Kernel SPO

Hong-Seok Choe

1 Idea Sketch

Consider the following toy example where we try to learn the RHS b using some contextual information:

$$\begin{array}{ll} \max & w_1 + 2w_2 \\ \text{s.t.} & w_1 + w_2 \leq b \\ & 0 \leq w_1, w_2 \leq 1 \end{array} \tag{P}$$

Let $\hat{b}$, $\tilde{b}$ denote the predicted b and realized b, resp. We want $\hat{b}$ to lie within the feasible region once the true value $\tilde{b}$ is revealed. Here, notice that the post-realization feasibility of (P) is captured by $1\{\hat{b} \leq \tilde{b}\}$. Indeed, in most predict-then-optimize problem settings, we want to ensure feasibility with some guarantees i.e., minimize P(infeasibility).

Obs. Suppose we have a black-box prediction scheme that returns $\hat{b}_n$ converging to $\tilde{b}$ at the rate of $\mathcal{O}(\frac{1}{\sqrt{n}})$. That is, $\|\hat{b}_n - \tilde{b}\| \leq \mathcal{O}(\frac{1}{\sqrt{n}})$. What can we say about the worst-case P(infeasibility)? If $\hat{b}_n = \tilde{b} + \frac{1}{\sqrt{n}}$ so that the estimates always overpredicts the true value $\tilde{b}$, then we have almost sure infeasibility $P(\hat{b}_n - \tilde{b} > 0) = 1$. This motivates using $\epsilon-$infeasibility where ϵ is the desired level of infeasibility that we want to achieve.

Let $\mathcal{B}$ denote the space of possible values of b and consider the loss from infeasibility $\ell_{\text{infeasibility}} : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$. In our toy example,

$$\ell_{\text{infeasibility}}(\hat{b}, b) = 1\{\hat{b} > b\}$$

so the risk here is just $\mathbb{E}[1\{\hat{b} > b\}] = P$(infeasibility). We may want to formulate the ERM problem incorporating the infeasibility risk.

One naïve way is to set multiple objectives, which can be easily represented using weighted sum of loss functions:

$$w_{\text{regret}} \underbrace{\ell_{\text{regret}}(\hat{b}, b)}_{①} + w_{\text{infeasibility}} \underbrace{\ell_{\text{infeasibility}}(\hat{b}, b)}_{②} \tag{1}$$

① Canonical regret:

$$\left| \begin{array}{lcc} \min & c^\top w & \\ \text{s.t.} & w \in S(b) & \\ & w \in S_0 & \end{array} \right. - \left. \begin{array}{lcc} \min & c^\top w \\ \text{s.t.} & w \in S(\hat{b}) \\ & w \in S_0 \end{array} \right|$$

- How to obtain feasible $S(\hat{b})$?
- Tractable surrogate loss?
- Dual form and potential relationship with the SPO loss

② P(infeasibility):

$$1\{\hat{b} \notin S^{-1}(b)\} \text{ where } S^{-1}(b) := \{w : w \in S_0 \cap S(b)\}$$

- Tractable reformulation/surrogate loss

– Can we view this as a regularization such that it allows reformulation w.r.t some function class $\mathcal{H}$

$$\begin{array}{ll} \min & ① + (\text{strongly convex}) \\ \text{s.t.} & ② \leq \epsilon \end{array} \qquad \begin{array}{ll} \min\limits_{f \in \mathcal{H}} & L_n(f) \\ \text{s.t.} & \|f\|_{\mathcal{H}}^2 \leq \epsilon \end{array}$$

Key Idea: RKHS embedding that captures $P(\text{infeasibility})$

2 RHKS Overview (From STAT300B Lecture Notes)

2.1 Motivation for Kernel Methods

Linear models are easy to solve, where the prediction is a function of the inner product $\langle w, x \rangle$ between a weight vector $w \in \mathbb{R}^d$ and an input $x \in \mathbb{R}^d$. Consider replacing $\langle w, x \rangle$ with $\langle w, \phi(x) \rangle$ where $\phi : \mathcal{X} \to \mathbb{R}^d$ is an arbitrary feature map. By making $\phi(x)$ complex enough, we can represent non-linear functions.

Consider ERM:

$$\hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (y^i - \langle w, \phi(x^i) \rangle)^2$$

$$\nabla \hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} (y^i - \langle w, \phi(x^i) \rangle) \phi(x^i)$$

Notice any gradient methods will involve linear combinations of feature vectors of the data points

$$w = \sum_{i=1}^{n} \alpha_i \phi(x^i)$$

and we can make predictions using inner products of feature maps

$$\langle w, \phi(x) \rangle = \sum_{i=1}^{n} \alpha_i \langle \phi(x^i), \phi(x) \rangle$$

This motivates the kernel trick. Observe that $\langle x_i, x_j \rangle$ can be replaced with $k(x_i, x_j)$ for any valid kernel function k. Furthermore, we can swap any kernel function with any other kernel function.

Example (Quadratic Features)

- Input $x \in \mathbb{R}^b$
- Feature map ($\mathcal{O}(b^2)$ dimension)

$$\phi(x) = [x_1^2, \dots x_b^2, \sqrt{2}x_1x_2, \dots \sqrt{2}x_1x_b, \dots, \sqrt{2}x_{b-1}x_b]$$

- $\langle w, \phi(x) \rangle$ takes $\mathcal{O}(b^2)$ time
- $\langle \phi(x), \phi(x') \rangle = \langle x, x' \rangle^2$ takes $\mathcal{O}(b)$ time. For n data points, the total complexity is $\mathcal{O}(bn)$

Key takeaway: Kernel offers a computationally efficient way of working with high dimensional $\phi(x)$ implicitly.

2.2 Kernels: Definition and Examples

Definition 1 (Kernels). *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite kernel (or simply, a kernel) if and only if for every finite set of points $x_1, \dots x_n \in \mathcal{X}$, the kernel matrix $K \in \mathbb{R}^{n \times n}$ is defined by $K_{ij} = k(x_i, x_j)$.*

Example (Gaussian Kernel)

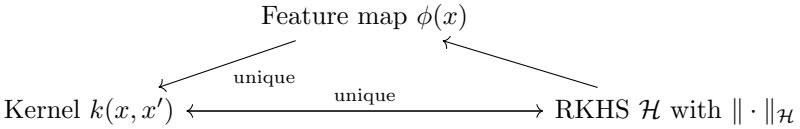
$$k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$$

- The bandwidth parameter σ^2 governs how smooth the functions should be (larger $\sigma^2 \implies$ more smoothness)
- The corresponding dimensionality is infinite

Let $k_1, \ldots k_m$ be valid kernel functions defined over $\mathcal{X}$, and let $\alpha_1, \ldots, \alpha_m$ be nonnegative coefficients. The following are valid kernels:

- $k(x, z) = \sum_{i=1}^m \alpha_i k_i(x, z)$ (Closed under nonnegative linear multiplication)
- $k(x, z) = \prod_{i=1}^m k_i(x, z)$ (Closed under multiplication)
- $k(x, z) = k_1(f(x), f(z))$ for any function $f : \mathcal{X} \to \mathbb{R}$
- $k(x, z) = g(x)g(z)$ for any function $g : \mathcal{X} \to \mathbb{R}$
- If $|\mathcal{X}| = n$, $k(x, z) = x^\top A^\top A z$ for any matrix $A \in \mathbb{R}^{m \times n}$

2.3 Three Views of Kernel Methods



- Feature map ϕ: maps from a data point $x \in \mathcal{X}$ to an element of an inner product space, or the feature vector (properties of single data points)
- Kernel k: takes two data points $x, x' \in \mathcal{X}$ and returns a real number (a similarity measure)
- RKHS $\mathcal{H}$: a set of functions $f : \mathcal{X} \to \mathbb{R}$ equipped with a norm $\|\cdot\|_{\mathcal{H}}$ for measuring the complexity of functions (prediction functions f)

Definition 2 (Hilbert Space)**.** *A Hilbert space $\mathcal{H}$ is a complete vector space with an inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that satisfies the following properties:*

- *Symmetry: $\langle f, g \rangle = \langle g, f \rangle$*
- *Linearity: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$*
- *Positive definiteness: $\langle f, f \rangle \geq 0$ with equality only if $f = 0$*

The inner product gives a canonical norm: $\|f\|_{\mathcal{H}} \stackrel{def}{=} \sqrt{\langle f, f \rangle}$

Definition 3 (Feature Map)**.** *Given a Hilbert space $\mathcal{H}$, a feature map $\phi : \mathcal{X} \to \mathcal{H}$ takes input $x \in \mathcal{X}$ to infinite feature vectors $\phi(x) \in \mathcal{H}$*

Theorem 1 (Feature Map to Kernel)**.** *Let $\phi : \mathcal{X} \to \mathcal{H}$ be a feature map. Then, $k(x, x') \stackrel{def}{=} \langle \phi(x), \phi(x') \rangle$ is a valid kernel.*

Theorem 2 (Kernel to Feature Map)**.** *For every kernel k, there exists a hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \phi(x), \phi(x') \rangle$*

2.4 Reproducing Kernel Hilbert Space

Not all Hilbert space over functions $f : \mathcal{X} \to \mathbb{R}$ is suitable for ML. We want to learn f which allows us to perform pointwise evaulations (i.e., predictions). RKHSes contain functions f s.t. function evaluations are bounded linear operators.

