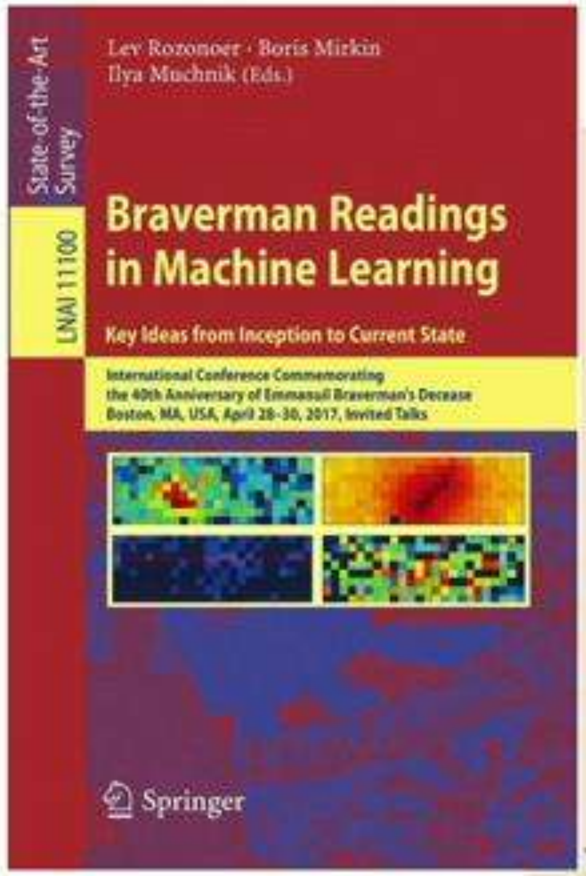
CONVEXITY « À LA CARTE »

Léon Bottou



Geometrical Insights for Implicit Generative Modeling

Leon Bottou [a,b], Martin Arjovsky [b,a], David Lopez-Paz [a], Maxime Oquab [a,c]

Abstract

Learning algorithms for implicit generative models can optimize a variety of criteria that measure how the data distribution differs from the implicit model distribution, including the Wasserstein distance, the Energy distance, and the Maximum Mean Discrepancy criterion. A careful look at the geometries induced by these distances on the space of probability measures reveals interesting differences. In particular, we can establish surprising approximate global convergence guarantees for the 1-Wasserstein distance, even when the parametric generator has a nonconvex parametrization.

arXiv:1712.07822
sections 6.1 and 6.3

Summary

1. Convex optimization « à la carte »
2. Approximation properties, global minimization, parametrization bias.
3. The case of implicit generative models.

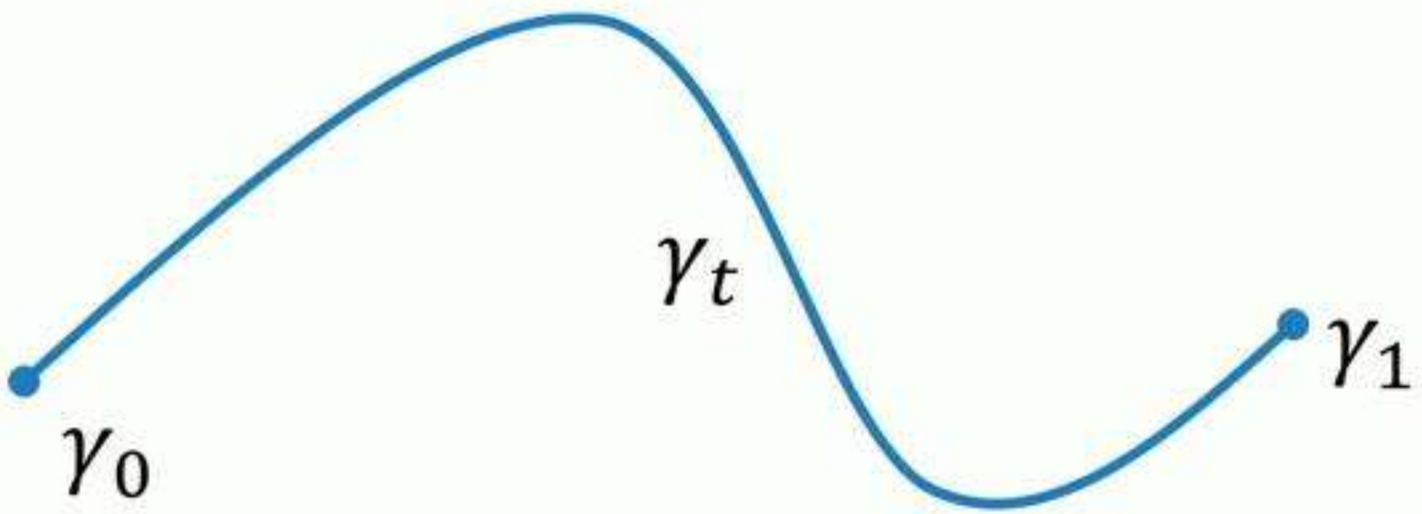
1- Convexity « à la carte »

