Chapter 1: Mathematical Prerequisites

Throughout the course, we use the following notation conventions:

- $P, Q, \dots$ denote probability distributions (measures). $\widehat{P}, \widehat{Q}, \dots$ denote estimated distributions.
- We write $P \ll Q$ if measure P is absolutely continuous with respect to Q (meaning $Q(A) = 0 \Rightarrow P(A) = 0$ for any measurable set A).
- If a distribution P is absolutely continuous with respect to a reference measure (typically Lebesgue measure), we write its density (defined as the Radon-Nikodym derivative) as p, i.e., $\mathrm{d}P(x) = p(x)\mathrm{d}x$.
- We use $\mathbb{E}_{X\sim P}[\phi(X)]$ or $\int \phi(x)\mathrm{d}P(x)$ interchangeably to denote expectations.
- Commonly used distributions are denoted as Gaussian ($\mathsf{N}(\mu, \Sigma)$), Uniform ($\mathrm{Unif}([0,1])$), and Bernoulli ($\mathrm{Ber}(p)$).
- Random variables as capitalized, e.g., X and Y. We denote $P(X, Y)$ as the joint distribution of (X, Y), and $P(X)$ (resp. $P(Y)$) as the marginal distribution of X (resp. Y).

1 Discrepancy between Distributions

A recurring task in generative modeling is to compare two probability distributions: the data distribution P_{data} and a learned distribution $\widehat{P}$. This comparison is formalized through *discrepancy measures*. Importantly, different discrepancies capture different notions of "closeness" and are generally *not equivalent*. In generative modeling, this choice matters: it affects how one defines a training objective, how one evaluates a model, and what kind of theoretical guarantees are meaningful.

1.1 Definitions

In this section we introduce four major measures: KL divergence, total variation distance, Wasserstein distance, and f-divergence.

Definition 1.1 (KL divergence)**.** Let P, Q be probability measures with $P \ll Q$. The Kullback–Leibler divergence is

$$\mathrm{KL}(P, Q) = \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P \in [0, +\infty],$$

where $\mathcal{X}$ is the support. In particular,

- if P, Q are discrete distributions, then $\mathrm{KL}(P, Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$;
- if P, Q admit densities p, q, then $\mathrm{KL}(P, Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$.

KL divergence quantifies how far Q deviates from P in a *log-likelihood* sense. It is nonnegative (why?) and equals 0 if and only if $P = Q$ (almost surely), but it is NOT symmetric and is not a metric.

KL divergence is central to maximum likelihood estimation and information theory. Diffusion models are often derived from minimizing the KL divergence between the learned distribution and the empirical data distribution.

Definition 1.2 (Total Variation (TV) distance)**.** For probability measures P, Q, the total variation distance is defined as

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|,$$

where A is a measurable set.

By the definition, TV distance measures the largest possible discrepancy between probabilities assigned to the same event. It therefore controls the worst-case difference in behavior under P v.s. Q over all measurable sets, and more generally over all bounded test functions (see Proposition 1.8). TV distance appears frequently in stochastic processes and mixing time analysis.

Definition 1.3 (Wasserstein distance)**.** Let $c(\cdot, \cdot)$ be a transportation cost and let $p \geq 1$. For probability measures P, Q with finite p-th moments, the p-Wasserstein distance is

$$W_p(P, Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y)^p \mathrm{d}\pi(x, y) \right)^{1/p},$$

where $\Pi(P, Q)$ is the set of couplings of (P, Q), i.e., probability measures π on $\mathcal{X} \times \mathcal{X}$ whose marginals are P and Q, respectively.

Wasserstein distance looks quite technical and evaluating it is usually challenging. However, they are known as *optimal transport* distances, which quantify how much "effort"—measured by the cost function—it takes to transport probability mass. Unlike KL/TV, Wasserstein distance depends on the cost function. A commonly used cost function is the Euclidean distance and in the sequel, we adopt this cost function.

Definition 1.4 (f-divergence)**.** Let function $f : \mathbb{R}_+ \to \mathbb{R}$ be convex with $f(1) = 0$. For $P \ll Q$, the f-divergence is defined as

$$D_f(P, Q) = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q.$$

In particular,

- if P, Q are discrete distributions, then $D_f(P, Q) = \sum_{x \in \mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right)$;
- if P, Q have densities p, q, then $D_f(P, Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$.