Definition 4 (Bounded Functional)**.** *Given a Hilbert space $\mathcal{H}$, a functional $L : \mathcal{H} \to \mathbb{R}$ is bounded if and only if there exists $M < \infty$ s.t.*

$$|L(f)| \leq M\|f\|_{\mathcal{H}} \text{ for all } f \in \mathcal{H}$$

Definition 5 (Evaluation Functional)**.** *Let $\mathcal{H}$ be a Hilbert space consisting of functions $f : \mathcal{X} \to \mathbb{R}$. For each $x \in \mathcal{X}$, we define the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$ as $L_x(f) \stackrel{def}{=} f(x)$*

Definition 6 (Reproducing Kernel Hilbert Space)**.** *A RHKS $\mathcal{H}$ is a Hilbert space over functions $f : \mathcal{X} \to \mathbb{R}$ s.t. for each $x \in \mathcal{X}$, the evaluation functional L_x is bounded.*

Theorem 3 (RHKS to Kernel)**.** *Every RKHS $\mathcal{H}$ over functions $f : \mathcal{X} \to \mathbb{R}$ defines a unique kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called the reproducing kernel of $\mathcal{H}$*

Proof. From Riesz representation theorem, all bounded linear functionals L on a Hilbert space can be expressed as $L(f) = \langle R, f\rangle$ for a unique $R \in \mathcal{H}$. Taking R_x as a unique representer of L_x, we thus have a reproducing property:

$$f(x) = \langle R_x, f\rangle \text{ for all } f \in \mathcal{H}$$

Consider $k(x, x') \stackrel{\text{def}}{=} R_x(x')$ and $\phi(x) \stackrel{\text{def}}{=} R_x$. We get the valid kernel. □

Theorem 4 (Moore-Aronszajn)**.** *For every kernel k, there exists a unique RKHS $\mathcal{H}$ with reproducing kernel k.*

Diagram Revisited:

- $f(x) = \sum_{i=1}^{\infty} \alpha_i k(x_1, x)$, $R_x = k(x, \cdot)$: Moore-Aronszajn establishes connection between kernels and RKHSes
- $\phi(x) = R_x$: can set feature map (not unique) to map x to the representer of x in RKHS
- $k(x, x') = \langle \phi(x), \phi(x')\rangle$: every kernel k corresponds to some inner product (via RKHS) and vice-versa

2.5 Learnings with Kernels

Kernels k provide a space of functions $\mathcal{H}$; this is the hypothesis class. Generally, a learning problem can be formulated as the following optimization problem:

$$f^* = \arg\min_{f \in \mathcal{H}} L\left(\{(x_i, y_i, f(x_i)\}_{i=1}^n\right) + Q\left(\|f\|_{\mathcal{H}}^2\right)$$

where

- $L : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R})^n \to \mathbb{R}$ is an arbitrary loss function on n data points
- $Q : [0, \infty) \to \mathbb{R}$ is a strictly increasing function (regularization)

The optimization problem seems daunting, but the representer theorem allows us to rewrite the minimizers as a linear combination of the kernel functions evaluated at the training points.

Theorem 5 (Representer Theorem)**.** *Let V denote the span of the representers of the training points:*

$$V \stackrel{def}{=} span\left(\{k(x_i, \cdot) : i = 1, \ldots, n\}\right) = \left\{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : \alpha \in \mathbb{R}^n\right\}$$

Then all minimizers f^* satisfy $f^* \in V$

Remarks:

- L need not be convex
- The representer theorem tells us that α's exist, but finding α's depends on L and the regularizer

2.6 Other Related Keywords

- Kernel approximations: deals with inexpensive computation
  - Random features (data-independent): write the kernel function as an integral, and using Monte Carlo approximations of this integral.
  - Nyström method (data-dependent): sample a subset of the n points and use these points to approximate the kernel matrix.
- Universality: deals with general purpose kernel k, in the sense that k can be used to solve any learning problem given sufficient data

3 Kernel Embedding

Kernels can also be used to represent and answer questions about probability distributions without having to explicitly estimate them.

Definition 7 (Maximum Mean Discrepancy)**.** *Assume P and Q are probability measures defined on some locally compacy Hausdorff space $\mathcal{X}$ (e.g., $\mathbb{R}^b$). The maximum mean discrepancy for some set of functions $\mathcal{F}$ is defined as*

$$D(P, Q, \mathcal{F}) \stackrel{def}{=} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{x \in P}[f(x)] - \mathbb{E}_{x \in Q}[f(x)] \right)$$

Motivating Example We want to test whether two probability distributions P and Q are the same by only observing expectations under the distributions. Given a distribution P, we can look at various moments of the distribution $\mathbb{E}_{x \sim P}[f(x)]$ for various functions f. Notice that the problem boils down to finding $\mathcal{F}$ s.t. $D(P, Q, \mathcal{F}) = 0 \implies P = Q$. For instance, if we knew P and Q were Gaussian, then it suffices to take $\mathcal{F} = \{x : x \to x, x \to x^2\}$, since the first two moments define a Gaussian distribution. Without any assumptions, however, we need a much larger class of functions $\mathcal{F}$ to work with general distributions P and Q.

Theorem 6 (Dudley, 1984)**.** *If $\mathcal{F} = C_0(\mathcal{X})$ (the set of all continuous bounded functions), then $D(P, Q, \mathcal{F}) = 0 \implies P = Q$*

Theorem 7 (Steinwart, 2001)**.** *Let $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, where $\mathcal{H}$ is the RKHS defined by a universal kernel k. Then, $D(P, Q, \mathcal{F}) = 0 \implies P = Q$*

3.1 Computing $D(P, Q, \mathcal{F})$

So $D(P, Q, \mathcal{F})$ contains sufficient information for testing whether two distributions are equal. We can actually compute $D(P, Q, \mathcal{F})$ in a closed form using properties of RKHS.

- First, by the reproducing property and linearity of the inner product

  $$\mathbb{E}_{x \in P}[f(x)] = \mathbb{E}_{x \in P}[\langle k(\cdot, x), f \rangle] = \Big\langle \underbrace{\mathbb{E}_{x \in P}[k(x, \cdot)]}_{\stackrel{\text{def}}{=} \mu_P}, f \Big\rangle$$

  Here, $\mu_P \in \mathcal{H}$ is the RKHS embedding of the probability distribution P

- Note that we can write

  $$D(P, Q, \mathcal{F}) = \sup_{f \in \mathcal{F}} \langle \mu_P - \mu_Q, f \rangle = \|\mu_P - \mu_Q\|_{\mathcal{H}}$$

  where the sup is obtained by setting f to be a unit vector in the direction of $\mu_P - \mu_Q$

- Rewriting the squared norm in terms of kernel evaluations,

$$\begin{aligned}\|\mu_P - \mu_Q\|_{\mathcal{H}}^2 &= \langle \mu_P, \mu_P \rangle - \langle \mu_P, \mu_Q \rangle - \langle \mu_Q, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle \\ &= \mathbb{E}_{x,x'} \langle \phi(x), \phi(x') \rangle - \mathbb{E}_{x,y} \langle \phi(x), \phi(y) \rangle - \mathbb{E}_{y,x} \langle \phi(y), \phi(x) \rangle + \mathbb{E}_{y,y'} \langle \phi(y), \phi(y') \rangle \\ &= \mathbb{E}_{P \times P}[k(x, x')] - \mathbb{E}_{P \times Q}[k(x, y)] - \mathbb{E}_{Q \times P}[k(y, x)] + \mathbb{E}_{Q \times Q}[k(y, y')]\end{aligned}$$

Finite sample case with $x_1, \ldots, x_n \sim P$ and $y_1, \ldots, y_n \sim Q$ (independently drawn)

- Unbiased estimator of $D(P, Q, \mathcal{F})$ as a U-statistic:

$$\hat{D}_n(P, Q, \mathcal{F}) = \frac{1}{\binom{n}{2}} \sum_{i<j} [k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)]$$

- To test the null hypothesis $P = Q$, we need to know the approximate distribution of $\hat{D}_n(P, Q, \mathcal{F})$, a random variable that is a function of the data points (Under null, $\hat{D}_n \to 0$ as $n \to \infty$)
  (a) We can derive a finite sample guarantee (i.e., complexity bound of $|\hat{D}_n(P, Q, \mathcal{F}) - D(P, Q, \mathcal{F})|$)
  (b) We can show that $\hat{D}_n(P, Q, \mathcal{F})$ is asymptotically normal with some variance, and use normal approximation

3.2 Conditional Kernel Embedding (Song et al. 2013, Muandet et al. 2020)

For a random variable in domain $\mathcal{X}$ distribution $P(X)$, suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the positive definite kernels with corresponding RKHS $\mathcal{H}_{\mathcal{X}}$, the kernel embedding of a kernel k for $\mathcal{X}$ is defined as

$$\mu_X = \mathbb{E}_X[k(\cdot, x)] = \int k(\cdot, x) dP(x) \in \mathcal{H}_{\mathcal{X}} \implies \hat{\mu}_X = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

Remark:

- We can obtain kernel mean shrinkage estimator (cf. James-Stein estimator)

$$\hat{\mu}_\alpha := \alpha f^* + (1 - \alpha) \hat{\mu}_P$$

  by solving

$$\hat{\mu}_P = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|g - k(x_i, \cdot)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{H}}^2$$

  Duality between Stein estimation in statistics and Tikhonov regularization?