Background

Curves

Let $\mathfrak{X}$ be a Polish metric space.

A continuous mapping $\gamma : t \in [0,1] \subset \mathbb{R} \mapsto \gamma_t \in \mathfrak{X}$ defines a curve in $\mathfrak{X}$ that connects γ_0 to γ_1.

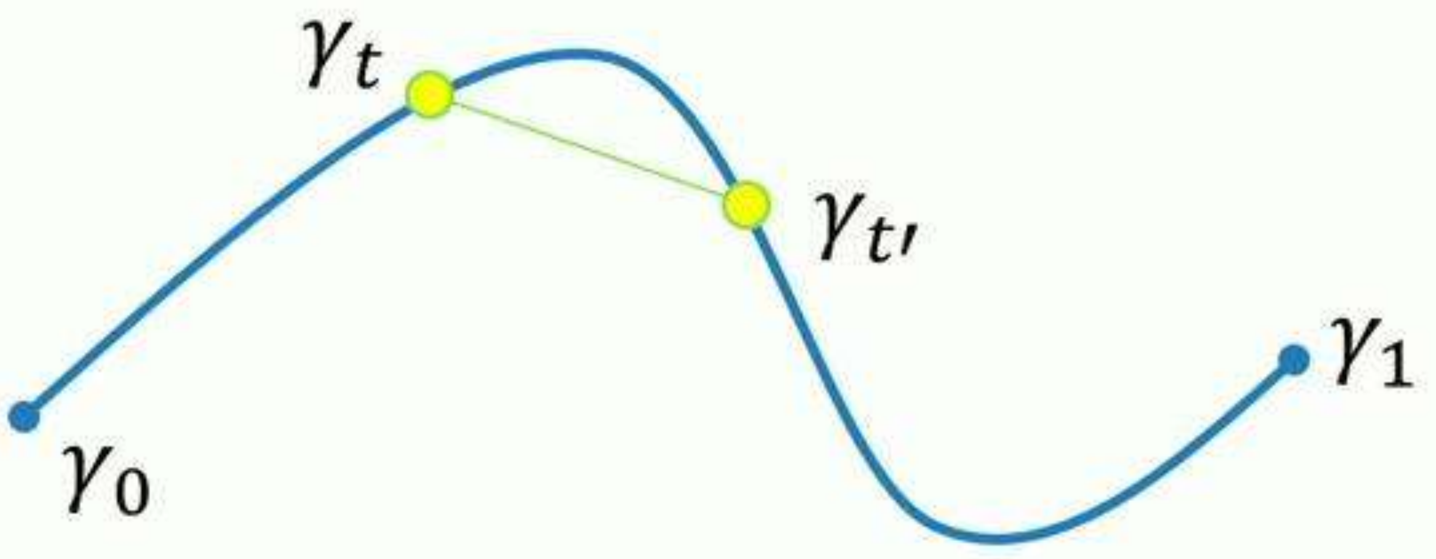


Background

Bounded speed curve

Such a curve has bounded speed if there is $K > 0$ such that

$$\forall \ 0 \leq t \leq t' \leq 1 \quad d(\gamma_t, \gamma_{t'}) \leq K\,(t' - t)$$