With the flexibility on the choice of function f, f-divergence forms a broad class of discrepancies that compare distributions. Many classical divergences are instances of f-divergences.

Example 1.5 (KL and TV as f-divergences)**.**

- KL divergence is an f-divergence with $f(t) = t \log t$:

$$D_f(P, Q) = \int \frac{\mathrm{d}P}{\mathrm{d}Q} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P = \mathrm{KL}(P, Q).$$

- Total variation is an f-divergence with $f(t) = \frac{1}{2}|t - 1|$:

$$D_f(P, Q) = \frac{1}{2} \int \mathrm{d}Q \left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right| = \frac{1}{2} \int |\mathrm{d}P - \mathrm{d}Q| = \mathrm{TV}(P, Q),$$

where the last equality invokes Proposition 1.7.

Remark 1.6 (Metrics v.s. divergences)**.** Total variation and Wasserstein distances are metrics on suitable spaces of probability measures (TV always; W_p on measures with finite p-th moments). In contrast, KL divergence and general f-divergences are usually *not* metrics: they are typically asymmetric and do not satisfy the triangle inequality. Nevertheless, divergences are often more directly connected to likelihood-based learning objectives.

1.2 Operational Characterizations

The following characterizations explain what TV and Wasserstein control in terms of discrepancies of expectations. These forms are ubiquitous in theory and in algorithm design.

Proposition 1.7 (TV equals half the L^1 distance of densities)**.** For distributions P, Q, it holds that

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int_{\mathcal{X}} |\mathrm{d}P - \mathrm{d}Q|.$$

In particular,

- if P, Q are discrete distributions, then $\int_{\mathcal{X}} |\mathrm{d}P - \mathrm{d}Q| = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$;
- if P, Q admit densities p, q, then $\int_{\mathcal{X}} |\mathrm{d}P - \mathrm{d}Q| = \int |p(x) - q(x)| \mathrm{d}x$.

Proof. We present a proof for the continuous distributions and leave the discrete case as an exercise. The goal is to show

$$\sup_A |P(A) - Q(A)| = \frac{1}{2} \int |p(x) - q(x)| \mathrm{d}x.$$

We define a set

$$A^* = \{x \in \mathcal{X} : p(x) \geq q(x)\},$$

where is measurable. We first show that $\sup_A P(A) - Q(A) = P(A^*) - Q(A^*)$. Indeed, for any measurable set A,

$$P(A) - Q(A) = \int_A (p(x) - q(x)) \mathrm{d}x = \int_{A \cap A^*} (p(x) - q(x)) \mathrm{d}x + \int_{A \cap (A^*)^{\complement}} (p(x) - q(x)) \mathrm{d}x.$$

On A^*, the integrand is nonnegative; on $(A^\star)^{\complement}$, however, it is nonpositive. Hence, we have

$$P(A) - Q(A) \leq \int_{A \cap A^*} (p(x) - q(x)) \mathrm{d}x \leq \int_{A^\star} (p(x) - q(x)) \mathrm{d}x = P(A^*) - Q(A^*).$$

Therefore, it holds that $\sup_A P(A) - Q(A) = P(A^*) - Q(A^*)$.

Similarly, $\inf_A P(A) - Q(A) = P((A^*)^\complement) - Q((A^*)^\complement)$, and we claim that

$$P(A^*) - Q(A^*) = -\big(P((A^*)^\complement) - Q((A^*)^\complement)\big).$$

As a result, we have

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)| = P(A^*) - Q(A^*).$$

The last step is to plug in the density functions of P, Q:

$$\begin{aligned}
\sup_A |P(A) - Q(A)| &= P(A^*) - Q(A^*) \\
&= \int_{A^*} (p(x) - q(x))\mathrm{d}x \\
&= \frac{1}{2}\left[\int_{A^*} (p(x) - q(x))\mathrm{d}x - \int_{(A^*)^\complement} (p(x) - q(x))\mathrm{d}x\right] \\
&= \frac{1}{2}\int |p(x) - q(x)|\mathrm{d}x.
\end{aligned}$$

The proof is complete. □

It is also convenient to use a variational formula for representing discrepancy measures, which is critical for deriving many generative models such as Generative Adversarial Networks (GANs). These variational forms are also known as the dual characterizations.

Proposition 1.8 (Dual characterization of TV)**.** For probability measures P, Q, we have,

$$\mathrm{TV}(P, Q) = \sup_{\|\phi\|_\infty \le 1/2} \left|\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]\right|.$$

Here, $\|\phi\|_\infty = \max_x |\phi(x)|$.

