CUHK STAT6050: Statistical Learning Theory Spring 2026

Lecture 9: Maximum Mean Discrepancy (MMD)

RKHS, Kernel Test, Reproducing Property

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 Motivation

A central problem in statistics and machine learning is to determine whether two probability distributions are the same. Formally, we are given:

Problem. Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables defined on a topological space $\mathscr{X}$, with respective Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$. Given observations $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$ and $\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$, independently and identically distributed (i.i.d.) from $\mathbb{P}$ and $\mathbb{Q}$, respectively, can we decide whether $\mathbb{P}\neq\mathbb{Q}$?

Where there is no ambiguity, we use the shorthand notation $\mathbf{E}_{\mathbb{P}}(f(\mathbf{X}))$ and $\mathbf{E}_{\mathbb{Q}}(f(\mathbf{Y}))$ to denote expectations with respect to $\mathbb{P}$ and $\mathbb{Q}$, respectively.

This problem arises in:

- **Two-sample testing**: test $H_0:\mathbb{P}=\mathbb{Q}$ vs. $H_1:\mathbb{P}\neq\mathbb{Q}$.
- **Covariate shift / transfer learning**: assessing domain shift between source and target.
- **Generative model evaluation**: comparing real data distribution to a learned generative model.

Our goal is to construct a criterion that takes on a unique and distinctive value only when $\mathbb{P}=\mathbb{Q}$. Following [Gretton et al., 2012], this will be defined via a lemma of Dudley (2002).

2 MMD via Bounded Continuous Functions

2.1 Definition

We measure the discrepancy between $\mathbb{P}$ and $\mathbb{Q}$ by comparing expectations of test functions from some function class $\mathscr{F}$.

Definition 2.1 (Maximum Mean Discrepancy [Gretton et al., 2012])**.** Let $\mathscr{F}$ be a class of functions $f:\mathscr{X}\to\mathbb{R}$. The **Maximum Mean Discrepancy** (MMD) is defined as:

$$\mathrm{MMD}[\mathscr{F},\mathbb{P},\mathbb{Q}]:=\sup_{f\in\mathscr{F}}\left(\mathbf{E}_{\mathbb{P}}(f(\mathbf{X}))-\mathbf{E}_{\mathbb{Q}}(f(\mathbf{Y}))\right).$$

The key question is: which function class $\mathscr{F}$ makes MMD a *metric* on probability distributions, i.e., $\mathrm{MMD}[\mathscr{F},\mathbb{P},\mathbb{Q}]=0$ if and only if $\mathbb{P}=\mathbb{Q}$?

2.2 MMD as a Metric: The Role of $\mathscr{F}$

The following result (Lemma 9.3.2 of Dudley, 2002) characterizes which function classes make MMD a metric.

Lemma 2.2 (Dudley, 2002)**.** *Let $\mathscr{X}$ be a separable metric space and let $\mathscr{F} = \mathscr{C}_b(\mathscr{X})$ be the space of bounded continuous functions on $\mathscr{X}$. Then for Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$:*

$$\mathrm{MMD}[\mathscr{C}_b(\mathscr{X}), \mathbb{P}, \mathbb{Q}] = 0 \quad \Longleftrightarrow \quad \mathbb{P} = \mathbb{Q}.$$

Remark 2.3. This shows that the supremum over all bounded continuous functions perfectly characterizes equality of distributions. However, $\mathscr{C}_b(\mathscr{X})$ is too large to be computationally tractable, and the empirical estimate of MMD over $\mathscr{C}_b(\mathscr{X})$ does not converge at a useful rate. We need a smaller, more structured function class.