Definitions

Let $\mathcal{C}$ denote a family of curves in $\mathfrak{X}$

- **A subset $\boldsymbol{\mathcal{F}} \subset \boldsymbol{\mathfrak{X}}$ is convex with respect to $\boldsymbol{\mathcal{C}}$**
  when, for every pair $x, y \in \mathcal{F}$, there is a curve $\gamma \in \mathcal{C}$ that is entirely contained in $\mathcal{F}$, that is,

$$\forall t \in [0,1] \quad \gamma_t \in \mathcal{F}$$

- **A real function $\boldsymbol{f} : \boldsymbol{\mathfrak{X}} \to \mathbb{R}$ is convex with respect to $\boldsymbol{\mathcal{C}}$**
  when for every curve $\gamma \in \mathcal{C}$, the restriction of f to the curve is convex, that is,

$$\forall\, t, a, b \in [0,1] \quad f\big((1-t)\gamma_a + t\,\gamma_b\big) \leq \ (1-t)\, f(\gamma_a) + t\, f(\gamma_b)$$

Definitions

Let $\mathcal{C}$ denote a family of curves in $\mathfrak{X}$

We recover normal convexity when $\mathfrak{X}$ is a Euclidean space and $\mathcal{C}$ contains all line segments.

- **A subset $\boldsymbol{\mathcal{F}} \subset \boldsymbol{\mathfrak{X}}$ is convex with respect to $\boldsymbol{\mathcal{C}}$**
  when, for every pair $x, y \in \mathcal{F}$, there is a curve $\gamma \in \mathcal{C}$ that is entirely contained in $\mathcal{F}$, that is,

$$\forall t \in [0,1] \quad \gamma_t \in \mathcal{F}$$

- **A real function $\boldsymbol{f} : \boldsymbol{\mathfrak{X}} \to \mathbb{R}$ is convex with respect to $\boldsymbol{\mathcal{C}}$**
  when for every curve $\gamma \in \mathcal{C}$, the restriction of f to the curve is convex, that is,

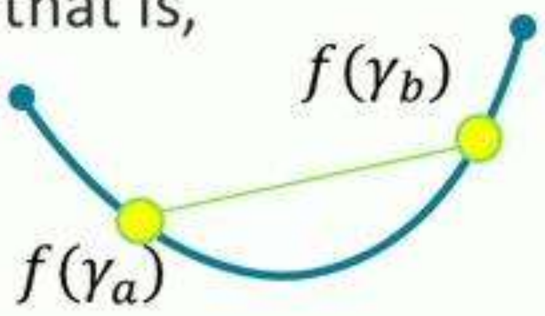
$$\forall\, t, a, b \in [0,1] \quad f\big((1-t)\gamma_a + t\,\gamma_b\big) \leq \ (1-t)\, f(\gamma_a) + t\, f(\gamma_b)$$

Definitions

Let $\mathcal{C}$ denote a family of curves in $\mathfrak{X}$

- **A subset $\mathcal{F} \subset \mathfrak{X}$ is convex with respect to $\mathcal{C}$**
  when, for every pair $x, y \in \mathcal{F}$, there is a curve $\gamma \in \mathcal{C}$ that is entirely contained in $\mathcal{F}$, that is,
  $$\forall t \in [0,1] \quad \gamma_t \in \mathcal{F}$$

- **A real function $f : \mathfrak{X} \to \mathbb{R}$ is convex with respect to $\mathcal{C}$**
  when for every curve $\gamma \in \mathcal{C}$, the restriction of f to the curve is convex, that is,
  $$\forall\, t, a, b \in [0,1] \;\; f\left(\gamma_{(1-t)a+tb}\right) \leq \; (1-t)\, f(\gamma_a) + t\, f(\gamma_b)$$