For two random variable X and Y, suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are respectively the positive definite kernels with corresponding RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. The kernel embedding for the marginal distribution $P(Y \mid X = x)$ is:

$$\mu_{Y|x} = \mathbb{E}_Y[l(\cdot, y)|x] = \int l(\cdot, y) dP(y \mid x) \in \mathcal{H}_{\mathcal{Y}}$$

Then for the conditional probability $P(Y \mid X)$, the kernel embedding is defined as a conditional operator $C_{Y|X} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ that satisfies

$$\mu_{Y|x} = C_{Y|X} k(x,) \implies \hat{\mu}_{Y|x} = \hat{C}_{Y|X} \phi(x)$$

If we have a data set $\{x_i, y_i\}_{i=1}^m$, which are i.i.d drawn from $P(X, Y)$, the conditional kernel embedding operator can be estimated by:

$$\hat{C}_{Y|X} = \Psi (K + \lambda I)^{-1} \Phi$$

where $\Psi = (l(y_1, \cdot), \cdots, l(y_m, \cdot))$ and $\Phi = (k(x_1, \cdot), \cdots, k(x_m, \cdot))$ are implicitly formed feature matrix, K is the Gram matrix for samples from variable X, i.e., $(K)_{ij} = k(x_i, x_j)$, and λ is a regularization parameter to avoid overfitting.

Remark:

- The definition of conditional kernel embedding provides a way to measure probability $P(Y \mid X)$ as an operator between the spaces $\mathcal{H}_\mathcal{Y}$ and $\mathcal{H}_\mathcal{X}$.
- Reproducing property:

$$\mathbb{E}_P[f(x)] = \langle f, \mu_P \rangle, \quad \forall f \in \mathcal{H}_\mathcal{X}$$
$$\mathbb{E}_{Y|x}[g(Y) \mid X = x] = \langle g, \mu_{Y|x} \rangle, \quad \forall g \in \mathcal{H}_\mathcal{Y}$$

- The conditional mean embedding operator can be alternatively formulated using function-valued least squares regression problem:

$$\hat{C}_{Y|X} = \underset{C:\mathcal{F}\to\mathcal{F}}{\arg\min} \sum_{i=1}^{m} \|\psi(y_i) - C\phi(x_i)\|_\mathcal{F}^2 + \lambda \|C\|_{HS}^2$$

  where $\|C\|_{HS}$ denotes the Hilbert-Schmidt norm (generalized Frobenius norm) of operator C
- Consider the conditional expectation mapping $h \to \mathbb{E}[h(Y) \mid X = x]$. This can be represented as

$$\mathbb{E}[h(Y) \mid X = x] = \langle h, \mu(x) \rangle_{\mathcal{H}_\mathcal{Y}}$$

  $\mu(x) \in \mathcal{H}_\mathcal{Y}$ is the conditional mean embedding of $P(Y \mid X = x)$ and

$$\hat{\mu}(x) := \sum_{i=1}^{n} \alpha_i(x) l(y_i, \cdot)$$

  where $\alpha_i(x) = \sum_{j=1}^{n} W_{ij} K(x_j, x)$, $W := (K + \lambda n I)^{-1}$, and $K = (K(x_i, x_j))_{ij}$.
  This can viewed as a vector-valued regression problem (Grünewälder et al. 2012) with the training data $\{(x_i, l(y_i, \cdot))\}_{i=1}^n := \{(x_i, z_i)\} \in \mathcal{X} \times \mathcal{H}_\mathcal{Y}$.

$$\hat{\mathcal{E}}_\lambda(f) = \sum_{i=1}^{n} \|z_i - f(x_i)\|_{\mathcal{H}_\mathcal{Y}}^2 + \lambda \|f\|^2$$

  Kernel Objective Prediction Method (Bertsimas and Koduri, 2022) uses similar approach:
  A natural optimization problem for the conditional mean embedding is to find a function $\mu : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ that minimizes the following objective

$$\mathcal{E}[\mu] = \sup_{\|g\|_{\mathcal{H}_\mathcal{Y}} \le 1} \mathbb{E}_X \left[\left(\mathbb{E}_Y[g(Y) \mid X] - \langle g, \mu(X) \rangle \right)^2 \right]$$

  We do not observe $\mathbb{E}_Y[g(Y) \mid X]$, so come up with surrogate loss function that upper bounds $\mathcal{E}[\mu]$

$$\mathcal{E}_s[\mu] = \mathbb{E}_{(X,Y)} \left[\|l(Y, \cdot) - \mu(X)\|^2 \right]$$

  whose empirical counterpart is

$$\hat{\mathcal{E}}_s[\mu] = \sum_{i=1}^{n} \|l(y_i, \cdot) - \mu(x_i)\|_{\mathcal{H}_\mathcal{Y}}^2 + \lambda \|\mu\|^2$$

3.3 Recovering Information from Kernel Embeddings

Given a kernel mean embedding μ_P, can we recover essential properties of P from μ_P?

1. Distributional Pre-Image Problem:
   Consider the case when we only have access to the estimate $\hat{\mu}_X$ which lies in a high-dimensional feature space. Let P_θ be an arbitrary distribution parametrized by θ and μ_{P_θ} be its mean embedding in $\mathcal{F}$. One can find P_θ by the following minimization problem

$$\theta^* = \arg\min_{\theta \in \Theta} \|\hat{\mu}_X - \mu_{P_\theta}\|_\mathcal{H}^2$$

subject to appropriate constraints on the parameter vector θ.

Now we consider more generalized version, the reduced set problem :

Assume we are given a function $g \in \mathcal{H}$ as a linear combination of the images of input points $x_i \in X$, i.e., $g = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, which is the kernel mean embedding of the finite signed measure $\nu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ whose supports are the points $x_1, \ldots, x_n$. That is, $g = \int \phi(y) d\nu(y)$. Given the reduced set vector $z_1, \ldots, z_m$ where $m \ll n$, the reduced set problem amounts to finding another finite signed measure $\mu = \sum_{j=1}^{m} \beta_j \delta_{z_j}$ whose supports are $z_1, \ldots, z_m$ that approximates well the original measure ν. From the distributional pre-image problem, the reduced set methods can be viewed as an approximation of a finite signed measure by another signed measure whose supports are smaller.

2. Kernel Herding:

   Herding directly generates pseudo-samples in a fully deterministic fashion and in a way that asymptotically matches the empirical moments of the data. Kernel herding algorithm extends the herding algorithm in the following way:

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \langle w_t, \phi(x) \rangle$$
$$w_{t+1} = w_t + \mathbb{E}_{x \sim P}[\phi(x)] - \phi(x_{t+1})$$

   Under technical assumptions, this can be seen to greedily minimize the squared error

$$\mathcal{E}_T^2 := \left\| \mu_P - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \right\|_{\mathcal{H}}^2$$

   Equivalence between Frank-Wolfe method solving the following optimization problem:

$$\min_{g \in \mathcal{M}} J(g) = \frac{1}{2} \|g - \mu\|^2$$

   Here, $\mathcal{M}$ is a marginal polytope and μ is the trivial solution. The Frank-Wolfe iterates are

$$\bar{g}_{t+1} \in \arg\min_{g \in \mathcal{M}} \langle g_t - \mu, g \rangle$$
$$g_{t+1} = (1 - \rho_t) g_t + \rho_t \bar{g}_{t+1}$$

3.4 Other Related Keywords

- Cross Covariance Operators:

  $C_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ can be defined in termes of the tensor product $\varphi(Y) \otimes \phi(X)$ in a tensor product feature space $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ as

$$C_{YX} := \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] = \mu_{P_{YX}}$$

  and the covariance is simply

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathrm{Cov}[g(Y), f(X)]$$

  The integral expression for C_{YX} is

$$(C_{YX} f)(\cdot) = \int_{\mathcal{X} \times \mathcal{Y}} l(\cdot, y) f(x) dP_{XY}(x, y)$$

- Kernel Dependency Measures:

  When the dependence is non-linear, one of the most successful nonparametric measures is the Hilbert Schmidt Independence Criterion (HSIC). Let $\mathcal{H}$ and $\mathcal{G}$ be separable RKHSs on X and Y with reproducing kernels k and l, respectively.

$$HSIC(\mathcal{H}, \mathcal{G}, k, l) = \|C_{XY}\|_{HS}^2$$

4 Related Works

4.1 Data-Driven Optimization: A RKHS Approach (Bertsimas and Koduri, 2022)

Goal: $\min_w \mathbb{E}\left[c(w;\xi) \mid X = x_0\right]$

- Bertsimas & Kallus: Solve

$$\min_w \sum_{i=1}^{n} W(x^i, x^0) c(w;\xi^i)$$

where $W(x^i, x^0)$ are closeness measures that can be viewed as weights given on the data points (locally) that capture conditional distribution of ξ given $X = x_0$

- Bertsimas & Koduri: Globally learn objective values via kernels without resorting to conditional distribution (Note: the kernelized formulations work because of the Representer theorem)

  1. Kernel Objective Prediction Method: Solve

$$\min_{h(w,\cdot)} \mathbb{E}\left[\left(c(w;\xi) - h(w,x)\right)^2\right]$$

  Empirical version:

$$\min_{h(w,\cdot)\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \left(c(w;\xi^i) - h(w,x^i)\right)^2 + \lambda\|h\|_{\mathcal{H}}^2$$

  Kernelize: $h(w,\cdot) = \sum_{j=1}^{n} \alpha^j(w) K(x^j,\cdot)$, where $\alpha^j(w)$ are scalars.