3 MMD via RKHS

3.1 Realization in an RKHS

Following [Gretton et al., 2012], we restrict $\mathscr{F}$ to the unit ball of a **Reproducing Kernel Hilbert Space** (RKHS) $\mathscr{H}_K$ associated with a positive definite kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$:

$$\mathscr{F} = \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \leq 1\}.$$

Definition 3.1 (MMD in RKHS)**.** Given a kernel K with RKHS $\mathscr{H}_K$, the MMD is:

$$\mathrm{MMD}[\mathscr{H}_K, \mathbb{P}, \mathbb{Q}] = \sup_{\|f\|_{\mathscr{H}_K} \leq 1} \left(\mathbf{E}_{\mathbb{P}}(f(\mathbf{X})) - \mathbf{E}_{\mathbb{Q}}(f(\mathbf{Y}))\right).$$

3.2 Mean Embedding and Closed-Form Expression

The RKHS structure allows us to represent the action of a distribution on $\mathscr{H}_K$ via a single element.

Definition 3.2 (Mean embedding)**.** The **mean embedding** of $\mathbb{P}$ in $\mathscr{H}_K$ is:

$$\mu_{\mathbb{P}} := \mathbf{E}_{\mathbb{P}}(K(\mathbf{X}, \cdot)) \in \mathscr{H}_K,$$

provided $\mathbf{E}_{\mathbb{P}}(\sqrt{K(\mathbf{X}, \mathbf{X})}) < \infty$. By the reproducing property, $\mathbf{E}_{\mathbb{P}}(f(\mathbf{X})) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}_K}$ for all $f \in \mathscr{H}_K$.

Proposition 3.3 (Closed-form MMD)**.** *The MMD in RKHS has the closed form:*

$$\begin{aligned} \mathrm{MMD}[\mathscr{H}_K, \mathbb{P}, \mathbb{Q}] &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}_K} \\ &= \left(\mathbf{E}_{\mathbb{P}\otimes\mathbb{P}}(K(\mathbf{X}, \mathbf{X}')) + \mathbf{E}_{\mathbb{Q}\otimes\mathbb{Q}}(K(\mathbf{Y}, \mathbf{Y}')) - 2\mathbf{E}_{\mathbb{P}\otimes\mathbb{Q}}(K(\mathbf{X}, \mathbf{Y}))\right)^{1/2}, \end{aligned} \tag{1}$$

where $\mathbf{X}, \mathbf{X}' \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ *and* $\mathbf{Y}, \mathbf{Y}' \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.

Sketch. By the reproducing property, $\mathbf{E}_{\mathbb{P}}(f(\mathbf{X})) = \langle f, \mu_{\mathbb{P}} \rangle$ and $\mathbf{E}_{\mathbb{Q}}(f(\mathbf{Y})) = \langle f, \mu_{\mathbb{Q}} \rangle$. Hence the supremum over the unit ball is achieved by $f^* \propto \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}$, giving $\mathrm{MMD} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|$. Expanding:

$$\begin{aligned}\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle \\ &= \mathbf{E}_{\mathbb{P}\otimes\mathbb{P}}(K(\mathbf{X}, \mathbf{X}')) - 2\,\mathbf{E}_{\mathbb{P}\otimes\mathbb{Q}}(K(\mathbf{X}, \mathbf{Y})) + \mathbf{E}_{\mathbb{Q}\otimes\mathbb{Q}}(K(\mathbf{Y}, \mathbf{Y}')).\end{aligned}$$ □

Example 3.4 (Linear kernel)**.** *For $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$:*

$$\mathrm{MMD}^2[\mathscr{H}_K, \mathbb{P}, \mathbb{Q}] = \|\mathbf{E}_{\mathbb{P}}(\mathbf{X}) - \mathbf{E}_{\mathbb{Q}}(\mathbf{Y})\|^2,$$

i.e. the squared distance between the means. This only detects first-order differences.

Example 3.5 (Gaussian kernel)**.** *For $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2)$, the kernel captures differences in all moments of $\mathbb{P}$ and $\mathbb{Q}$, making it suitable for detecting any distributional shift.*

3.3 Characteristic Kernels

In general, restricting from $\mathscr{C}_b(\mathscr{X})$ to $\mathscr{H}_K$ may lose the property that $\mathrm{MMD} = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$. The kernel must be *characteristic*.