Before we present a proof, we remark that the function ϕ can be viewed as a test function concentrating on the difference between P and Q. The class of test functions plays a vital role in defining discrepancies. The richer the test function class, the stronger the discrepancy measure. Of course, we don't want to use constant test functions, as they can tell nothing about the difference in P, Q.

Proof. In order to show some equality like $a = b$, we usually show $a \le b$ and $b \le a$. Here, one direction is straightforward. We have

$$\begin{aligned}
\sup_{\|\phi\|_\infty \le 1/2} |\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]| &= \sup_{\|\phi\|_\infty \le 1/2} \left|\int \phi(x)\mathrm{d}(P(x) - Q(x))\right| \\
&\le \int \frac{1}{2}|\mathrm{d}P(x) - \mathrm{d}Q(x)| \\
&= \mathrm{TV}(P, Q). \qquad (1.1)
\end{aligned}$$

For the other direction, we construct a function ϕ acting like an indicator function. Recall we define set A^* in Proposition 1.7. We let

$$\phi^*(x) = \frac{1}{2} \cdot \mathbb{1}\{x \in A^*\},$$

where $\mathbb{1}$ is the indicator function. It is clear that $\|\phi^*\|_\infty \leq 1/2$. By the same argument in Proposition 1.7, we can show

$$\mathrm{TV}(P, Q) = |\mathbb{E}_P[\phi^*(X)] - \mathbb{E}_Q[\phi^*(X)]| \leq \sup_{\|\phi\|_\infty \leq 1/2} |\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]|. \tag{1.2}$$

Combining (1.1) and (1.2), we obtain the desired equality. □

We now shift to Wasserstein distance and demonstrate that the variational formula is unifying by changing the test function class. Although it is possible to derive dual characterizations for Wassersteion W_p ($p \geq 2$) distances, they are highly complicated. A nice result is established for W_1 distance.

Theorem 1.9 (Kantorovich-Rubinstein duality for Wasserstein W_1 distance)**.** For probability measures P, Q and cost function $c(x, y) = \|x - y\|_2$, it holds that

$$W_1(P, Q) = \sup_{\mathrm{Lip}(\phi) \leq 1} |\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]|,$$

where $\mathrm{Lip}(\phi) \leq 1$ means $|\phi(x) - \phi(y)| \leq \|x - y\|_2$ for all $x, y \in \mathcal{X}$.

The primal definition of W_1 is an infimum over couplings, which is typically hard to evaluate. The dual form expresses W_1 as the worst-case difference of expectations over 1-Lipschitz test functions. This dual viewpoint is fundamental both theoretically and algorithmically (e.g., it motivates adversarial objectives based on Lipschitz critics). A full proof requires tools from convex analysis and optimal transport; we treat Theorem 1.9 as a standard result. A good reference book is [3].

1.3 Integral Probability Metrics (IPMs)

The dual characterizations above suggest a unifying viewpoint: many useful discrepancies can be expressed as a supremum of expectation differences over a *class of test functions*. This leads to the notion of *Integral Probability Metrics* (IPMs).

Definition 1.10 (Integral probability metric)**.** Let $\mathcal{F}$ be a class of measurable functions $\phi : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_P|\phi(X)| < \infty$ and $\mathbb{E}_Q|\phi(X)| < \infty$ for all $\phi \in \mathcal{F}$. The IPM induced by $\mathcal{F}$ is

$$\mathsf{d}_{\mathcal{F}}(P, Q) = \sup_{\phi \in \mathcal{F}} |\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]|.$$

The choice of $\mathcal{F}$ determines what differences between P and Q are detectable. If $\mathcal{F}$ contains very oscillatory functions, then the IPM can be sensitive to fine-grained differences (e.g., support mismatch). If $\mathcal{F}$ contains only smooth functions, the IPM emphasizes coarse, geometric differences (e.g., shifts in mean, mass transportation).

Table 1 summarizes several frequently used IPMs by listing the test function class $\mathcal{F}$ and the resulting metric. The central message is that *the discrepancy is fully determined by the choice of the test function class.*

Table 1: A cheat sheet of integral probability metrics. Different choices of the test class $\mathcal{F}$ induce different notions of distributional discrepancy.