  2. Kernel Optimizer Prediction Method: Solve

$$\min_{w(\cdot)} \mathbb{E}\left[c(w(x),\xi)\right]$$

  Empirical version:

$$\min_{w^1(\cdot),\ldots,w^d(\cdot)\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} c(w^1(x^i),\ldots,w^d(x^i);\xi^i) + \lambda\sum_{t=1}^{d}\|w^t\|_{\mathcal{H}}^2$$

  Kernelize: $\hat{w}^t = \sum_{i=1}^{n} K(x^i,\cdot)\alpha_i^t$
  Obs: Formulation is very similar to that of ICEO:

$$\begin{aligned} \min_{f\in\mathcal{H}} \quad & \frac{1}{n}\sum_{i=1}^{n} c(w_i,\xi^i) + \rho\phi(w_i) \\ \text{s.t.} \quad & w_i = w_\rho(f(x^i)) \end{aligned}$$

4.2 Solving Chance-Constrained Optimization Under Nonparametric Uncertainty Through Hilbert Space Embedding (Gopalakrishnan et al. 2019)

Setting: Chance-constrained optimization

$$\begin{aligned} \min \quad & g(u) \\ \text{s.t.} \quad & P(f(\xi,u)\le 0) \ge \epsilon \\ & u \in \mathcal{U} \end{aligned} \tag{CCO}$$

Here, ξ is a random variable and $\mathcal{U}$ is a (ground) convex set.

Consider distributions $f(\xi, u)$ parametrized by u, which is denoted $P_f(u)$. We can embed $P_f(u)$ in RKHS so that it allows representation using kernel functions $k(\cdot, \cdot)$:

$$\mu_{P_f(u)} := \sum_{i=1}^{n} \alpha_i k(f(\xi^i, u), \cdot)$$

where $\{\xi^i\}_{i=1}^n$ is our given data.

Define a desired distribution $f(\tilde{\xi}, u^*)$ s.t $P(f(\tilde{\xi}, u^*) \leq 0) \approx 1$ where $\tilde{\xi} \sim P_\xi^*$ (true distribution of ξ) and u^* is any solution of CCO. We can also embed P_f^* in RHKS, which is denoted by $\mu_{P_f^*}$.

Kernel embedding allows reformulating CCO in the following form using MMD

$$\begin{aligned} \min \ \ & \rho_1 \|\mu_{P_f(u)} - \mu_{P_f^*}\|_{\mathcal{H}}^2 + \rho_2 g(u) \\ \text{s.t. } & u \in \mathcal{U} \end{aligned}$$

which has a tractable form using kernel tricks.

Remarks:

- Note that we do not know the true embedded distribution $\mu_{P_f^*}$. This can be resolved by taking the "reduced set," a subset of samples that retains information of true distribution. Once the reduced set is determined, we can then embed to RKHS to obtain $\hat{\mu}_{P_f^*}$
- We can use $\epsilon-$infeasibility in lieu of the chance constraint
- CCO reformulation using MMD somewhat captures how empirical kernel mean (i.e., generalized moment) deviates from the true embedded mean where we have almost sure feasibility.

4.3 Worst-Case Risk Quantification Under Distributional Ambiguity Using Kernel Mean Embedding in Moment Problem (Zhu et al. 2020)

Setting: Kernel Mean Embedding Moment Problem (KME-MP)

$$\begin{aligned} (1) \ \max_{P,\mu} \quad & \int \ell(x) dP(x) \\ \text{s.t.} \quad & \|\mu - \mu_{\hat{P}}\|_{\mathcal{H}} \leq \epsilon \\ & \int \phi(x) dP(x) = \mu \\ & P \in \mathcal{P}, \mu \in \mathcal{H} \end{aligned} \qquad \begin{aligned} (2) \ \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i \ell(z_i) \\ \text{s.t.} \quad & \left\| \sum_{i=1}^{N} \alpha_i \phi(z_i) - \sum_{j=1}^{M} \frac{1}{M} \phi(x_j) \right\|_{\mathcal{H}} \leq \epsilon \\ & \Big(\text{or} \quad \alpha^\top K_z \alpha - 2 \frac{1}{M} \alpha^\top K_{zx} 1 + \frac{1}{M^2} 1^\top K_x 1 \leq \epsilon^2 \\ & \sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0 \Big) \end{aligned}$$

Here, our data is $\{x_j\}_{j=1}^M$, and z_i's are expansion points (obtained by sampling/reduced set type methods)

Convergence result: If P^* is the solution of (1) and α_i^* is the solution of (2) when $\{z_i\}_{i=1}^n$ are sampled from P^*, then

$$\sum_{i=1}^{N} \alpha_i^* \ell(z_i) \to \int \ell(x) dP(x) \text{ as } N \to \infty$$

Remarks:

- (2) can be reformulated with convex quadratic constraints using kernel tricks: $K_z := [k(z_i, z_j)]_{ij}$, $K_{zx} := [k(z_i, x_j)]_{ij}$, $K_x := [k(x_i, x_j)]_{ij}$ are gram matrices

- $\sum_{i=1}^N \alpha_i \phi(z_i)$ can be viewed as putting probability mass α_i on z_i. So this is "transporting" the mass that incurs the worst cost.
- E.g. evaluating the worst-case probability with ambiguity set $\mathcal{C}$ for the underlying distribution P:

$$\begin{aligned} (1') \ \max_{P} \quad & P(X \notin A) \\ \text{s.t.} \quad & P \in \mathcal{C} \subseteq \mathcal{P} \end{aligned} \qquad \begin{aligned} (2') \ \max_{\alpha} \quad & \sum_{i=1}^N \alpha_i 1\{z_i \notin A\} \\ \text{s.t.} \quad & \alpha^\top K_z \alpha - 2\frac{1}{M}\alpha^\top K_{zx} 1 + \frac{1}{M^2} 1^\top K_x 1 \leq \epsilon^2 \\ & \sum_{i=1}^N \alpha_i = 1, \alpha_i \geq 0 \end{aligned}$$

5 More Kernel Ideas

- Spectral filtering:

  Regularized least squares problem can be formulated in matrix form :

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \equiv (K + n\lambda I)c = Y$$

  because RKHS allows expressing the RLS estimator as $f_S^\lambda(X) = \sum_{i=1}^n c_i k(x, x_i)$ If the problem is ill-conditioned, and solving this RLS problem amounts to stabilizing a possibly ill-conditioned matrix inversion problem.

- Kernel DRO (Zhu et al. 2020):

  Empirical distribution $\hat{P} = \sum_{i=1}^N \frac{1}{N} \delta_{\xi_i}$ where $\{\xi_i\}_{i=1}^n$ are data samples. Variational duality can be established:

$$\begin{aligned} (P) \ & \min_{\theta} \sup_{d_{\mathcal{F}}(P, \hat{P}) \leq \epsilon} \int \ell(\theta, \xi) dP(\xi) \\ (D) \ & \min_{\theta, \lambda \geq 0, f_0 \in \mathbb{R}, f \in \mathcal{F}} f_0 + \frac{1}{N} \sum_{i=1}^n \lambda f(\xi_i) + \lambda \xi \\ & \text{s.t. } \ell(\theta, \xi) \leq f_0 + \lambda f(\xi), \quad \forall \xi \in \mathcal{X} \end{aligned}$$

- Shape constraints (e.g. directional monotonicity, convexity):

  Certain shape constraints on the prediction function can be represented with differential operators, thereby inducing a special RKHS.

- Equivalence with Gaussian process:

  GP induces minimizing norm in RKHS (Kimeldorf-Wahba correspondence)

6 TO-DO

- Surrogate loss for regret
- What do I want to prove (consistency results...)
- Coarsing output space

7 Related Work

7.1 Uncertain Convex Programs: Randomized Solutions and Confidence Levels (Calafiore and Campi, 2005)

Goal:

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t. } f(x,\delta) \le 0, \quad \delta \in \Delta \\ &\quad x \in \mathcal{X} \end{aligned} \tag{1}$$

Definition 8 (Violation Probability)**.** *Let $x \in \mathcal{X}$ be a candidate solution for (1). The probability of violation of x is defined as*

$$V(x) \doteq P\{\delta \in \Delta : f(x,\delta) > 0\}$$

Definition 9 ($\epsilon-$level Solution)**.** *Let $\epsilon \in [0,1]$. We say that $x \in \mathcal{X}$ is an ϵ-level robustly feasible solution if $V(x) \le \epsilon$.*

Definition 10 (Sampled Convex Program)**.** *Let $\delta^{(1)}, \ldots, \delta^{(N)}$ be N independent identically distributed samples extracted according to probability P. The sampled convex program derived from (1) is*

$$\begin{aligned} \mathrm{SCP}_N : \min\ & c^\top x \\ \text{s.t. } & f(x,\delta^{(i)}) \le 0, \quad i = 1, \ldots, N \\ & x \in \mathcal{X} \end{aligned}$$

Theorem 8. *Let $\hat{x}_N$ be the (unique) solution to SCP_N. Then,*

$$\mathbb{E}_{P^N}\left[V\left(\hat{x}_N\right)\right] \le \frac{n}{N+1}$$

where n is the size of x, and $P^N = P \times \cdots \times P$ is the probability measure in the space Δ^N of the multi-sample extraction $\delta^{(1)}, \ldots, \delta^{(N)}$.