Definition 3.6 (Characteristic kernel [Gretton et al., 2012])**.** A bounded measurable kernel K is called **characteristic** if the mean embedding $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective, i.e.,

$$\mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \quad \Longrightarrow \quad \mathbb{P} = \mathbb{Q}.$$

Equivalently, $\mathrm{MMD}[\mathscr{H}_K, \mathbb{P}, \mathbb{Q}] = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Theorem 3.7 (Gaussian kernel is characteristic [Steinwart, 2001, Gretton et al., 2012])**.** *The Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2)$ on $\mathbb{R}^d$ is characteristic.*

Remark 3.8. The Gaussian kernel is in fact *universal* (see Lecture 8, Section 3), which is a strictly stronger property implying it is also characteristic. The Laplacian kernel is likewise characteristic.

4 Kernel Two-Sample Test

4.1 Empirical Estimator

Given i.i.d. samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \sim \mathbb{P}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim \mathbb{Q}$, we consider the U-statistic estimator of MMD:

Unbiased estimator (U-statistic).

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{m(m-1)} \sum_{i \neq j} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} K(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} K(\mathbf{x}_i, \mathbf{y}_j),$$

which can be computed in $O((m+n)^2)$ time.

4.2 Statistical Properties

Finite-sample test based on the unbiased estimator

Apply McDiarmid's inequality directly to $\widehat{\mathrm{MMD}}_u^2(\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{y}_1,\ldots,\mathbf{y}_n)$. The **bounded difference** when replacing one $\mathbf{x}_i$ is:

$$c_{x,i} = \sup\left|\widehat{\mathrm{MMD}}_u^2(\ldots,\mathbf{x}_i,\ldots) - \widehat{\mathrm{MMD}}_u^2(\ldots,\mathbf{x}_i',\ldots)\right| \leq \frac{4\kappa}{m},$$

and similarly $c_{y,j} \leq 4\kappa/n$. Thus:

$$\sum_{i=1}^{m} c_{x,i}^2 + \sum_{j=1}^{n} c_{y,j}^2 \leq m \cdot \frac{16\kappa^2}{m^2} + n \cdot \frac{16\kappa^2}{n^2} = 16\kappa^2\left(\frac{1}{m}+\frac{1}{n}\right).$$

McDiarmid's inequality gives:

$$\Pr\left\{\widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2 > t\right\} \leq \exp\left(-\frac{2t^2}{16\kappa^2(1/m+1/n)}\right) = \exp\left(-\frac{t^2 m_{\mathrm{eff}}}{8\kappa^2}\right),$$

where $m_{\mathrm{eff}} := mn/(m+n)$ is the **harmonic mean sample size**. When $m = n$, $m_{\mathrm{eff}} = m/2 \approx m_2$, recovering Theorem 10.

Let $\kappa := \sup_{\mathbf{x}\in\mathscr{X}} K(\mathbf{x},\mathbf{x})$ be an upper bound on the kernel diagonal (assume $0 \leq K(\mathbf{x}_i,\mathbf{x}_j) \leq \kappa$). The following concentration result (Theorem 10 in [Gretton et al., 2012]) provides a finite-sample bound.

Theorem 4.1 (Concentration of $\widehat{\mathrm{MMD}}_u^2$, Theorem 10 in [Gretton et al., 2012])**.** *Assume* $0 \leq K(\mathbf{x}_i,\mathbf{x}_j) \leq \kappa$*. Then for any* $t > 0$*:*

$$\Pr_{\mathbf{X},\mathbf{Y}}\left\{\widehat{\mathrm{MMD}}_u^2[\mathscr{H}_K,\mathbf{X},\mathbf{Y}] - \mathrm{MMD}^2[\mathscr{H}_K,\mathbb{P},\mathbb{Q}] > t\right\} \leq \exp\left(\frac{-t^2 m_2}{8\kappa^2}\right),$$

where $m_2 := \lfloor m/2 \rfloor$*. The same bound applies for deviations of* $-t$ *and below.*

Corollary 4.2 (Level-α test, Corollary 11 in [Gretton et al., 2012])**.** *A hypothesis test of level* α *for* $H_0 : \mathbb{P} = \mathbb{Q}$ *has the* ***acceptance region****:*

$$\widehat{\mathrm{MMD}}_u^2[\mathscr{H}_K,\mathbf{X},\mathbf{Y}] < \frac{4\kappa}{\sqrt{m}}\sqrt{\log(\alpha^{-1})}.$$

The test rejects H_0 *when* $\widehat{\mathrm{MMD}}_u^2$ *exceeds this threshold.*

Question. How to improve the bound? Using Hoeffding decomposition to write as a sum of independent random variables!