Name	**Test function class $\mathcal{F}$**	**IPM $d_{\mathcal{F}}(P, Q)$**
Total variation (TV)	$\mathcal{F}_{\mathrm{TV}} = \{\phi : \|\phi\|_\infty \le 1/2\}$	$\mathrm{TV}(P, Q)$ (Prop. 1.8)
Wasserstein-1	$\mathcal{F}_{W_1} = \{\phi : \mathrm{Lip}(\phi) \le 1\}$ on $(\mathcal{X}, \|\cdot\|_2)$	$W_1(P, Q)$ (Thm. 1.9)
MMD (kernel IPM)	$\mathcal{F}_{\mathrm{MMD}} =$ RKHS unit ball	$\mathrm{MMD}(P, Q)$
Neural IPM	$\mathcal{F}_{\mathrm{NN}} =$ neural networks	$\mathsf{d}_{\mathcal{F}_{\mathrm{NN}}}(P, Q)$ (GANs)

- TV uses the richest class among the above (all bounded tests), so it is extremely sensitive to support mismatch. This strength can be a weakness in high dimensions though.
- Wasserstein-1 restricts tests to be Lipschitz, which incorporates geometry and is often more stable when supports do not overlap, but it requires moment/geometry assumptions for finiteness.
- MMD is an IPM defined by a Reproducing Kernel Hilbert Space (RKHS); it is smooth, easy to estimate from samples, and can be *characteristic* (i.e., $\mathrm{MMD}(P, Q) = 0$ implies $P = Q$) for appropriate kernels. MMD is widely used for two-sample test and has a simple empirical estimator. Interested readers should refer to the seminal paper [1] for details.
- Neural IPMs arise in adversarial learning: one chooses a parameterized critic class (often neural networks) and trains the critic to maximize the discrepancy while training the generator to minimize it.

1.4 Relationships and Conversions

We articulate some key relationships (how one discrepancy upper bounds another) and non-relationships (counterexamples showing that a bound in one discrepancy need not imply a bound in another).

We first present several important inequalities that convert a bound in one discrepancy into a bound in another. These results are frequently used when a training objective controls one divergence, but evaluation or theory is stated in another.

Lemma 1.11 (Pinsker's inequality)**.** If distributions $P \ll Q$, then it holds that

$$\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2} \mathrm{KL}(P, Q)}.$$

Proof. Let $r = \frac{\mathrm{d}P}{\mathrm{d}Q}$ be the Radon-Nikodym derivative so that $\int r \, \mathrm{d}Q = 1$, and write

$$\mathrm{KL}(P, Q) = \int r \log r \, \mathrm{d}Q.$$

We also have

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q| = \frac{1}{2}\int |r-1|\mathrm{d}Q.$$

Recall that these are the corresponding f-divergence representations of KL and TV. Define the function

$$\psi(t) = t\log t - t + 1 \quad \text{with} \quad t \geq 0.$$

Some calculus check shows that $\psi(t) \geq 0$ (we observe $\psi(1) = 0$ and $\psi'(t) = \log t$.) We claim that

$$\left(\frac{4}{3} + \frac{2}{3}t\right)\psi(t) \geq (t-1)^2 \quad \text{for} \quad t \geq 0,$$

and show $g(t) \leq 0$. To see this, we let

$$g(t) = (t-1)^2 - \left(\frac{4}{3} + \frac{2}{3}t\right)\psi(t).$$

It is clear that $g(1) = 0$, $g'(1) = 0$, and $g''(t) = -\frac{4}{3t}\psi(t) \leq 0$, which says g is a concave function. For any t, we apply a Taylor expansion to $g(t)$ around 1. Let ξ interpolates between t and 1; we have

$$\begin{aligned} g(t) &= g(1) + g'(1)(x-1) + \frac{g''(\xi)}{2}(x-1)^2 \\ &= -\frac{4}{6t}\psi(t) \\ &\leq 0. \end{aligned}$$

Now we have

$$\begin{aligned} \mathrm{TV}(P,Q) &= \frac{1}{2}\int |r-1|\,\mathrm{d}Q \\ &\leq \frac{1}{2}\int \sqrt{\left(\frac{4}{3} + \frac{2}{3}r\right)\psi(r)}\mathrm{d}Q \\ &\leq \frac{1}{2}\sqrt{\int \left(\frac{4}{3} + \frac{2}{3}r\right)\mathrm{d}Q}\sqrt{\int \psi(r)\mathrm{d}Q} \qquad \text{(Cauchy-Schwarz)} \\ &= \sqrt{\frac{1}{2}\int \psi(r)\mathrm{d}Q}. \end{aligned}$$

