From Monge to GANs

A Primer on Optimal Transport

Probability distrib°: $(\mathcal{X}, d)$ metric space eg $\mathcal{X} \subset \mathbb{R}^d$, $d(x,y) = \|x-y\|$

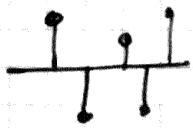
For simplicity, $\mathcal{X}$ compact (otherwise needs moment conditions)

Radon measures: $\alpha \in \mathcal{M}(\mathcal{X})$: $A \subset \mathcal{X} \rightsquigarrow \alpha(A) \in \mathbb{R}$ Borel measure with regularity condition

$\mathcal{M}(\mathcal{X}) = \mathcal{C}(\mathcal{X})^*$ dual of cont. funct° $\int_A d\alpha(x)$

ie $f \in \mathcal{C}(\mathcal{X}) \rightsquigarrow \int_{\mathcal{X}} f(x)\, d\alpha(x)$ is linear continuous

ex: $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ discrete measure $\int f \delta_x = f(x)$



2 "ways" to discretize → weight $(a_i)_i$: EULER; → points $(x_i)_i$: LAGRANGE

ex: $d\alpha(x) = \rho(x)\, dx \iff \frac{d\alpha}{dx} = \rho$

$\int f d\alpha = \int f \cdot \rho\, dx$

Proba measures: $\left[\alpha \geq 0 \text{ and } \int d\alpha = 1\right] \iff$ [Law of random variable X]

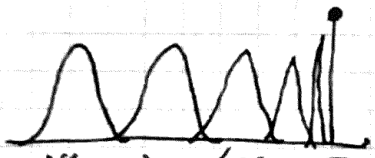
not° : $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ $\quad \mathbb{E}_X(f(X)) = \int f(x)\, d\alpha(x)$.

Weak* topology: $\alpha_n \xrightarrow{*} \alpha \iff \forall f,\ \int f d\alpha_n \to \int f d\alpha$

(cv in law) $X_n \xrightarrow{\mathcal{L}} X \iff \forall f,\ \mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$

$\iff \forall A \subset \mathcal{X},\ \mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$

ex:



$N(x_n, \sigma_n) \xrightarrow[x_n \to x]{\sigma_n \to 0} \delta_x$

$\delta_{x_n} \xrightarrow{*} \delta_x$

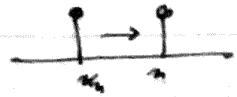
Comparing distribut°: $\varphi : \mathbb{R} \to \mathbb{R}$, convex, $\varphi(1) = 0$

$$D_\varphi(\alpha|\beta) \triangleq \int_{\mathcal{X}} \varphi\left(\frac{d\alpha}{d\beta}(x)\right) d\beta(x) + \alpha^\perp(\mathcal{X})\, \varphi'_\infty \qquad \text{Cizar } \varphi\text{-div}$$

Prop $D_\varphi(\cdot|\cdot) \geq 0$ is jointly cvx, $D_\varphi(\alpha|\beta) = 0 \iff \alpha = \beta$

ex:
- $\varphi(s) = s\log(s)$ $\quad D_\varphi = KL$ = rel. entropy
- $\varphi(s) = |1-s|$ $\quad D_\varphi(\alpha,\beta) = \left\|\frac{d\alpha}{dx} - \frac{d\beta}{dx}\right\|_{L^1}$ TV
- $\varphi(s) = \sqrt{1-s^2}$ $\quad D_\varphi(\alpha,\beta) = \left\|\sqrt{\frac{d\alpha}{dx}} - \sqrt{\frac{d\beta}{dx}}\right\|_{L^2}^2$ Hellinger^2 (Hellinger is a dist)
- $\varphi(s) = s\log(s) - (s+1)\log\left(\frac{s+1}{2}\right)$ $\quad D_\varphi(\alpha,\beta) = KL\left(\alpha \middle| \frac{\alpha+\beta}{2}\right) + KL\left(\beta \middle| \frac{\alpha+\beta}{2}\right)$ ($JS^{1/2}$ is a distance)