For $m = n$, pair as $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ and define:

$$h(\mathbf{z}_i, \mathbf{z}_j) := K(\mathbf{x}_i, \mathbf{x}_j) + K(\mathbf{y}_i, \mathbf{y}_j) - K(\mathbf{x}_i, \mathbf{y}_j) - K(\mathbf{x}_j, \mathbf{y}_i),$$

with $\mathbf{E}[h] = \mathrm{MMD}^2$ and $h \in [-2\kappa, 2\kappa]$. Group into $m_2 = \lfloor m/2 \rfloor$ disjoint pairs:

$$\widehat{\mathrm{MMD}}_u^2 \approx \frac{1}{m_2} \sum_{i=1}^{m_2} \underbrace{h(\mathbf{z}_{2i-1}, \mathbf{z}_{2i})}_{\text{i.i.d., range } [-2\kappa, 2\kappa],\ \mathrm{Var}=\sigma^2}.$$

Apply Bennett's inequality (variance-aware) to get:

$$\Pr\left\{ \widehat{\mathrm{MMD}}_u^2 - \mathrm{MMD}^2 > t \right\} \leq \exp\left(-\frac{m_2 t^2/2}{\sigma^2 + 2\kappa t/3} \right),$$

where $\sigma^2 = \mathrm{Var}(h(\mathbf{z}_i, \mathbf{z}_j))$. This improves over Theorem 10 when $\sigma^2 \ll \kappa^2$.

Asymptotic test based on the unbiased estimator

Theorem 4.3 (Asymptotic distribution under H_0, Theorem 12 in [Gretton et al., 2012])**.** *Let*

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) := k(\mathbf{x}_i, \mathbf{x}_j) - \mathbf{E}_{\mathbb{P}} k(\mathbf{x}_i, \mathbf{X}) - \mathbf{E}_{\mathbb{P}} k(\mathbf{X}, \mathbf{x}_j) + \mathbf{E}_{\mathbb{P}\otimes\mathbb{P}} k(\mathbf{X}, \mathbf{X}')$$

be the centered kernel. Under $H_0 : \mathbb{P} = \mathbb{Q}$, as $m, n \to \infty$ with $t = m + n$ and $m/t \to p_\infty \in (0,1)$:

$$t\, \widehat{\mathrm{MMD}}_u^2 \xrightarrow{d} \sum_{l=1}^{\infty} \tilde{\lambda}_l \left[\left(p_\infty^{-1/2} a_l - (1-p_\infty)^{-1/2} b_l \right)^2 - p_\infty^{-1} - (1-p_\infty)^{-1} \right],$$

where $\{a_l\}$ and $\{b_l\}$ are independent sequences of i.i.d. $\mathcal{N}(0,1)$ variables, and $\tilde{\lambda}_l$ are the eigenvalues of the centered kernel operator:

$$\int \tilde{k}(\mathbf{x}, \mathbf{x}')\, \psi_l(\mathbf{x})\, d\mathbb{P}(\mathbf{x}) = \tilde{\lambda}_l\, \psi_l(\mathbf{x}').$$

*This distribution is generally intractable, so **permutation tests** or **bootstrap** are used to obtain critical values in practice.*

More details on the asymptotic distribution and its implications for test power can be found in [Gretton et al., 2012].