Definitions

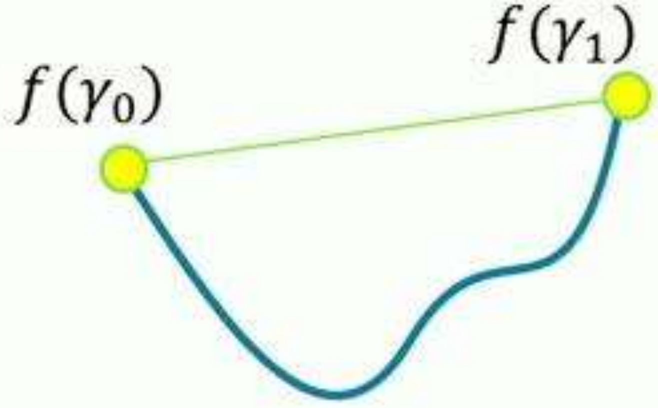
Let $\mathcal{C}$ denote a family of curves in $\mathfrak{X}$

- **A subset $\mathcal{F} \subset \mathfrak{X}$ is convex with respect to $\mathcal{C}$**
  when, for every pair $x, y \in \mathcal{F}$, there is a curve $\gamma \in \mathcal{C}$ that is entirely contained in $\mathcal{F}$, that is,

$$\forall t \in [0,1] \quad \gamma_t \in \mathcal{F}$$

- **A real function $f : \mathfrak{X} \to \mathbb{R}$ is endpoints-convex with respect to $\mathcal{C}$**
  when, for every curve $\gamma \in \mathcal{C}$

$$\forall\, t \in [0,1] \quad f(\gamma_t) \leq (1-t)\, f(\gamma_0) + t\, f(\gamma_1)$$



Convex optimization « à la carte »

Theorem

Let $\mathcal{F} \subset \mathfrak{X}$ be convex with respect to $\mathcal{C}$.

Let the cost function $f : \mathfrak{X} \to \mathbb{R}$ be endpoints-convex with respect to $\mathcal{C}$.

Then:

- $\forall\, M \geq \min_{\mathcal{F}} f,$ the level sets $L(f, \mathcal{F}, \mathrm{M}) = \{x \in \mathcal{F} \;\; s.t.\; f(x) \leq M\}$ are connected.
- If $\mathcal{C}$ only contains bounded speed curves, all local minima of f in $\mathcal{F}$ are global.

Proof (1)

Let $x, y \in L(f, \mathcal{F}, \mathrm{M})$.

- Since $\mathcal{F}$ is convex w.r.t. $\mathcal{C}$, there is a curve $\gamma \in \mathcal{C}$ connecting x to y such that
$$\forall t \in [0,1] \;\; \gamma_t \in \mathcal{F}$$
- Since f is endpoints-convex w.r.t. $\mathcal{C}$, for all $t \in [0,1]$,
$$f(\gamma_t) \leq (1-t)f(x) + t\, f(y) \leq M \quad \Rightarrow \quad \gamma_t \in L(f, \mathcal{F}, \mathrm{M})$$

Therefore $L(f, \mathcal{F}, \mathrm{M})$ is path-connected.

Proof (2)

- A point $x \in \mathcal{F}$ is a local minimum of f in $\mathcal{F}$ iff there is $\epsilon > 0$ such that, for all $x' \in \mathcal{F}$, $d(x,x') < \epsilon \implies f(x') \geq f(x)$.
- Reasoning by contradiction, assume there is $y \in \mathcal{F}$ such that $f(y) < f(x)$.
- Let $\gamma \in \mathcal{C}$ be a bounded speed curve contained $\mathcal{F}$ and connecting x to y :
$$\forall\ 0 \leq t \leq t' \leq 1 \quad d(\gamma_t, \gamma_{t'}) \ \leq K\ (t' - t)$$
- Therefore $f\left(\gamma_{\epsilon/2K}\right) \geq f(\gamma_0) = f(x)$
- But endpoints convexity means $f\left(\gamma_{\epsilon/2K}\right) \leq \left(1 - \frac{\epsilon}{2K}\right) f(x) + \frac{\epsilon}{2K} f(y) < f(x)$!!!