The last step is to observe that

$$\int \psi(r)\mathrm{d}Q = \int r\log r\mathrm{d}Q - \int (r-1)\mathrm{d}Q = \int r\log r\mathrm{d}Q = \mathrm{KL}(P,Q).$$

The proof is complete. □

The first Pinsker's inequality is exact in the sense that there exist probability measures P and Q for which it becomes equality. However,it is nontrivial only if $\mathrm{KL}(P,Q) \leq 2$, since we always have $\mathrm{TV}(P,Q) \leq 1$. A nontrivial result for larger KL divergence is

$$\mathrm{TV}(P,Q) \leq 1 - \frac{1}{2}\exp(-\mathrm{KL}(P,Q)).$$

For simplicity, we omit the proof here and a formal statement can be found in Lemma 2.6 of [2].

The second result is about TV and Wasserstein distance.

Proposition 1.12 (TV controls W_1 on bounded metric spaces)**.** If the diameter of support $\mathcal{X}$ is bounded, i.e., $\operatorname{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} \|x-y\|_2 < \infty$, then

$$W_1(P,Q) \le 2\operatorname{diam}(\mathcal{X}) \cdot \mathrm{TV}(P,Q).$$

Proof. By the dual characterization in Theorem 1.9, we have

$$W_1(P,Q) = \sup_{\mathrm{Lip}(\phi)\le 1} \left|\mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)]\right|.$$

Fix any 1-Lipschitz ϕ. Choose $x_0 \in \mathcal{X}$ and define $\widetilde{\phi}(x) = \phi(x) - \phi(x_0)$. Then $\widetilde{\phi}$ is still 1-Lipschitz and it holds that

$$\mathbb{E}_P[\widetilde{\phi}(X)] - \mathbb{E}_Q[\widetilde{\phi}(X)] = \mathbb{E}_P[\phi(X)] - \mathbb{E}_Q[\phi(X)].$$

Moreover, for any $x \in \mathcal{X}$, it holds that

$$|\widetilde{\phi}(x)| = |\phi(x) - \phi(x_0)| \le \|x - x_0\|_2 \le \operatorname{diam}(\mathcal{X}),$$

hence $\|\widetilde{\phi}\|_\infty \le \operatorname{diam}(\mathcal{X})$. Now applying Proposition 1.8 (with $\widetilde{\phi}/\operatorname{diam}(\mathcal{X})$) yields

$$\left|\mathbb{E}_P[\widetilde{\phi}(X)] - \mathbb{E}_Q[\widetilde{\phi}(X)]\right| \le 2\mathrm{TV}(P,Q) \cdot \|\widetilde{\phi}\|_\infty \le 2\operatorname{diam}(\mathcal{X}) \cdot \mathrm{TV}(P,Q).$$

Taking the supremum over all 1-Lipschitz function ϕ proves the assertion. □

1.5 Non-implications and Counterexamples

The inequalities above should not be misinterpreted as equivalences. In general, no single discrepancy dominates all others. The following counterexamples are particularly instructive.

Example 1.13 (Small Wasserstein distance does not imply small TV)**.** Let $\mathcal{X} = \mathbb{R}$. Consider P concentrated on a single point $x = 0$ and Q concentrationed on $x = \epsilon$. Then

$$W_1(P,Q) = |\epsilon| \to 0 \quad \text{as} \quad \epsilon \to 0,$$

but

$$\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)| = 1 \quad \text{for any} \quad \epsilon \ne 0.$$

An intuitive explanation is that Wasserstein distance captures the *distance* between supports (here a shift by ϵ), while TV captures *whether supports overlap* (here they do not). Thus a tiny geometric shift can yield maximal TV.

Example 1.14 (Small TV does not imply small Wasserstein distance on unbounded spaces)**.** Let $P = (0,1)$ be Gaussian on $\mathbb{R}$ and define a contaminated distribution

$$Q_{\epsilon,M} = (1-\epsilon)\mathsf{N}(0,1) + \epsilon\ \mathsf{N}(M,1), \quad \text{with} \quad 0 < \epsilon < 1 \text{ and } M \ne 0.$$

For fixed small ϵ, the Wasserstein distance between P and $Q_{\epsilon,M}$ can be made arbitrarily large by taking M large. Intuitively, an ϵ-fraction of the probability mass must move a distance $|M|$. On the other hand, TV is small since only an ϵ portion of the distribution is perturbed.