5 Applications of MMD

5.1 Domain Adaptation

In transfer learning, the source distribution $\mathbb{P}_s$ (labeled) and target distribution $\mathbb{P}_t$ (unlabeled) often differ — this is called **covariate shift**. MMD provides a principled measure of this domain gap:

$$\mathrm{MMD}[\mathscr{H}_K, \mathbb{P}_s, \mathbb{P}_t] = \| \mu_{\mathbb{P}_s} - \mu_{\mathbb{P}_t} \|_{\mathscr{H}_K}.$$

Setup. Given labeled source data $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s} \sim \mathbb{P}_s$ and unlabeled target data $\{\mathbf{x}_j^t\}_{j=1}^{n_t} \sim \mathbb{P}_t$, the goal is to learn a predictor f that generalizes to the target domain.

MMD-regularized ERM. Minimize the joint objective:

$$\min_f \underbrace{\frac{1}{n_s}\sum_{i=1}^{n_s} \ell(f(\mathbf{x}_i^s), y_i^s)}_{\text{source empirical risk}} + \lambda \cdot \underbrace{\widehat{\mathrm{MMD}}_u^2[\mathscr{H}_K, \mathbf{X}^s, \mathbf{X}^t]}_{\text{domain mismatch penalty}},$$

where the second term penalizes distributional shift between source and target feature distributions.

Intuition. If f maps inputs to a representation space, minimizing MMD encourages the representation distributions of $\mathbb{P}_s$ and $\mathbb{P}_t$ to align. A small MMD means the source-trained model is likely to transfer well. This idea underpins methods such as Deep Adaptation Networks (DAN) [Long et al., 2015] and DANN.

5.2 GANs and Integral Probability Metrics

Original GAN and f-divergences. Let $\mathbf{X} \sim \mathbb{P}_{\text{real}}$ be a real data sample, $\mathbf{Z} \sim p_Z$ be a noise input (e.g. $\mathcal{N}(0, I)$), and $G_\theta(\mathbf{Z}) \sim \mathbb{P}_{G_\theta}$ be a generated sample. The discriminator $D_\phi : \mathscr{X} \to [0,1]$ estimates the probability that its input is real.

The original GAN [Goodfellow et al., 2014] trains via:

$$\min_{G_\theta} \max_{D_\phi} \underbrace{\mathbf{E}_{\mathbb{P}_{\text{real}}}[\log D_\phi(\mathbf{X})]}_{\text{real samples scored high}} + \underbrace{\mathbf{E}_{p_Z}[\log(1 - D_\phi(G_\theta(\mathbf{Z})))]}_{\text{fake samples scored low}}.$$

At the optimal discriminator D_ϕ^*, the generator minimizes $\mathrm{JSD}(\mathbb{P}_{\text{real}} \| \mathbb{P}_{G_\theta})$.

More generally, **f-GAN** [Nowozin et al., 2016] unifies GAN objectives via f-divergences. Using the **Fenchel dual** $f^*(t) = \sup_u [tu - f(u)]$ of a convex function f, the f-divergence $D_f(\mathbb{P} \| \mathbb{Q})$ admits the variational (saddle-point) form:

$$D_f(\mathbb{P} \| \mathbb{Q}) = \sup_{T : \mathscr{X} \to \mathrm{dom}(f^*)} \left[\mathbf{E}_{\mathbb{P}}[T(\mathbf{X})] - \mathbf{E}_{\mathbb{Q}}[f^*(T(\mathbf{X}'))] \right],$$

where $\mathbf{X} \sim \mathbb{P}$ (real) and $\mathbf{X}' \sim \mathbb{Q}$ (generated), and T plays the role of the discriminator. For the choice $f(u) = u \log u - (u+1) \log \frac{u+1}{2}$, this recovers the JS-divergence and the original GAN objective.

Problem with f-divergences. When $\mathbb{P}_{\text{real}}$ and $\mathbb{P}_{G_\theta}$ have *disjoint supports* (common during early GAN training):

$$\mathrm{JSD}(\mathbb{P}_{\text{real}} \| \mathbb{P}_{G_\theta}) = \log 2 \quad \text{(constant, gradient = 0)}.$$

The generator receives no useful gradient signal — this is the **vanishing gradient problem** of the original GAN.

