**Motivation**: Histograms / Density $\longrightarrow$ Comparing them, optimizing over them
$\longrightarrow \neq$ Vectors, $\neq$ Functions !! "WEAK OBJECTS"

**Measures**: • On a set $X$, $A \subset X \rightsquigarrow \mu(A) \in \mathbb{R}$ ($\mathbb{C}^{(+\infty)}$) size / Mass

(1)

**Positive**: $\mu(A) \geq 0$  Proba: $\mu(X) = 1$

$[\mu(A \cup B) = \mu(A) + \mu(B)$ if $A \cap B = \emptyset]$ — Lebesgue "area", Dirac, Density, Sums of Diracs!
$\hookrightarrow$ should extend to countable union. Der. / Cont.

**Radon measure**: to speak about convergence, one needs a distance (of pts & measures)

(1) Needs all balls (small enough) $\mu(B) < +\infty$  (2) $\mu(A) = \sup\{\mu(K) : \text{compact } K \subseteq A\}$

$\Rightarrow$ Allows to define $\int f \, d\mu$ for $f \in \mathcal{C}(X)$ continuous

**Proba measures & Random variables**: $Z : (\Omega, \mathbb{P}) \to X$ random var ($\hookrightarrow$ proba)

associated measure $\mu(A) = \mathbb{P}_Z(A) \triangleq \mathbb{P}(Z \in A)$

$Z$ is "push-forwarding" $\mathbb{P}$ to $\mathbb{P}_X$.

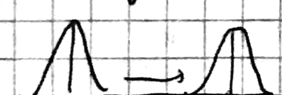
**Convergence of {random variables / Measures}**: convergence in law $\iff$ weak $*$ convergence

$\mu_n \xrightarrow{*} \mu \iff \forall A, \mu_n(A) \to \mu(A)$

$\mathbb{P}_{Z_n} \xrightarrow{\mathcal{L}} \mathbb{P}_Z \iff \forall f \in \mathcal{C}(X), \int f d\mu_n \to \int f d\mu$

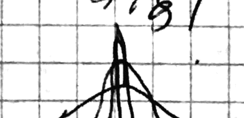
⚠ Weaker than the $\boxed{\text{STRONG}}$ convergence of the density $\mu_n = g_n dx$, $\mu = g dx$: $g_n \xrightarrow{L^1} g \Rightarrow \mu_n \to \mu$, $\not\Leftarrow$
($\mu_n$ & $\mu$ might not ~~be~~ have densities!)

**Example**:

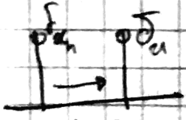


$\iff m_n \to m$

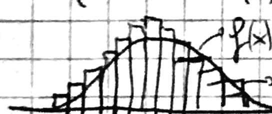
$N(m_n, \sigma)$  $N(m, \sigma)$



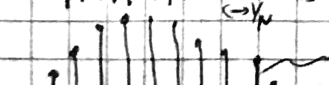
$N(0, \sigma_n) \to \delta_0$, $\sigma_n \to 0$



$x_n \to x$



$f(x)dx$, $f_n(x)dx$, $\longleftrightarrow 1/n$ — Finite element (fun.) approx:



$\longrightarrow \sum \left[\int_{k/n}^{\frac{k+1}{n}} f(x) dx\right] \delta_{k/n}$  Discrete approx.
$\hookrightarrow f(k/n) \cdot 1/n$

**Key question**: Quantifying the speed of convergence? $D(\mu, \nu)$ "distance like"

First requirement: $\mu_n \to \mu \iff D(\mu_n, \mu) \xrightarrow{n} 0$
$\iff D$ defines the weak topology

WARNING: in "$\infty$" dimension (and in particular if $D$ is not a norm)
$(D_1, D_2)$ same topology $\not\Leftrightarrow$ equivalent $\iff \exists C, D_2/C \leq D_1 \leq C D_2$
ie convergence rates depend on the distance!

# Statistical divergence

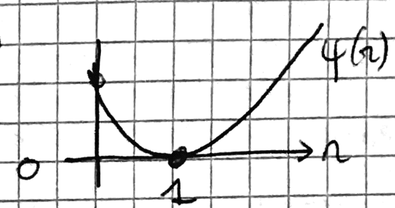


**1) Comparing density:** if $\begin{cases}\mu = f\,dx \\ \nu = g\,dx\end{cases}$ then compute $D(\mu,\nu) = \left(\int |f(x)-g(x)|^p\,dx\right)^{1/p} = \|f-g\|_{L^p}$

Pbm: • strong assumption
• Not continuous with weak topology
$D(\mu+\varepsilon\delta_0,\nu)$ not defined

**2) Comparing relative density:** $\varphi$-divergence / Csiszár divergence

Comparing $\frac{d\mu}{d\nu} = f$ (ie $d\mu(x) = f(x)\,d\nu$) with $1$




$$D(\mu|\nu) \triangleq \int_X \varphi\left(\frac{d\mu}{d\nu}(x)\right) d\nu(x)$$

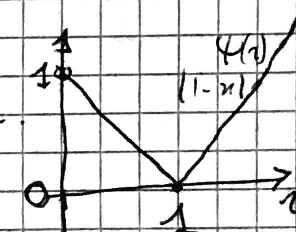
$\begin{cases}\varphi \text{ convex} \\ \varphi(1)=0\end{cases}$ $\quad D(\mu|\nu)=0 \Leftrightarrow \mu=\nu$

$D$ convex of $(\mu,\nu)$ !

example: $\varphi(x) = |x-1|$

$$D(\mu|\nu) = \underbrace{\int \left|\frac{d\mu}{d\nu} - 1\right| d\nu}_{TV(\mu-\nu)} = \int \left|\frac{d\mu}{dx} - \frac{d\nu}{dx}\right| dx = \left\|\frac{d\mu}{dx} - \frac{d\nu}{dx}\right\|_{L^1}$$

→ it is a norm !!