This example amplifies the intuition: TV measures the *amount* of mass that changes, but not *how far* it moves; Wasserstein measures both amount and distance of transport.

Example 1.15 (KL can be infinite while other notions are finite)**.** If $P \not\ll Q$, then $\mathrm{KL}(P, Q) = +\infty$. For instance, let P be the point mass concentrated on $x = 0$ and $Q = \mathsf{N}(0, 1)$ on $\mathbb{R}$. Then $\mathrm{KL}(P, Q) = +\infty$, while $W_1(P, Q) < \infty$ and TV is well-defined (indeed $\mathrm{TV}(P, Q) = 1$).

KL (and many f-divergences) depends on the Radon-Nikodym derivative $\mathrm{d}P/\mathrm{d}Q$ and therefore requires absolute continuity. When P assigns mass to a set where Q assigns zero mass, $\mathrm{d}P/\mathrm{d}Q$ is ill-defined and KL becomes infinite. This sensitivity is one reason why KL can be too strong in settings with singular measures or near-disjoint supports.

1.6 (Optional) More relationships under Additional Assumptions

The counterexamples above convey an important message: *without additional structure*, different discrepancies are incomparable. However, in many learning problems we *do* have extra information about the distributions of interest—for example, bounded support, controlled tails, smooth densities, or bounded density ratios. Under such assumptions, one can often *convert* one discrepancy into another. The purpose of this subsection is to give a high-level impression of when such conversions are possible, rather than to present a full technical treatment.

Case 1: bounded support and bounded test functions. Proposition 1.12 already illustrates a typical pattern: if $\mathcal{X}$ is bounded, then Lipschitz functions are automatically bounded (after centering), which allows one to translate between IPMs induced by different test function classes. More generally, on bounded domains, many distances become closer to each other because geometry and tail behavior are no longer issues.

Case 2: density ratios bounded away from 0 and ∞. A major obstacle for comparing f-divergences is that they depend on the Radon-Nikodym derivative $r = \mathrm{d}P/\mathrm{d}Q$, and r may be unbounded (or even undefined). If we impose a *ratio bound*

$$0 < m \le \frac{\mathrm{d}P}{\mathrm{d}Q} \le M < \infty \qquad Q \text{ almost surely for constants } m, M,$$

then many divergences become comparable.

Proposition 1.16 (Informal: local comparability of f-divergences)**.** Assume $P \ll Q$ and $m \le \mathrm{d}P/\mathrm{d}Q \le M$ Q-almost surely. Then for any two convex functions f, g with $f(1) = g(1) = 0$, there exist constants $0 < c_{f,g} \le C_{f,g} < \infty$ depending only on (f, g, m, M) such that

$$c_{f,g}\, D_g(P, Q) \;\le\; D_f(P, Q) \;\le\; C_{f,g}\, D_g(P, Q).$$

This proposition should be read as a qualitative statement: *if the density ratio never becomes too small or too large, then controlling one f-divergence controls all others up to constants.* In reinforcement learning and control, such density ratio is closely relevant to the concentrability coefficient, which measures the mismatch between trajectory measures under different behavior policies.

Case 3: tail conditions enable transport-type inequalities. On unbounded spaces, W_p depends strongly on tail behavior. If Q has light tails and strong concentration (e.g., sub-Gaussian behavior), then KL control can imply Wasserstein control through *transportation inequalities*. These results are widely used in probability, but their hypotheses are typically stated in terms of functional inequalities (e.g., log-Sobolev or Poincaré).

Proposition 1.17 (Informal: KL can control Wasserstein under concentration)**.** For certain reference measures Q with strong concentration, one can prove inequalities of the form

$$W_2(P,Q)^2 \;\le\; C\,\mathrm{KL}(P,Q),$$

for all $P \ll Q$, where C depends only on Q.

There is deep technical exposure underneath this result. Interested readers may refer to Theorem 22.15 in [3].

Case 4: converting between IPMs. Since IPMs are defined by test function classes, there is a simple monotonicity principle: if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $d_{\mathcal{F}_1}(P,Q) \le d_{\mathcal{F}_2}(P,Q)$. This observation is often used to reason about relative strength.

References

[1] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[2] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[3] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.