W-GAN: replacing f-divergence with an IPM. **Wasserstein GAN** [Arjovsky et al., 2017] replaces the JS-divergence with the **Wasserstein-1 distance**, which is an Integral Probability Metric (IPM):

$$W_1(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_{\text{Lip}}\le 1} \left(\mathbf{E}_{\mathbb{P}}(f(\mathbf{X})) - \mathbf{E}_{\mathbb{Q}}(f(\mathbf{Y}))\right).$$

(This is the Kantorovich–Rubinstein dual of the optimal transport problem.) The W-GAN objective is:

$$\min_{G_\theta} \max_{\|D_\phi\|_{\text{Lip}}\le 1} \mathbf{E}_{\mathbb{P}_{\text{real}}}(D_\phi(\mathbf{X})) - \mathbf{E}_{\mathbb{P}_{G_\theta}}(D_\phi(G_\theta(\mathbf{Z}))).$$

Crucially, $W_1 > 0$ and has *non-zero gradients* even when supports are disjoint, solving the vanishing gradient problem.

MMD-GAN: RKHS as the function class. **MMD-GAN** [Li et al., 2017] replaces the Lipschitz constraint with the RKHS unit ball:

$$\min_{G_\theta} \text{MMD}^2[\mathscr{H}_K, \mathbb{P}_{\text{real}}, \mathbb{P}_{G_\theta}] = \min_{G_\theta} \|\mu_{\mathbb{P}_{\text{real}}} - \mu_{\mathbb{P}_{G_\theta}}\|^2_{\mathscr{H}_K}.$$

The RKHS constraint gives a **closed-form differentiable** objective (no adversarial optimization needed), and the kernel can be learned end-to-end.

Unified IPM view.

GAN variant	Function class $\mathscr{F}$	Distance	Training
GAN / f-GAN	Unconstrained D (via Fenchel dual)	f-divergence	Adversarial
W-GAN	1-Lipschitz NN	W_1 (IPM)	Adversarial
MMD-GAN	RKHS unit ball	MMD (IPM)	Closed-form

5.3 Generative Model Evaluation

MMD provides a principled distance between the real data distribution $\mathbb{P}_{\text{real}}$ and the distribution $\mathbb{P}_{G_\theta}$ induced by a generative model G_θ:

$$\text{MMD}^2[\mathscr{H}_K, \mathbb{P}_{\text{real}}, \mathbb{P}_{G_\theta}] = \|\mu_{\mathbb{P}_{\text{real}}} - \mu_{\mathbb{P}_{G_\theta}}\|^2_{\mathscr{H}_K}.$$

Empirical evaluation. Given n_r real samples $\{\mathbf{x}_i^r\}$ and n_g generated samples $\{\mathbf{x}_j^g\} = \{G_\theta(\mathbf{z}_j)\}$, the unbiased MMD estimator provides a practical evaluation metric:

$$\widehat{\text{MMD}}^2_u[\mathbf{X}^r, \mathbf{X}^g; K] = \frac{1}{n_r(n_r-1)} \sum_{i\neq j} K(\mathbf{x}_i^r, \mathbf{x}_j^r) + \frac{1}{n_g(n_g-1)} \sum_{i\neq j} K(\mathbf{x}_i^g, \mathbf{x}_j^g) - \frac{2}{n_r n_g} \sum_{i,j} K(\mathbf{x}_i^r, \mathbf{x}_j^g).$$

Advantages over likelihood-based metrics.

- **Model-agnostic**: Works for any implicit generative model (GANs, flow-based, diffusion models) where the density $p_\theta(\mathbf{x})$ is unavailable.
- **No mode collapse detection bias**: Unlike FID (which relies on a pretrained Inception network), MMD with a characteristic kernel detects *all* distributional differences.
- **Statistical guarantees**: Concentration bounds (Theorem 10) provide confidence intervals on the estimated MMD.

References

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

[Li et al., 2017] Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30.

[Long et al., 2015] Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105.

[Nowozin et al., 2016] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29.

[Steinwart, 2001] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93.