Corollary 1. *Fix two real numbers $\epsilon \in [0,1]$ (level parameter) and $\beta \in [0,1]$ (confidence parameter) and let*

$$N \ge \frac{n}{\epsilon\beta} - 1$$

Then, with probability no smaller than $1-\beta$, the randomized problem SCP_N returns an optimal solution $\hat{x}_N$ which is ϵ-level robustly feasible.

Notice that SCP_N can be infeasible (i.e. $\cap_{i=1,\ldots,N}\left\{x : f\left(x,\delta^{(i)}\right) \le 0\right\} \cap \mathcal{X} = \emptyset$). If a random extraction of a multi-sample $\delta^{(1)}, \ldots, \delta^{(N)}$ is rejected when no optimal solution exists, another extraction is performed in such case.

Theorem 9. *Let $\Delta_E^N \subseteq \Delta^N$ be the set where a solution of SCP_N exists. If $P^N(\Delta_E^N) > 0$, then*

$$\frac{\mathbb{E}_{P^N}[V(\hat{x}_N) \cap 1(\Delta_E^N)]}{P(\Delta_E^N)} \le \frac{n}{N+1}$$

Moreover, in this case Corollary 1 still holds, provided that $1-\beta$ is intended as a lower bound on the conditional probability $P^N(\{V(\hat{x}_N) \le \epsilon\} \cap \Delta_E^N)/P^N(\Delta_E^N)$. (the measurability of Δ_E^N is taken as an assumption).

7.1.1 Extension: Incorporating Contextual Information

Given a new input x^0 and its realized parameter b^0, we have a convex optimization problem

$$CP(b^0) := \left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b^0) \leq 0 \right\}$$

Unlike oracle situation in which we have an access to all information, b^0 is usually unknown yet depends on the context x^0. Recall that prediction-then-optimization that solves $CP(\hat{b}^0)$ may have feasibility issues:

1. Infeasibility from Prediction: $S_0 \cap \{w : f(w, \hat{b}^0) \leq 0\} = \emptyset$
2. Post-Realization Infeasibility: $f(\hat{w}^*, b^0) > 0$ where $\hat{w}^*$ is the solution to the $CP(\hat{b}^0)$

Following the notations in C&C, we simply obtain the contextual uncertain convex problem as follows:

$$CUCP(x^0) := \left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b) \leq 0,\ b \in \Delta(x^0) \right\}$$

Note that this is a family of convex optimization problems whose constraints are parameterized by an uncertainty set that depends on the input x^0. Above formulation "relaxes" the two infeasibility issues:

1. Instead of working with the point-estimate $\hat{b}^0$, we have multiple candidates of b^0 that lie in the uncertainty set $\Delta(x^0)$. The assumption here is

   $$S_0 \cap \mathcal{W} \neq \emptyset \text{ where } \mathcal{W} = \bigcap_{b \in \Delta(x^0)} \{w : f(w, b) \leq 0\}$$

   In other words, the robust version of the problem

   $$RCP(x^0) := \left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b) \leq 0,\ \forall b \in \Delta(x^0) \right\}$$

   has a feasible solution.

2. We can instead control the violation (i.e., infeasibility) probability

   $$V(w) := P_{x^o} \left\{ b \in \Delta(x^0) : f(w, b) > 0 \right\}$$

   where P_{x^0} is the probability measure associated with its support $\Delta(x^0)$

7.1.2 Approach 1: Sampling from the Learned Conditional Distribution

If we know the true conditional distribution $P(B \mid X = x_0)$, we can simply set $P_{x^0} = P(B \mid X = x_0)$ and $\Delta(x^0) = \text{supp}(P_{x^0})$. Applying the result of C&C, the following randomized problem gives an $\epsilon-$level solution:

$$CSCP_N(x^0) := \left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b^{(i)}) \leq 0,\ i \in [N] \right\}$$

where $(b^{(1)}, \ldots, b^{(N)}) \sim P_{x^0}^N$ is the tuple of N i.i.d. samples extracted according to probability P_{x^0}. Under the true conditional distribution, we have the tail bound result with a confidence parameter β

$$P_{x^0}^N \left\{ V(\hat{w}_N) \leq \epsilon \right\} \geq 1 - \beta$$

where $\hat{w}_N$ is the solution to $CSCP_N(x^0)$. Note that using the data $b^1, \ldots, b^n$ does not work because each $b^i \sim P(B \mid X = x^i)$ while $b^0 \sim P(B \mid X = x^0)$.

One simple approach is to learn the distribution $\hat{P}_{x^0}$ from the data $\{(x^i, b^i)\}_{i=1}^n$. Let $(\hat{b}^{(1)}, \ldots, \hat{b}^{(N)}) \sim \hat{P}_{x^0}^N$

be the sampled tuple according to probability $\hat{P}_{x^0}$. Denoting $\hat{\hat{w}}_N$ the solution to the problem with sampled constraints, we have the same tail bound result

$$\hat{P}_{x^0}^N \left\{ V(\hat{\hat{w}}_N) \leq \epsilon \right\} \geq 1 - \beta$$

while the desired probability bound is $P_{x^0}^N \left\{ V(\hat{\hat{w}}_N) \leq \epsilon \right\}$ which is computed w.r.t the true conditional distribution. Notice that there are two sources of discrepancy.

- (Support Discrepancy) The induced support $\hat{\Delta}(x^0)$ of $\hat{P}_{x^0}$ may be different from true $\Delta(x^0)$ of P_{x^0}
- (Distributional Discrepancy) If the learning model is misspecified, it is very likely that $\hat{P}_{x^0} \not\to P_{x^0}$

Case Study: Logistic Regression We assume that we have n samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} P \in \mathcal{Q}$ and we consider a parametric family $\mathcal{P} = \{p_\theta : \theta \in \Theta\} \subset \mathcal{Q}$ for the purpose of approximating P. Note that P is not necessarily a member of $\mathcal{P}$. The m-estimator associated with a given a function $m_\theta(x)$ is

$$\arg\max_{\theta \in \Theta} M_n(\theta) \quad \text{where} \quad M_n(\theta) = P_n m_\theta = \frac{1}{n} \sum_{i=1}^{n} m_\theta \left(X_i, Y_i \right)$$

Let the pairs $z = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, and consider the logistic loss $m_\theta(z) = \log\left(1 + \exp\left(-y\theta^T x\right)\right)$, with population expectation $M(\theta) := \mathbb{E}\left[m_\theta(X, Y)\right]$ for $(X, Y) \sim P$. The case $m_\theta(x) = \log p_\theta(x)$ reduces m-estimators to the MLE.

- Step 1. If $\Theta \subset \mathbb{R}^d$ is a compact set and $\mathbb{E}[\|X\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

$$\sup_{\theta \in \Theta} |P_n m_\theta(X, Y) - M(\theta)| \xrightarrow{P} 0. \tag{ULLN}$$

- Step 2. (ULLN) + For all $\epsilon > 0$, $\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M\left(\theta_0\right)$. Then for any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n\left(\theta_0\right) - o_P(1)$ we have

$$\hat{\theta}_n \xrightarrow{P} \theta_0$$

- Step 3. Under parametric assumption, we have $\hat{P}_{x^0} := P_{x^0}^{\hat{\theta}_n}$ and $P_{x^0} := P_{x^0}^{\theta_0}$. From continuous mapping theorem,

$$\frac{1}{1 + \exp(-y\hat{\theta}_n^\top x^0)} \xrightarrow{P} \frac{1}{1 + \exp(-y\theta_0^\top x^0)}$$

  Not entirely sure this is the right way of establishing (weak) convergence of measures...

So if $\hat{P}_{x^0} \to P_{x^0}$ (converges in distribution, or uniform convergence of density functions?), $\hat{P}_{x^0}^N \to P_{x^0}^N$ and we can leverage the tail bound result.

Numerical experiment: How should we learn the conditional distribution? Assuming the relationship parameterized by θ,

$$P(B = b \mid X = x) = p_{\theta^*}(x, b)$$

any good continuous examples? (Logistic regression/Multinomial logistic regression only considers finite support of B)

7.1.3 Approach 2: Better Sampling with Kernel Herding / Reduced Set Method

Koduri (2021) suggested a local approach to the optimization problem with side-information. Given the original data $D = \{(x^i, b^i)\}_{i=1}^n$, we can find a $k-$NN of x^0 (w.r.t some metric or using CART/RF), a subset $\{(x^{(i)}, b^{(i)})\}_{i=1}^k \subset D$, and solve

$$\left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b^{(i)}) \leq 0,\ i \in [k] \right\}$$

Instead, we can try Kernel Herding/Reduced Set Method to generate/select samples that retains as much as information related to the input x^0. The goal is to extend the proposed vanilla methods to deal with conditional distributions.