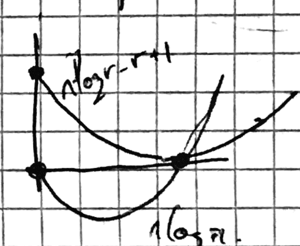



example: $\varphi(x) = x\log x$
KL divergence

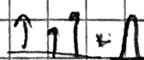
$$D(\mu|\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu$$

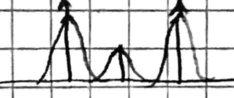
("Generalized" KL) $\left[\varphi(x) = x\log x - x + 1\right]$ $\quad D(\mu|\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu + \int (d\nu - d\mu)$

↳ • useful to make it also a Bregman divergence
↳ • same if $\int d\mu = \int d\nu$




**3) Hilbertian norm using lifting:** For simplicity, $X = \mathbb{R}^d$





Convolution / Kernel density estimator: $\mu * h$ as a density $\quad \int h(x-y)\,d\mu(y)$
($h(x)$ smooth $\int h = 1$)

ex $\mu = \sum a_k \delta_{x_k} \to \sum a_k h(x - x_k)$

$$D(\mu,\nu)^2 = \text{"}\|h*\mu - h*\nu\|_{L^2}^2\text{"} = \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} h(x-y)\,d\xi(y)\right]^2 dx \triangleq \|\mu-\nu\|_k^2$$

→ seems intractable even for discrete $\mu$ and $\nu$ $\quad$ Max Mean Discrepancy (MMD)

$$= \int_{\mathbb{R}^d} \iint h(x-y)\,h(x-y')\,d\xi(y)\,d\xi(y')\,dx \quad \boxed{\text{FUBINI}}$$

$$= \iint K(y,y')\,d\xi(y)\,d\xi(y') \qquad K(y,y') \triangleq \int_{\mathbb{R}^d} h(x-y)\,h(x-y')\,dx$$

↳ Kernel

$\boxed{\text{ex}}$ $h(x) = e^{-\|x\|^2/2\sigma^2} \rightsquigarrow K(y,y') = \exp(-\|y-y'\|^2/4\sigma^2)$

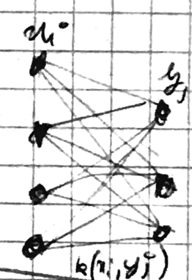
$\boxed{\text{Ex}}$ $K(y,y') = -\|y-y'\|$ also correspond to $h(x) = \frac{1}{\|x\|^{\dots}}$

So $D(\mu,\nu)^2 = \iint k(y,y')(d\mu(y) - d\nu(y))(d\mu(y') - d\nu(y'))$

$\boxed{\text{ex}}$ $\mu = \sum a_k \delta_{x_k}$, $\nu = \sum b_k \delta_{y_k}$ $\to$ $D(\mu,\nu)^2 = -2\sum_{ij} k(x_i, y_j) a_i b_j$ $= \mathbb{E}(k(X,X'))$

$+ \sum_{ii'} k(x_i, x_{i'}) a_i a_{i'}$ $+ \mathbb{E}(k(Y,Y'))$

$+ \sum_{jj'} k(y_j, y_{j'}) b_j b_{j'}$ $-2\,\mathbb{E}(k(X,Y))$




$\boxed{\text{Thm: } \|\cdot\|_k \text{ metrizes weak convergence}}$

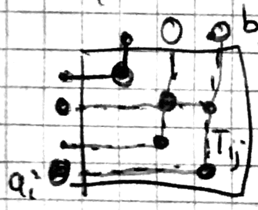
---

## Optimal Transport

Special discrete case : $\mu = \sum_i a_i \delta_{x_i}$ $\nu = \sum_j b_j \delta_{y_j}$ "Grain of masses"

$a_i, b_j \geq 0$, $\sum a_i = \sum_j b_j \; (=1)$.

Coupling : $\Pi(a,b) = \{ P \in \mathbb{R}_+^{n \times m} : \underbrace{\forall_i \sum_j P_{ij} = a_i}_{P\mathbb{1} = a}, \; \underbrace{\forall_j \sum_i P_{ij} = b_j}_{P^T\mathbb{1} = b} \}$ cvx Polytop




OT : $\mathcal{C}_c(\mu,\nu) = \min_{P \in \Pi(a,b)} \sum c_{ij} P_{ij}$ → Lin Prog / → Combinatorial optim / → Sinkhorn approx

Wasserstein dist : if $c_{ij} = d(x_i, y_j)^p$, $p \geq 1$, $W_p(\mu,\nu) \triangleq \mathcal{C}_c(\mu,\nu)^{1/p}$

Thm : $W_p$ is a distance & it metrizes weak convergence

---

MMD : ⊕ Simple ⊕ Good Sample cxity ⊖ Reflects less the distance $\|\mu - \mu(\cdot - \tau)\| = O(\sqrt{\delta})$

Sinkhorn

OT : ⊖ Cpx ⊖ Bad Sample cpxity ⊕ More geometrical

$|D(\mu,\nu) - D(\hat{\mu}_n, \hat{\nu}_n)| \sim \frac{1}{n^q}$ MMD: $q = 1/2$ OT: $q = 1/d$