Simple machine learning example (1)

- Let $\mathfrak{X}$ be the continuous functions from $\Omega \subset \mathbb{R}^{d_{in}}$ to $\mathbb{R}^{d_{out}}$
- Let $\mathcal{F} \subset \mathfrak{X}$ be a family of functions $F_\theta : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ parametrized by θ.
- Let ℓ: $\mathbb{R}^{d_{out}} \times \mathbb{R}^{d_{out}}$ be a loss function, *convex in its first argument*.
- Let the training examples $(x_1, y_1) \ldots (x_n, y_n) \in \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}$
- Define the empirical cost function

$$f : F \in \mathfrak{X} \mapsto f(F) = \frac{1}{n} \sum_i \ell(F(x_i), y_i)$$

I did not write "parametric"

Simple machine learning example (2)

- Let the curves in $\mathcal{C}$ represent <u>mixtures</u> of any two functions of $\mathfrak{X}$

$$\forall\, F, G \in \mathfrak{X}, \quad \forall t \in [0,1], \;\; \gamma_t^{FG} = (1-t)F + t\,G$$

Line segments in $\mathfrak{X}$!

- Cost function f is trivially convex w.r.t. $\mathcal{C}$

- If $\mathcal{F}$ is convex w.r.t. $\mathcal{C}$, the theorem applies

  - *Linear models: YES*
  - *Kernel models : YES*
  - *Neural networks : ALMOST?*

Neural networks (1)

Why ALMOST?

- If an overparametrized neural network can approximate anything (e.g. Cybenko89, Hornik89) then there should be weights θ_t that make F_{θ_t} arbitrarily close to $\gamma_t^{FG} = (1-t)F + t\,G$.

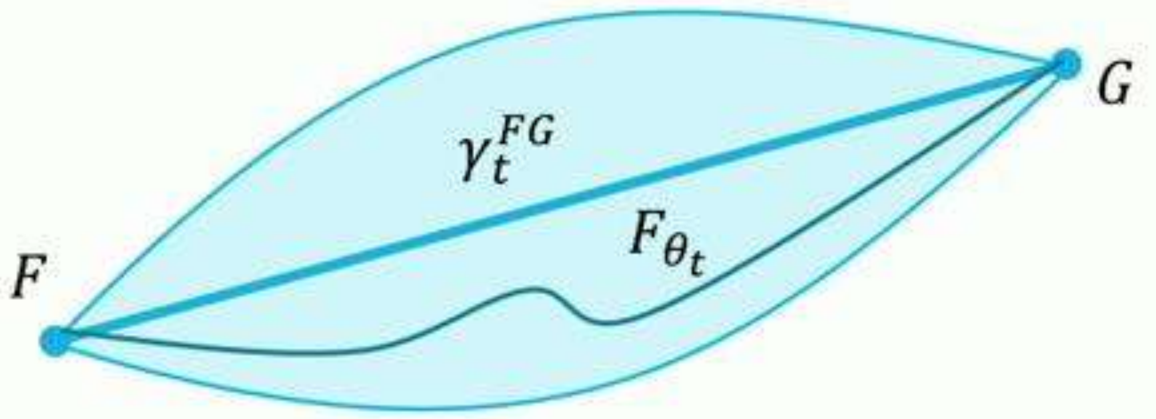
This is not sufficient!

- A$\Longrightarrow$B does not generally mean that AlmostA $\Longrightarrow$AlmostB.
- This is where curves can help.

Neural networks (2)

- For the sake of the argument, assume that we can find θ_t such that

$$d\left(\gamma_t^{FG}, F_{\theta_t}\right) \leq \; R\, t\,(1-t)$$



Proving this is cumbersome
—I won't even try—
but the point is that R
gets smaller when the net gets
bigger and approximates better

- Let $\mathcal{C}$ contain all curves contained in such cigar shaped regions
- By construction $\mathcal{F}$ is convex w.r.t. $\mathcal{C}$.
- But is the cost function f endpoints-convex w.r.t. $\mathcal{C}$?