②



<u>Pbm</u> : D_{KL} doesn't metrize cv in law : $x_n \to x$, $\begin{cases} D_{KL}(\alpha|\beta) = +\infty \\ \|\alpha-\beta\|_{TV} = 2 \end{cases}$ $\alpha = \delta_{x_n}$, $\beta = \delta_x$

$\rightsquigarrow$ One wants $D(\alpha|\beta)$ such that $\alpha_n \rightharpoonup \alpha \iff D(\alpha_n|\alpha) \to 0$

<u>Kernel norms</u> : aka MMD (maximum Mean Discrepancy)

$k(x,y)$ positive kernels : $\forall (x_i)_{i=1}^n$ $(k(x_i,x_j))_{i=1}^n \succ 0$

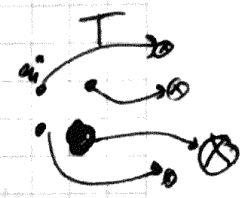
<u>then</u> : $\|\xi\|_k^2 \triangleq \iint_{\mathcal{X}^2} k(x,y)\, d\xi(x)\, d\xi(y)$ is a norm

<u>ex</u> : $\xi = \sum_{i=1}^m a_i \delta_{x_i}$ $\|\xi\|_k^2 = \sum_{i,j} a_i a_j k(x_i,x_j) \rightsquigarrow n^2$ open

<u>Prop</u> : if k universal (ie $\operatorname{span} k(x,\cdot)$ dense in $\mathcal{C}(\mathcal{X})$) then $\|\alpha-\beta\|_k^2$ metrizes weak convergence

<u>Ex</u> : Gaussian, ED

<u>Monge OT</u> : <u>Push-Forward</u> : $T : \mathcal{X} \to \mathcal{X}$ $T_\# : \delta_x \to \delta_{p(x)}$

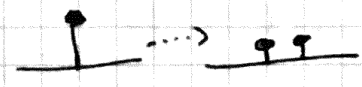


$\begin{cases} \text{By linearity : } T_\# : \sum a_i \delta_{x_i} \longrightarrow \sum a_i \delta_{p(x_i)} \\ \text{For measure } \beta = T_\# \alpha \text{ defined by } \forall f \in \mathcal{C}(\mathcal{X}),\ \int f(y)\, d\beta(y) = \int f(T(x))\, d\alpha(x) \\ \text{For random variables } Y \sim \beta = T_\# \alpha \text{ is defined as } Y = T(X). \end{cases}$

For density $\rightsquigarrow$ painful.

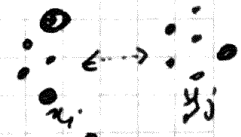
<u>Monge Pbm</u> : $\min_{T : \mathcal{X} \to \mathcal{X}} \left\{ \int_{\mathcal{X}} d(x, T(x))^p\, d\alpha(x) \ : \ T_\# \alpha = \beta \right\}$.

<u>Pbm</u> : In general, constraints are empty :



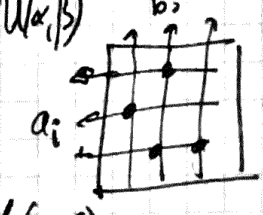
Non convex intractable pbm $\Rightarrow$ convex relaxation.

