interested in estimating CME from i.i.d data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ via solution to regularized LS problem [3]:

$$\hat{C}_{Y|X} \in \arg \min_{T:\mathcal{H}_K \to \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^{n} ||l(y_i, \cdot) - T[k(x_i, \cdot)]||_L^2 + \lambda ||T||_{\text{HS}}^2$$

- **Goal**: Establish rates for $||C_{Y|X} - \hat{C}_{Y|X}||_\gamma$ (in various norms) as $n \to \infty$ in the misspecified setting that are:

Data-Driven Learning

- In practical settings, cannot compute population CME
- Interested in estimating CME from i.i.d data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ via solution to regularized LS problem [3]:

$$\hat{C}_{Y|X} \in \arg \min_{T:\mathcal{H}_K \to \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^{n} ||l(y_i, \cdot) - T[k(x_i, \cdot)]||_L^2 + \lambda ||T||_{\text{HS}}^2$$

- **Goal**: Establish rates for $||C_{Y|X} - \hat{C}_{Y|X}||_\gamma$ (in various norms) as $n \to \infty$ in the misspecified setting that are:
  - adaptive to the "degree" of misspecification (β)

Data-Driven Learning

- In practical settings, cannot compute population CME
- Interested in estimating CME from i.i.d data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ via solution to regularized LS problem [3]:

$$\hat{C}_{Y|X} \in \arg \min_{T:\mathcal{H}_K \to \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^{n} ||l(y_i, \cdot) - T[k(x_i, \cdot)]||_L^2 + \lambda ||T||_{\text{HS}}^2$$

- **Goal**: Establish rates for $||C_{Y|X} - \hat{C}_{Y|X}||_\gamma$ (in various norms) as $n \to \infty$ in the misspecified setting that are:
  - adaptive to the "degree" of misspecification (β)
  - characterize the interplay between kernel complexity and CME regularity

Contributions

- Establish first explicit learning rates for the conditional mean embedding under misspecification in various Sobolev operator norms

Contributions

- Establish first explicit learning rates for the conditional mean embedding under misspecification in various Sobolev operator norms
  - Our framework notably does not impose the conventional Hilbert-Schmidt criterion [2, 7, 1] on the target CME, and only requires it to be bounded, with respect to some interpolation space.

Contributions

- Establish first explicit learning rates for the conditional mean embedding under misspecification in various Sobolev operator norms
  - Our framework notably does not impose the conventional Hilbert-Schmidt criterion [2, 7, 1] on the target CME, and only requires it to be bounded, with respect to some interpolation space.
- Demonstrate that in certain parameter regimes, these rates translate to uniform convergence rates for the mean embeddings in $\mathcal{H}_L$ of $P(Y|x)$ over all $x \in \mathcal{X}$

Visit our Poster!

Poster 3592 on March 28 (1:15 - 2:45 EDT)

References I

[1] Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. "Kernel dimension reduction in regression". In: *The Annals of Statistics* 37.4 (2009), pp. 1871–1905.

[2] Kenji Fukumizu et al. "Kernel measures of conditional dependence.". In: *NIPS*. Vol. 20. 2007, pp. 489–496.

[3] Steffen Grünewälder et al. "Conditional mean embeddings as regressors". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. 2012, pp. 1803–1810.

[4] Ilja Klebanov, Ingmar Schuster, and TJ Sullivan. "A rigorous theory of conditional mean embeddings". In: *SIAM Journal on Mathematics of Data Science* 2.3 (2020), pp. 583–606.

[5] Krikamol Muandet et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

References II

[6] Junhyung Park and Krikamol Muandet. "A measure-theoretic approach to kernel conditional mean embeddings". In: *Advances in Neural Information Processing Systems* 33 (2020).

[7] Le Song et al. "Hilbert space embeddings of conditional distributions with applications to dynamical systems". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 961–968.