Kernel Herding Consider a sampling problem where we wish to find an infinite sequence of points $x_1, x_2, \ldots$ from the set $\mathcal{X}$. Each x_i is called a pseudosample. We want to ensure that the error of the first T pseudosamples is minimized (or bounded at most α/T where α is a quantity that can depend on the problem parameters) from the sampling scheme. In the kernelized setting, kernel herding minimizes the squared error between expected feature values evaluated at the true distribution and the empirical distribution obtained from herding.

$$\mathcal{E}_T^2 = \left\| \mu_P - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \right\|_{\mathcal{H}}^2$$

Here, $\mu_P = \int_{\mathcal{X}} k(\cdot, x) dP(x)$ is the kernel mean embedding of probability measure P in into an RKHS $\mathcal{H}$ endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (and thus with the corresponding feature map $\phi(x) = k(\cdot, x)$). The kernel herding update rule is defined as follows:

$$\begin{cases} x_{t+1} = \arg\max_{x \in \mathcal{X}} \langle w_t, \phi(x) \rangle \\ w_{t+1} = w_t + \mu_P - \phi(x_{t+1}) \end{cases} \tag{1}$$

Under the assumption that the data is uniformly bounded in feature space, that is, for all $x \in \mathcal{X}$, $\|\phi(x)\|_{\mathcal{H}} \leq R$, we have the following proposition (Chen et al. 2010)

Proposition 1. *If μ_P is in the relative interior of the marginal polytope $\mathcal{M} := conv\{\phi(x) \mid x \in \mathcal{X}\}$, then*

$$\left\| \mu_P - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \right\|_{\mathcal{H}} \leq \frac{\|w_0\| + R/\gamma^*}{T}$$

where $\gamma^* \leq \left\langle \frac{w_t}{\|w_t\|}, \frac{c_t}{\|c_t\|} \right\rangle$ with $c_t := \arg\max_{c \in \mathcal{M} - \mu_p} \langle w_t, c \rangle$

Denoting $\hat{\mu}_T := \frac{1}{T} \sum_{t=1}^{T} \phi(x_t)$ and combining with the convergence theorem of empirical mean embedding (Theorem 3.4 in the review paper), we get

$$\begin{aligned} \|\hat{\mu}_P - \hat{\mu}_T\| &\leq \|\hat{\mu}_P - \mu_P\| + \|\mu_P - \hat{\mu}_T\| \\ &\leq \underbrace{\sqrt{\frac{C_k}{n}} + \sqrt{\frac{2C_k \log \frac{1}{\delta}}{n}}}_{\text{number of data points}} + \underbrace{\frac{\|w_0\| + R/\gamma^*}{T}}_{\text{number of generated samples}} \end{aligned}$$

where $C_k \geq \sup_{x \in \mathcal{X}} k(x, x)$ is a finite constant (Note that we can simply take $C_k = R$)

Remark: The first term is probability bound such that the event happens with probability at least $1 - \delta$. On the other hand, the second term is deterministic.

Equivalence between Frank-Wolfe method solving the following optimization problem (Bach et al. 2012):

$$\min_{g \in \mathcal{M}} J(g) = \frac{1}{2} \|g - \mu_P\|^2$$

Here, $\mathcal{M}$ is a marginal polytope and μ_P is the trivial solution, which is the target distribution. The Frank-Wolfe iterates are

$$\begin{cases} \bar{g}_{t+1} \in \arg\min_{g \in \mathcal{M}} \langle g_t - \mu_P, g \rangle \\ g_{t+1} = (1 - \rho_t) g_t + \rho_t \bar{g}_{t+1} \end{cases}$$

Two natural choices for ρ_t:

(a) Fixed Step Size $\rho_t = \frac{1}{t+1}$:

This gives the standard kernel herding as in (1) with uniform weights i.e., $\mu_P = \sum_{t=1}^{T} \frac{1}{T}\phi(x_t)$

(b) Line search to find the point in the segment with optimal value:

This gives non-uniform weights $\mu_P = \sum_{t=1}^{T} w_t \phi(x_t)$

If μ_P is in the relative interior of $\mathcal{M}$, we can take a ball of center μ_P and radius $d > 0$ that is included in $\mathcal{M}$. Under Slater condition, the convergence analysis of Frank-Wolfe gives

(a) Fixed Step Size $\rho_t = \frac{1}{t+1}$:

$$\frac{1}{2}\|g_T - \mu_P\|^2 \leq \frac{2R^4}{d^2T^2}$$

Same rate $O(1/T^2)$ as in Proposition 1.

(b) With Line Search:

$$\frac{1}{2}\|g_T - \mu_P\|^2 \leq R^2 \exp(-\frac{d^2T}{R^2})$$

which follows from the linear convergence rate.

Note that both choices of step sizes yield faster rates than random sampling.

Comments:

- We are not directly setting or estimating the support $\hat{\Delta}(x^0)$ here. Rather, it is implicitly assumed that the support is the pre-image of the marginal polytope $\mathcal{M}$?
- Herding is an infinite memory process on x_t (as opposed to a Markov process) because new samples depend on the entire history of samples generated. What happens if the sampled trajectory $b^{(1)}, \ldots, b^{(T)}$ do not yield feasibility

$$\bigcap_{t=1,\ldots,T} \{w : f(w, b^{(t)}) \leq 0\} = \emptyset$$

Things to be addressed

- Performance of vanilla approaches (without side-information) compared to the naïve sampling of C&C.
  - Qualitatively better solution with better samples?
  - Less number of data points required for consistency?
    The tail bound proof in C&C relies on combinatorial argument, especially the $N + 1$ in the denominator, so it seems that the required number of samples cannot be improved in general

Reduced Set Method Given $\{b^i\}_{i=1}^n$, we have empirical kernel mean embedding $\hat{\mu}_B = \frac{1}{n}\sum_{i=1}^{n} \phi(b^i)$.

1. Solve

$$\min_{\alpha} \|w^\top \alpha\|_1 \text{ s.t. } \left\| \sum_{i=1}^{n} \alpha_i \phi(b^i) - \frac{1}{n}\sum_{i=1}^{n} \phi(b^i) \right\|_{\mathcal{H}} \leq \epsilon$$

   and obtain $\mathcal{R} = \{i : a_i \neq 0, i = 1, \ldots, n\}$

2. Solve

$$\left\{ \min_{w \in S_0} c^\top w \text{ s.t. } f(w, b^i) \leq 0,\ i \in \mathcal{R} \right\}$$

Remark: Reduced set method can be viewed as a sparse approximation of empirical kernel mean embedding?

7.2 Contextual Decision-making under Parametric Uncertainty and Data-driven Optimistic Optimization (Cao and Gao, 2022)

Setting:

$$\mathbb{E}[R \mid X = x] = r(x; \theta^*)$$

where θ^* is an unknown true model parameter. Assuming

$$Y = f_{\theta^*}(X) + \epsilon$$

where ϵ is a zero-mean noise. From the data $\{(Y_i, X_i\}_{i=1}^n$, we want to estimate θ^* via ERM

$$\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$$

Supervised-Learning-Oracle-Based Parameter Uncertainty Set

Let $\psi(y, x, \epsilon; \theta) := \nabla_\theta \ell(y, f_\theta(x) + \epsilon) \in \mathbb{R}^d$. Consider the following uncertainty set

$$\left\{ (\theta, \epsilon) \in \mathbb{R}^d \times \mathbb{R}^n : \sum_{i=1}^n \psi(Y_i, X_i, \epsilon_i; \theta) = 0, \sum_{i=1}^n \frac{\epsilon_i^2}{\omega_{n,i}^2} \leq \rho_n^2 \right\} \tag{S}$$

Here, the projection on the ϵ-component $\{\epsilon \in \mathbb{R}^n : \frac{\epsilon_i^2}{\omega_{n,i}^2} \leq \rho_n^2\}$ is an ellipsoidal uncertainty set for the noises $\{\epsilon_i\}_{i=1}^n$, where $\omega_{n,i} > 0$ indicates the (estimated) standard deviation of ϵ_i.