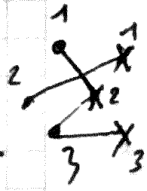
<u>Discrete Kantorovitch pbm</u> : $\alpha = \sum_i a_i \delta_{x_i} \longleftrightarrow \beta = \sum_j b_j \delta_{y_j}$



Coupling : $\left\{ P \in \mathbb{R}_+^{m \times m} : \sum_j P_{ij} = a_i,\ \sum_i P_{ij} = b_j \right\} = \mathcal{U}(\alpha,\beta)$

$W_p(\alpha,\beta)^p \triangleq \min \left\{ \sum P_{ij}\, d(x_i, y_j)^p : P \in \mathcal{U}(\alpha,\beta) \right\}$





$\rightsquigarrow$ Extends to arbitrary measures : replace $P \to \pi \in \mathcal{M}_+(\mathcal{X}^2)$

$\mathcal{U}(\alpha,\beta) = \left\{ \pi \in \mathcal{M}^+(\mathcal{X}^2) : \begin{matrix} \pi_1 = \alpha \\ \pi_2 = \beta \end{matrix} \right\}$. $\sum \rightsquigarrow \int_{\mathcal{X}}$

Examples : $\boxed{1D}$ $W_p(\sum_{i=1}^n \delta_{x_i}, \sum_{i=1}^n \delta_{y_i}) = \|\mathrm{sort}(x) - \mathrm{sort}(y)\|$, $W_2(N(m_1,\sigma_1), N(m_2,\sigma_2))^2 = \|m_1 - m_2\|^2 + \|\sigma_1 - \sigma_2\|^2$

Monge $\leftrightarrow$ Kanto : Discrete : $\sum_{i=1}^n \delta_{x_i} = \alpha$, $\sum_{i=1}^n \delta_{y_i} = \beta \Rightarrow \exists P$ ad° Permut° matrix (Birkhoff-Von Neumann)

Continuous (BRENIER) : π optimal supported on $x \mapsto T(x)$, $T(x) = \nabla\psi(x)$, ψ convex $\rightarrow$ MONGE AMPERE

Prop : W_p is a distance on $\mathcal{M}_+^1(X)$ which metrizes weak° conv

ex : $W_p(\delta_x, \delta_y) = d(x,y)$ $\longrightarrow$ not possible with MMD

Entropic Regularizat° : $\min_{P \in U(a,b)} \overbrace{\langle P, d^p \rangle}^{\sum_{ij} P_{ij} d(x_i,x_j)^p} + \varepsilon \underbrace{KL(P|a \otimes b)}_{\sum_{ij} P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right)}$

arbitrary measures

SCHRÖDINGER PBM $\longrightarrow W_{p\varepsilon}^p(\alpha,\beta) \triangleq \min_{\pi \in \mathcal{M}_+(X^2)} \left\{ \int d^p d\pi + \varepsilon KL(\pi|\alpha\otimes\beta) : \pi_1 = \alpha, \pi_2 = \beta \right\}$.

Prop : $\underset{\text{unique}}{P}$ optimal $\Leftrightarrow \begin{cases} \exists (u,v) \in \mathbb{R}^n \times \mathbb{R}^m & P_{ij} = K_{ij} u_i v_j a_i b_j \quad K = e^{-d^p/\varepsilon} \\ \text{and } P\mathbb{1} = a, & P^T \mathbb{1} = b \end{cases}$

Ca : one needs to find (u,v) st. $\begin{cases} P\mathbb{1} = (u \odot a) \odot (K \times (v \odot b)) = a \\ P^T \mathbb{1} = (v \odot b) \odot (K^T \times (u \odot a)) = b \end{cases}$

Sinkhorn Algorithm : $u \leftarrow \frac{1}{K(vb)} \Leftrightarrow v \leftarrow \frac{1}{K(ua)}$. Complexity per iter° : $O(n^2)$

THM/Sinkhorn (7) : converges!

log-domain : $(u,v) = (e^{f/\varepsilon}, e^{g/\varepsilon})$, (f,g) are "dual variables"

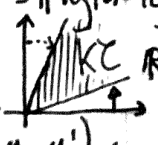
$$f_i \leftarrow -\varepsilon \underbrace{\log\left[\sum_j \exp\left(-\frac{d(x_i,x_j)^p - g_j}{\varepsilon}\right) b_j\right]}_{\text{LSE} \rightsquigarrow \text{stabilize}} = \mathrm{SoftMin}_\varepsilon^b \left[C(x_i,\cdot) - g(\cdot)\right]$$

$$g_j \leftarrow \mathrm{SoftMin}_\varepsilon^a \left[C(\cdot, y_j) - f(\cdot)\right] \qquad C = d^p$$