Neural network (3)

- With a Lipschitz assumption on the loss ℓ we can have something like

$$\begin{aligned} f\left(F_{\theta_t}\right) &\leq f(\gamma_t^{FG}) + \lambda\, t\,(1-t) \\ &\leq (1-t) f(F) + t\, f(G) + \lambda\, t\,(1-t) \end{aligned}$$

This holds because f is convex w.r.t. the mixture curves γ_t^{FG}.

- In fact, if the loss ℓ were μ-strongly convex
  we could even write $f\left(F_{\theta_t}\right) \leq (1-t) f(F) + t\, f(G) + (\lambda - \mu) t\,(1-t)$
  and apply the convexity a-la-carte theorem when $\mu \geq \lambda$!

- What about the general case?

Almost-convex optimization «à la carte»

Let $\mathcal{F} \subset \mathfrak{X}$ be convex with respect to $\mathcal{C}$.

For each $\gamma \in \mathcal{C}$, let the cost function $f : \mathfrak{X} \to \mathbb{R}$ satisfy

$$f(\gamma_t) \leq (1-t)\, f(\gamma_0) + t\, f(\gamma_1) + \lambda\, t(1-t)$$

Then:

- $\forall\, M \geq \left(\min_{\mathcal{F}} f\right) + \lambda,$ the level sets $L(f, \mathcal{F}, \mathrm{M})$ are connected.

Basically, any local minimum is
at most γ above the global minimum.

Neural network (3)

- With a Lipschitz assumption on the loss ℓ we can have something like

$$\begin{aligned} f\left(F_{\theta_t}\right) &\leq f(\gamma_t^{FG}) + \lambda\, t\,(1-t) \\ &\leq (1-t) f(F) + t\, f(G) + \lambda\, t\,(1-t) \end{aligned}$$

This holds because f is convex w.r.t. the mixture curves γ_t^{FG}.

- In fact, if the loss ℓ were μ-strongly convex
we could even write $f\left(F_{\theta_t}\right) \leq (1-t) f(F) + t\, f(G) + (\lambda - \mu) t\,(1-t)$
and apply the convexity a-la-carte theorem when $\mu \geq \lambda$!

- What about the general case?

Almost-convex optimization «à la carte»

Let $\mathcal{F} \subset \mathfrak{X}$ be convex with respect to $\mathcal{C}$.

For each $\gamma \in \mathcal{C}$, let the cost function $f : \mathfrak{X} \rightarrow \mathbb{R}$ satisfy