With the new uncertainty set S, we have the data-driven robust/optimistic optimization problem:

$$\begin{aligned} \min_{\theta \in \mathbb{R}^d} / \max_{\epsilon \in \mathbb{R}^n} \; & r(x; \theta) \\ \text{s.t.} \; & \sum_{i=1}^n \psi(Y_i, X_i, \epsilon_i; \theta) = 0 \\ & \sum_{i=1}^n \frac{\epsilon_i^2}{\omega_{n,i}^2} \leq \rho_n^2 \end{aligned}$$

8 Two Distribution Problem

Distance between probability measures are usually defined over the same $\sigma-$algebra (i.e., w.r.t same support Δ) One of the widely studied and well understood families of distances/divergences between probability measures is integral probability metrics (IPMs), defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int_\Delta f d\mathbb{P} - \int_\Delta f d\mathbb{Q} \right|$$

This induces distance measures by appropriately choodsing $\mathcal{F}$:

- Wasserstein distance: $\mathcal{F} = \{f : \|f\|_L \leq 1\}$ (i.e., the set of 1-Lipschitz functions)
- Total variation distance: $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$
- Maximum mean discrepancy distance: $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$

8.0.1 Problem Formulation

Suppose we have two measures $\hat{P}$ and P where we want to approximate the true measure P via $\hat{P}$. Given

$$d(\hat{P}, P) \leq R \ ,$$

we want to bound the violation probability under the true measure. Let $\hat{w}_N$ denote the solution to

$$\begin{aligned} &\min \ c^\top w \\ &\text{s.t. } f(w, \hat{\delta}^{(i)}) \leq 0, \quad i = 1, \dots, N \\ &\quad w \in S_0 \end{aligned}$$

where $(\hat{\delta}^{(1)}, \dots, \hat{\delta}^{(N)}) \sim \hat{P}^N$ (and indeed $(\hat{\delta}^{(1)}, \dots, \hat{\delta}^{(N)}) \in \Delta^N$ i.i.d sampled). We also define violation probabilities as follows:

$$\begin{aligned} \hat{V}(w) &= \hat{P}\{f(w, \delta) > 0, \delta \in \Delta\} && \text{(violation probability under estimated measure)} \\ V(w) &= P\{f(w, \delta) > 0, \delta \in \Delta\} && \text{(violation probability under true measure)} \end{aligned}$$

From C&C, we know that

$$\hat{P}^N\{\hat{V}(\hat{w}_N) > \epsilon\} \leq \frac{d}{(N+1)\epsilon}$$

where d denote the dimension of $S_0 \subseteq \mathbb{R}^d$. Suppose that $\|\hat{P} - P\|_{TV} \leq R$. Then, $\|\hat{P}^N - P^N\|_{TV} \leq NR$ (follows from tensorization property which also applies to Wasserstein-1 distance $W_1(P, Q)$; see this). Hence,

$$\begin{aligned} P^N(V(\hat{w}_N) \geq \epsilon) &\leq \hat{P}^N(V(\hat{w}_N) \geq \epsilon) + NR \\ &\leq \hat{P}^N(\hat{V}(\hat{w}_N) \geq \epsilon - R) + NR \\ &\leq \frac{d}{(N+1)(\epsilon - R)} + NR \end{aligned}$$

Solving for a confidence parameter β, for

$$N \geq \frac{-(\epsilon - R)(R - \beta) - \sqrt{(\epsilon - R)^2(R + \beta)^2 - 4dR(\epsilon - R)}}{2R(\epsilon - R)},$$

we have $V(\hat{w}_N) < \epsilon$ with probability at least $1 - \beta$, or equivalently

$$P^N(V(\hat{w}_N) < \epsilon) \geq 1 - \beta$$

As $R \to 0+$, notice that we get the same sample guarantee $N \geq \frac{d}{\epsilon\beta} - 1$. The implicit assumptions $\epsilon - R > 0$, $R - \beta < 0$, and $NR \leq 1$ can be satisfied if R is small enough (naturally holds as $\hat{P} \to P$)

If $R = O(n^{-1/2})$, this implies that

$$\lim_{n \to \infty} \frac{f(n)}{1} = \frac{d}{\epsilon\beta} - 1 < \infty$$

so $N = O(1)$

We want to derive the similar results when $\|\mu_{\hat{P}} - \mu_P\|_{\mathcal{H}} \leq R$.

Key ingredients ($\sup_{x \in \mathcal{X}} k(x, x) \leq C < \infty$, k measurable on $\mathcal{X}$, $\tilde{\rho} = \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}}$) :

- $\|\mu_P - \mu_Q\|_{\mathcal{H}} \leq \sqrt{C}\|P - Q\|_{TV}$
- $\|\mu_P - \mu_Q\|_{\mathcal{H}} \leq W(P, Q) \leq \sqrt{\|\mu_P - \mu_Q\|_{\mathcal{H}}^2 + 4C}$ if $(\mathcal{X}, \tilde{\rho})$ is separable
- $W(P, Q) \leq \text{diam}(\mathcal{X})\|P - Q\|_{TV}$

- In the specific case where $\mathcal{H} = L^2(\mathcal{X}, m)$ for m the normalized Lebesgue measure on $\mathcal{X}$, we know that $\{f \in C_b(\mathcal{X}), \|f\|_\infty \leq 1\}$ will be contained in $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_\mathcal{H} \leq 1\}$, and therefore $\|P - Q\|_{TV} \leq \|\mu_P - \mu_Q\|_\mathcal{H}$. Nevertheless this is a very extreme case, since we would need a very powerful kernel to approximate the whole L^2. (see this)
- Weak convergence implies MMD convergence if and only if the kernel is bounded and continuous. Convergence in MMD is often rather weak and can, at best, metrize weak convergence, but not convergence in total variation or KL divergence since those are known to be strictly stronger than weak convergence (see this)

Simply put, MMD does not have nice tensorization property to unravel i.i.d. product measure. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical distribution and $\mu_{P_n} = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$ be its kernel mean embedding. We know from Theorem 3.4 in the review paper, with proability at least $1 - \delta$

$$\|\mu_{P_n} - \mu_P\|_\mathcal{H} \leq \sqrt{\frac{C}{n}} + \sqrt{\frac{2C \log \frac{1}{\delta}}{n}}$$

Now consider the product measure P^N (N copies of i.i.d random variables) We can similarly extend above to the multi-dimensional case. Let $\mathcal{H}^N = \mathcal{H} \otimes \cdots \otimes \mathcal{H}$ be the RKHS of real-valued functions with domain $\Delta^N = \Delta \times \cdots \times \Delta$ associated with the kernel function

$$k^N((x_1, \ldots, x_N), (x'_1, \ldots, x'_N)) = k(x_1, x'_1) \cdot \cdots \cdot k(x_N, x'_N)$$

Or simply with the feature map $\phi^N : \Delta^N \to \mathcal{H}^N$ and $\phi : \Delta \to \mathcal{H}$, we get the expression

$$\phi^N((x_1, \ldots, x_N)) = \phi(x_1) \cdot \cdots \phi(x_N)$$

We claim that with probability at least $1 - \delta$,

$$\left\|\mu_{P_n^N} - \mu_{P^N}\right\|_{\mathcal{H}^N} \leq \sqrt{N\left(\frac{C}{n}\right)^N} + \sqrt{\frac{2N^2 C^N \log \frac{1}{\delta}}{n}}$$

The proof sketch follows from Prop. A.1 of Tolstikhin et al. (2017):

- Step 1. Apply McDiarmid's inequality:

  Let $f(x_1, \ldots, x_n) = \left\|\int_{\Delta^N} \phi^N(x) dP_n^N(x) - \int_{\Delta^N} \phi^N(x) dP^N(x)\right\|_{\mathcal{H}^N}$. Then, by reverse triangle inequality,

$$\begin{aligned}
|f(x_1, \ldots, x_n) - f(x'_1, \ldots, x_n)| &\leq \left\| \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right)^N - \left(\frac{1}{n}\left(\phi(x'_1) + \sum_{i=2}^n \phi(x_i)\right)\right)^N \right\|_{\mathcal{H}^N} \\
&\leq \left\| \frac{\phi(x_1) - \phi(x'_1)}{n} \right\|_{\mathcal{H}^N} N \cdot \sqrt{C}^{N-1} \leq \frac{2N\sqrt{C}^N}{n}
\end{aligned}$$

  Hence,

$$\left\|\mu_{P_n^N} - \mu_{P^N}\right\|_{\mathcal{H}^N} \leq \mathbb{E}\left[\left\|\mu_{P_n^N} - \mu_{P^N}\right\|_{\mathcal{H}^N}\right] + \sqrt{\frac{2N^2 C^N \log \frac{1}{\delta}}{n}}$$

- Upper bound the expectation term:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mu_{P_n^N} - \mu_{P^N}\right\|_{\mathcal{H}^N}\right] &= \mathbb{E}\left[\left\|\int_{\Delta^N} \phi^N(x) dP_n^N(x) - \int_{\Delta^N} \phi^N(x) dP^N(x)\right\|_{\mathcal{H}^N}\right] \\
&\leq \sqrt{\mathbb{E}\left[\left\|\int_{\Delta^N} \phi^N(x) dP_n^N(x) - \int_{\Delta^N} \phi^N(x) dP^N(x)\right\|_{\mathcal{H}^N}^2\right]} \\
&= \sqrt{\underbrace{\mathbb{E}\left[\left\|\int_{\Delta^N} \phi^N(x) dP_n^N(x)\right\|_{\mathcal{H}^N}^2\right]}_{(A)} + \left\|\int_{\Delta^N} \phi^N(x) dP_n^N(x)\right\|_{\mathcal{H}^N}^2 - 2\underbrace{\mathbb{E}\left[\langle \cdots, \cdots \rangle_{\mathcal{H}^N}\right]}_{(B)}}
\end{aligned}$$