Approximation of the c-transforms : $f^c \triangleq \min_x c(x,\cdot) - f(x)$.

linear convergence rates : Variation norm : $\|f\|_V = \inf_{c \in \mathbb{R}} \|f + c\|_\infty$ $\rightsquigarrow$ defined up to a constant

Hilbert metric : $d_H(u,u') = \|\log(u) - \log(v)\|_\infty$

Contract° : $\exists \eta \in]0,1[$, $d_H(Ku, Ku') \leq \eta\, d_H(u,u')$  $\mathbb{R}_+^n = \mathbb{R}^n_+ \setminus C$

$\rightsquigarrow$ linear rate of Sinkh. $d_H(1/u, 1/u') = d_H(u,u')$.

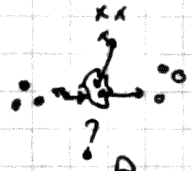
<u>Complexity</u> : $O(n^2)$ per iteration , $O(n)$ in some case (eg Gaussian cost)
Matrix/vector mult$^{\underline{n}}$ $\xrightarrow{\text{parallelis}^{\underline{n}}}$ Matrix/Matrix mult$^{\underline{n}}$

<u>BONUS</u> : Sinkhorn divergences: $W^p_{p\varepsilon}(\alpha\|\beta) - \frac{1}{2}W_\varepsilon(\alpha,\alpha) - \frac{1}{2}W_\varepsilon(\beta,\beta) \geq 0 \xrightarrow{\varepsilon\to+\infty} \|\alpha-\beta\|^2_{-dl^p}$!!!

<u>BONUS</u> : Sample complexity OT <u>vs</u> MMD <u>vs</u> Sinkhorn

<u>Barycenters</u> : $(\alpha,\beta) \rightsquigarrow W^p_p(\alpha,\beta)$ convex (look at the dual pbm).

[Cuturi/Agueh 2011] $\min_\alpha \underbrace{\sum_{k=1}^K \lambda_k W^p_p(\alpha,\beta_k)}_{\text{convex } f^{\underline{n}}}$



If $\beta_k = \delta_{x_k}$ $\rightarrow$ usual def$^{\underline{n}}$ $\min_x \sum_k \lambda_k \|x_k - x\|^p$

If β_k discrete $\Rightarrow$ α^* is discrete $\qquad$ β_k Gaussian $\Rightarrow$ α^* Gaussian

Entropic regul$^{\underline{n}}$: $\min_\alpha \sum_k^K \lambda_k W^p_{p\varepsilon}(\alpha,\beta_k) = \min_{\alpha,(\pi_k)_k} \sum_k \lambda_k KL(\pi_k | K \otimes \xi_0)$ Ref. meas.

$\rightsquigarrow$ Sinkhorn-like iterations $\qquad (\pi_k)_1 = \alpha \quad (\pi_k)_2 = \beta_k$

<u>Gradient-flows</u> : <u>In $\mathbb{R}^d$</u> : $\min_{x\in\mathbb{R}^d} f(x) \rightsquigarrow \frac{dx_t}{dt} = -\underbrace{\nabla f}_{\text{with respect to } \langle\cdot,\cdot\rangle_{\mathbb{R}^d}}(x_t)$

<u>Explicit</u>: $x_{k+1} = x_k - \tau\nabla f(x_k) \qquad (t \simeq \tau k)$

<u>Implicit</u>: $x_{k+1} + \tau\nabla f(x_{k+1}) = x_k \rightsquigarrow x_{k+1} = \underbrace{(Id+\tau\nabla f)^{-1}}_{\text{Prox}_{\tau f}}(x_k)$

"Proximal point algo" $\qquad$ <u>ie</u> : $x_{k+1} = \underset{x}{\text{Argmin}}\ \frac{1}{2}|x-x_k|^2 + \tau f(x)$

Advantages : no need for f to be smooth, no constraint on f,
Disadvantage : need to solve a sub-problem often intractable
Fundamental remark : derivative free $\rightarrow$ works over a metric-space !

$$\alpha_{k+1} \triangleq \underset{\alpha}{\text{Argmin}}\ d(\alpha_k,\alpha)^p + \tau f(\alpha)$$

Take $d(x,y) = \|x-y\|^2$, $p=2$ $\rightarrow$ Implicit Euler.

Take $d(\alpha,\beta) = W_2(\alpha,\beta)$, $p=2$ $\rightarrow$ Jordan-Kinderlehrer-Otto Wass. Flow

Continuous flow : let $\tau\to 0$, $\tau k \to t$ then $\alpha_k \xrightarrow{*} \alpha(t)$ for some class of function f

"measure valued" derivative : $f(\alpha+\varepsilon\xi) = f(\alpha) + \varepsilon \int f'(\alpha)\, d\xi + o(\varepsilon)$ $\quad f'(\alpha) \in \mathcal{C}(\mathcal{X})$

Formally, $\alpha(t)$ solves : $\dfrac{\partial \alpha}{\partial t} = -\mathrm{div}\left[\alpha \cdot \nabla\left(f'(\alpha)\right)\right]$

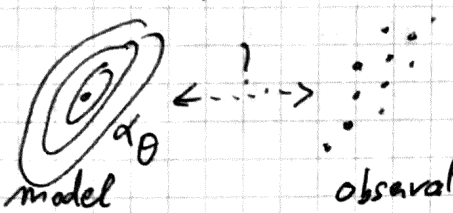
<u>Examples</u> : $f(\alpha) = \int u \, d\alpha \rightsquigarrow$ Advect$^{\underline{n}}$ $\quad \dfrac{\partial \alpha}{\partial t} = \mathrm{div}(\alpha \nabla u)$

$f(\alpha) = \int \log\left(\frac{d\alpha}{dx}\right) d\alpha \rightsquigarrow$ Heat $\quad \Delta \alpha = \dfrac{\partial \alpha}{\partial t}$

$f(\alpha) = \int h\left(\frac{d\alpha}{dx}\right) d\alpha \rightsquigarrow$ non-linear PDE, eg Porous Medium $\Delta \alpha$

$f(\alpha) = \iint k(x,y)\, d\alpha(x)\, d\alpha(y) \rightsquigarrow$ Pairwise inter$^{\underline{n}}$, Keller-Segel, etc

<u>Density-Fitting</u> :



$\beta = \frac{1}{n} \sum \delta_{x_i}$

Maximum-likelihood : $d\alpha_\theta = \underbrace{p_\theta \, dx}$ $\rightarrow$ fixed for all θ (only its support matters)

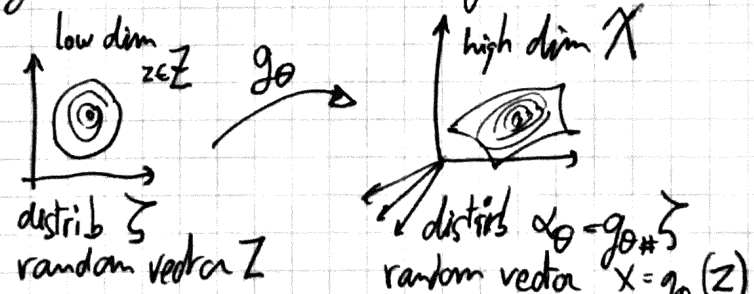
$\min_\theta \frac{1}{n} \sum_i \log(p_\theta(x_i)) \xrightarrow[n \to +\infty]{} KL(\beta_\infty | \alpha)$ $\quad$ <u>Ex</u> : $\begin{cases} \text{Gaussian estimat}^{\underline{n}} \\ \text{Mixture (EM)} \end{cases}$

<u>Pbm</u> : if support(α_θ) "moves" + p_θ might be intractable...