- (B)

$$\begin{aligned}
\mathbb{E}\left[\left\langle \int_{\Delta^N} \phi^N(x) dP_n^N(x), \int_{\Delta^N} \phi^N(x) dP^N(x) \right\rangle_{\mathcal{H}^N}\right] &\overset{(1)}{=} \mathbb{E}\left[\left\langle \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right)^N, \left(\int_\Delta \phi(x) dP(x)\right)^N \right\rangle_{\mathcal{H}^N}\right] \\
&\overset{(2)}{=} \mathbb{E}\left[\left\langle \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right), \left(\int_\Delta \phi(x) dP(x)\right) \right\rangle_{\mathcal{H}}^N\right] \\
&\overset{(3)}{=} \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^n \phi(x_i)\ , \int_\Delta \phi(x) dP(x) \right\rangle_{\mathcal{H}}\right]^N \\
&\overset{(4)}{=} \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\langle \phi(x_i), \int_\Delta \phi(x) dP(x) \right\rangle_{\mathcal{H}}\right)^N \\
&\overset{(5)}{=} \left\| \int_\Delta \phi(x) dP(x) \right\|_{\mathcal{H}}^{2N}
\end{aligned}$$

  (1) Definition
  (2) Tensor product property
  (3) Independence
  (4) Linearity of expectation
  (5) Reproducing property

- (A)

$$\begin{aligned}
\mathbb{E}\left[\left\| \int_{\Delta^N} \phi^N(x) dP_n^N(x) \right\|_{\mathcal{H}^N}^2\right] &= \left(\mathbb{E}\left[\left\| \int_\Delta \phi(x) dP_n(x) \right\|_{\mathcal{H}}^2\right]\right)^N \\
&= \left(\frac{1}{n}\mathbb{E}_{X\sim P}\left[\|\phi(X)\|_{\mathcal{H}}^2\right] + \frac{n-1}{n}\mathbb{E}_{X\sim P, Y\sim P}\left\langle \phi(X), \phi(Y)\right\rangle_{\mathcal{H}}\right)^N \\
&= \left(\frac{\mathbb{E}_{X\sim P}\left[\|\phi(X)\|_{\mathcal{H}}^2\right] - \left\|\int_\Delta \phi(x) dP(x)\right\|_{\mathcal{H}}^2}{n} + + \left\| \int_\Delta \phi(x) dP(x) \right\|_{\mathcal{H}}^2\right)^N
\end{aligned}$$

- Let $a = \left\|\int_\Delta \phi(x) dP(x)\right\|_{\mathcal{H}}^2$, and we know that $a \le C$. Putting together the pieces, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mu_{P_n^N} - \mu_{P^N}\right\|_{\mathcal{H}^N}\right] &\le \sqrt{\left(\frac{C-a}{n}\right)^N - a^N} \\
&\le \sqrt{\left(\frac{C}{n} + a\right)^N - a^N} \\
&\le \sqrt{\frac{C}{n} N \left(\frac{C}{n}\right)^{N-1}}
\end{aligned}$$

Note that

$$\left|P_n^N\{V(x) > \epsilon\} - P^N\{V(x) > \epsilon\}\right| = \left| \int_{\Delta^N} 1\{V(x) > \epsilon\} dP_n^N(x) - \int_{\Delta^N} 1\{V(x) > \epsilon\} dP^N(x) \right|$$

and

$$|\hat{V}(x) - V(x)| = \left| \int_\Delta 1\{f(w,x) > \epsilon\} dP_n(x) - \int_\Delta 1\{f(w,x) > \epsilon\} dP(x) \right|$$

Assuming both indicator functions are well represented within the unit RKHS ball $\{f : |f\|_{\mathcal{H}} \leq 1\}$, we get

$$\begin{aligned} P^N(V(\hat{w}_N) \geq \epsilon) &\leq P_n^N(V(\hat{w}_N) \geq \epsilon) + \sqrt{N\left(\frac{C}{n}\right)^N} + \sqrt{\frac{2N^2C^N \log\frac{3}{\beta}}{n}} \\ &\leq \hat{P}^N(\hat{V}(\hat{w}_N) \geq \epsilon - \sqrt{\frac{C}{n}} - \sqrt{\frac{2C\log\frac{3}{\beta}}{n}}) + \sqrt{N\left(\frac{C}{n}\right)^N} + \sqrt{\frac{2N^2C^N \log\frac{3}{\beta}}{n}} \\ &\leq \frac{d}{(N+1)(\epsilon - \sqrt{\frac{C}{n}} - \sqrt{\frac{2C\log\frac{3}{\beta}}{n}})} + \sqrt{N\left(\frac{C}{n}\right)^N} + \sqrt{\frac{2N^2C^N \log\frac{3}{\beta}}{n}} \end{aligned}$$

Solving $\beta/3 = RHS$, we have $V(\hat{w}_N) < \epsilon$ with probability at least $1-\beta$ using union bound. Assuming $C = 1$ (holds for Gaussian kernel) we get

$$\begin{aligned} (RHS) &\leq \frac{d}{(N+1)\delta} + \sqrt{N\left(\frac{1}{n}\right)^N} + \sqrt{\frac{2N^2\log\frac{3}{\beta}}{n}} \\ &\leq \frac{d}{(N+1)\delta} + \frac{1}{\sqrt{e\log n}} + N\sqrt{\frac{2\log\frac{3}{\beta}}{n}} \end{aligned}$$

which has a closed form solution.

To sum up, if $\|P_n - P\| = R = O(n^{-1/2})$ (or $O_P(n^{-1/2})$), assuming that we fix hyperparameters β and ϵ, we obtain

- Using metric $\|\cdot\|_{TV}$:

$$N \gtrsim f(n) := \frac{-(\epsilon - n^{-1/2})(n^{1/2} - \beta) - \sqrt{(\epsilon - n^{1/2})^2(n^{-1/2}+\beta)^2 - 4dR(\epsilon - n^{-1/2})}}{2n^{-1/2} + (\epsilon - n^{-1/2})} \implies f(n) = O(1)$$

- Using metric $\|\cdot\|_{\mathcal{H}}$:

$$N \gtrsim g(n) \implies g(n) = O_P(1)$$

$$\frac{-\left(\sqrt{\frac{2\log\frac{\beta}{3}}{n}} - \frac{\beta}{3} + \frac{1}{\sqrt{e\log n}}\right)\left(\epsilon - \frac{1}{\sqrt{n}} - \sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right) - \sqrt{\left(\sqrt{\frac{2\log\frac{\beta}{3}}{n}} - \frac{\beta}{3} + \frac{1}{\sqrt{e\log n}}\right)^2\left(\epsilon - \frac{1}{\sqrt{n}} - \sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right)^2 - 4\left(\sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right)\left(\epsilon - \frac{1}{\sqrt{n}} - \sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right)\left(d - \left(\frac{\beta}{3} - \frac{1}{\sqrt{e\log n}}\right)\left(\epsilon - \frac{1}{\sqrt{n}} - \sqrt{\frac{2\log}{n}}\right.\right.}}{2\left(\sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right)\left(\epsilon - \frac{1}{\sqrt{n}} - \sqrt{\frac{2\log\frac{\beta}{3}}{n}}\right)}$$

Here, we assume that the kernel is bounded and continuous with $\sup_{x\in\mathcal{X}} k(x,x) = 1$ plus other regularity assumptions to well-define the N-fold tensor product RKHS. (Gaussian kernel will suffice)

Intuitively, if we have more and more data so that we well-approximate the true measure, we will need less samples N to get ϵ-feasible solution of our original optimization problem, and vice versa. Yet, $P_n \to P$ with rate $n^{-1/2}$ simply implies convergence $P_n^N \to P^N$ of i.i.d. product measure at the same rate, so as $n \to \infty$, N becomes asymptotically constant which is the canonical guarantee $\frac{d}{\epsilon\beta} - 1$ of Calafiore and Campi.

In the analysis, we need to compute the N-fold product measure, and the above finite sample guarantees are derived under the i.i.d sampling assumption. In non-i.i.d settings, including kernel herding, it seems hard to work with the distribution of entire sample trajectory.

- Phase 1: n data $\to$ Empirical measure P_n
- Phase 2: N i.i.d. sample from $P_n \to$ Tail bound of true violation probability $P^N(V(w) > \epsilon)$ using P_n^N
  T samples from kernel herding (deterministic sample trajectory that well-approximates P or P_n) $\to$ Can we give sample bound $T \gtrsim f(\epsilon)$ s.t. $V(w) > \epsilon$?

8.0.2 Other Keywords

- MMD Coreset:

  The worst-case error between sample and target expectations over the RKHS unit ball is given by the kernel MMD:

  $$MMD_k(P, P_n) := \sup_{\|f\|_k \leq 1} |Pf - P_n f|$$

  and we call a sequence of points $(x_i)_{i=1}^n$ an (n, ϵ)-MMD coreset for (k, P) if $MMD_k(P, P_n) \leq \epsilon$.

  An i.i.d. sample from P yields an order $(n^{-1/2}, n^{-1/4})$-MMD coreset with high probability. (Tolstikhin et al., 2017)

- Kernel Thinning

  Given a target distribution P on $\mathbb{R}^d$, a reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and a sequence of $\mathbb{R}^d$-valued input points $\mathcal{S}_{in} = (x_i)_{i=1}^n$, the goal is to identify a thinned MMD coreset, a subsequence $\mathcal{S}_{out}$ of size $n^{-1/2}$ satisfying $MMD_k(P, \mathcal{S}_{out}) = o(n^{-1/4})$. Here, the requirement is that $MMD_k(P, \mathcal{S}_{in}) = O(n^{-1/2})$ which can be achieved by i.i.d sampling (or kernel herding)