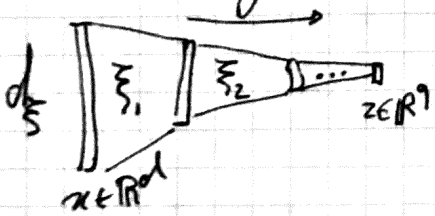
Minimum divergence estimator : $\min_\theta D(\alpha_\theta | \beta)$

<u>Generative model</u> :

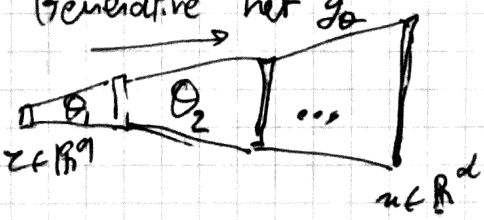
$\rightsquigarrow$ Singular distribut$^{\underline{n}}$



Discriminatory deep-net



Generative net g_θ



<u>φ-divergence</u> : not ok, not weak* continuous
OT / MMD : not tractable, need to know a metric
$\Big\} \rightarrow$ use duality!

Other approach : $\left.\begin{matrix} \text{OT} \rightarrow \text{Sinkhorn} \\ \text{MMD} \end{matrix}\right\}$ + metric learning (cf Aude Genevay)

$$\alpha \longrightarrow D(\alpha,\beta) \text{ convex}, \qquad D(\alpha,\beta) = \sup_{f:X\to\mathbb{R}} \underbrace{\langle f,\alpha\rangle}_{\int f(x)\,d\alpha(x)} - \underbrace{D^*(f,\beta)}_{\text{Legendre transfo w.r.t. } \alpha}$$

<u>Example</u>: φ-div: $D^*_\varphi(f,\beta) = \int \varphi^*(f(x))\,d\beta(x)$

TV: $\varphi(x) = |1-x| \longrightarrow \varphi^*(t) = \iota_{[-\frac{1}{2},\frac{1}{2}]}(t)$

KL: $\varphi(x) = x\log x \longrightarrow \varphi^*(t) = \exp(t-1)$

JS: $\varphi(x) = x\log x - (x+1)\log(x+1) \longrightarrow \varphi^*(t) = -\log(1-e^t) + \iota_{\mathbb{R}^-}(t)$

<u>Example</u>: MMD: $D(\alpha,\beta) = \frac{1}{2}\|\alpha-\beta\|_K^2$, $\quad D^*(f,\beta) = \int f\,d\beta - \frac{1}{2}\|f\|^2_{RKHS}$

<u>Example</u>: W_1: $D(\alpha,\beta) = W_1(\alpha,\beta)$ $\quad D^*(f,\beta) = \int f\,d\beta - \iota_{Lips}(f)$

$\hookrightarrow \|\nabla f\|_\infty \leq 1$

<u>Min-Max fit</u>: $\displaystyle\min_\theta \max_f \int_X f\,d\alpha_\theta - D^*(f,\beta)$

$$\alpha_\theta = \int_Z f(g_\theta(z))\,d\zeta(z)$$

<u>Generative Adversarial Networks</u>: restrict $f = d_\xi$ to be discr. net

[Goodfellow et al]

$$\min_\theta \max_\xi \int_Z d_\xi(g_\theta(z))\,d\zeta(z) - D^*(d_\xi,\beta)$$

for φ-div. $\longrightarrow = \frac{1}{n}\sum_i \varphi^*(d_\xi(x_i))$

<u>Concl°</u>: $\rightarrow$ hard to solve but impressive result

$\rightarrow$ Restricting $f = d_\xi$: not anymore a true divergence

$\rightarrow$ Unclear which D is best (is OT really useful?)