$$f(\gamma_t) \leq (1-t)\, f(\gamma_0) + t\, f(\gamma_1) + \lambda\, t(1-t)$$

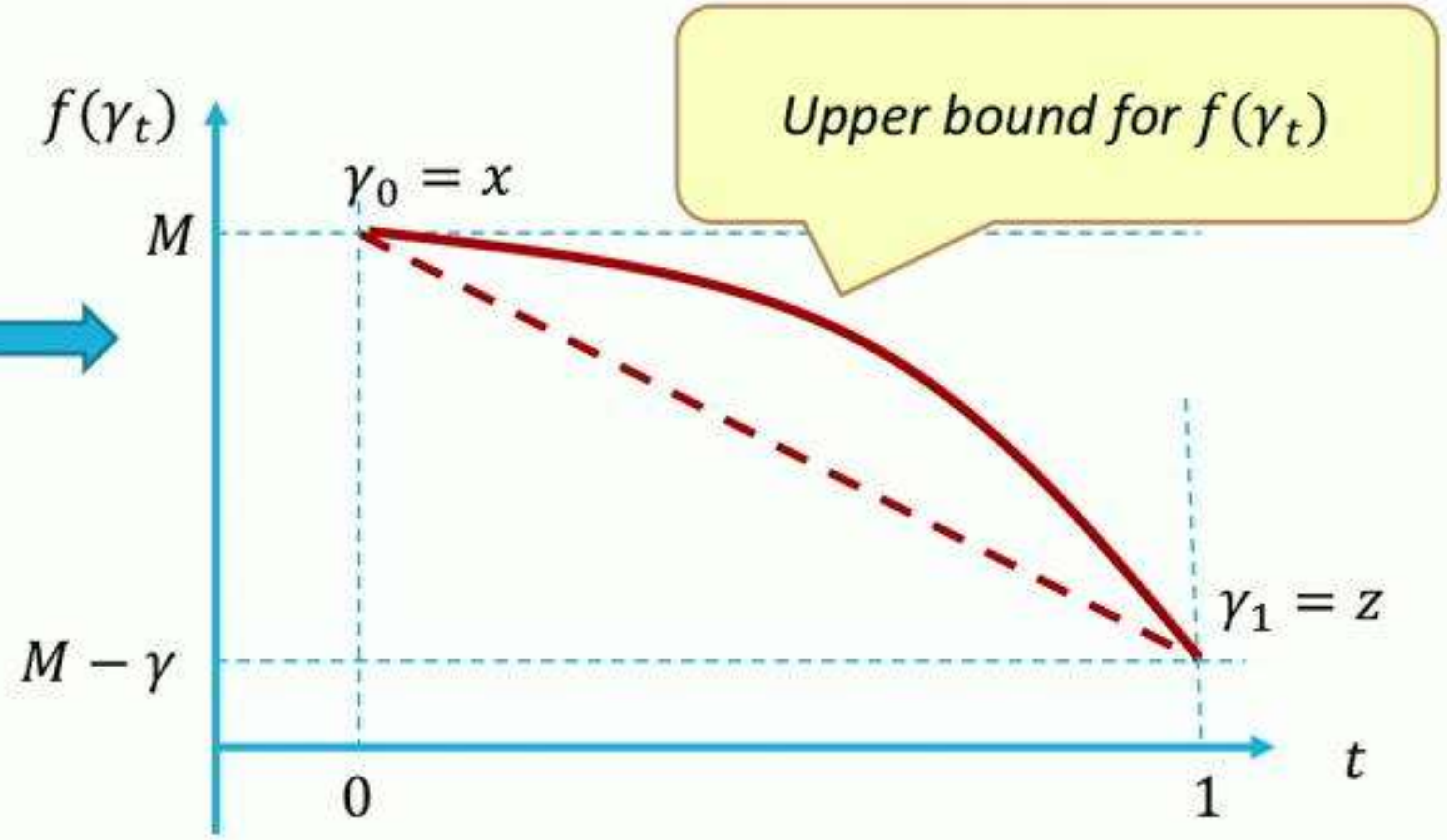
Then:

- $\forall\, M \geq \left(\min_{\mathcal{F}} f \right) + \lambda,$ the level sets $L(f, \mathcal{F}, \mathrm{M})$ are connected.

Basically, any local minimum is at most γ above the global minimum.

Proof

- Let $x, y \in L(f, \mathcal{F}, \mathrm{M})$ with $M \geq \left(\min_{\mathcal{F}} f\right) + \gamma$. We have $f(x) \leq M$ and $f(y) \leq M$.
- Pick $z \in L(f, \mathcal{F}, M)$ such that $f(z) \leq \mathrm{M} - \gamma$.
- Find a curve $\gamma \in \mathcal{C}$ connecting x to z such that $\forall t \in [0,1],\ \gamma_t \in \mathcal{F}$.
- Observe that $\gamma_t \in L(f, \mathcal{F}, M)$
- Similarly curve $\gamma' \in \mathcal{C}$ connecting z to y
- Concatenate curves γ and γ' to form a path that connects x to y without leaving $L(f, \mathcal{F}, M)$.



Almost-convex optimization «à la carte»

Let $\mathcal{F} \subset \mathfrak{X}$ be convex with respect to $\mathcal{C}$.

For each $\gamma \in \mathcal{C}$, let the cost function $f : \mathfrak{X} \to \mathbb{R}$ satisfy

$$f(\gamma_t) \leq (1-t)\, f(\gamma_0) + t\, f(\gamma_1) + \lambda\, t(1-t)$$

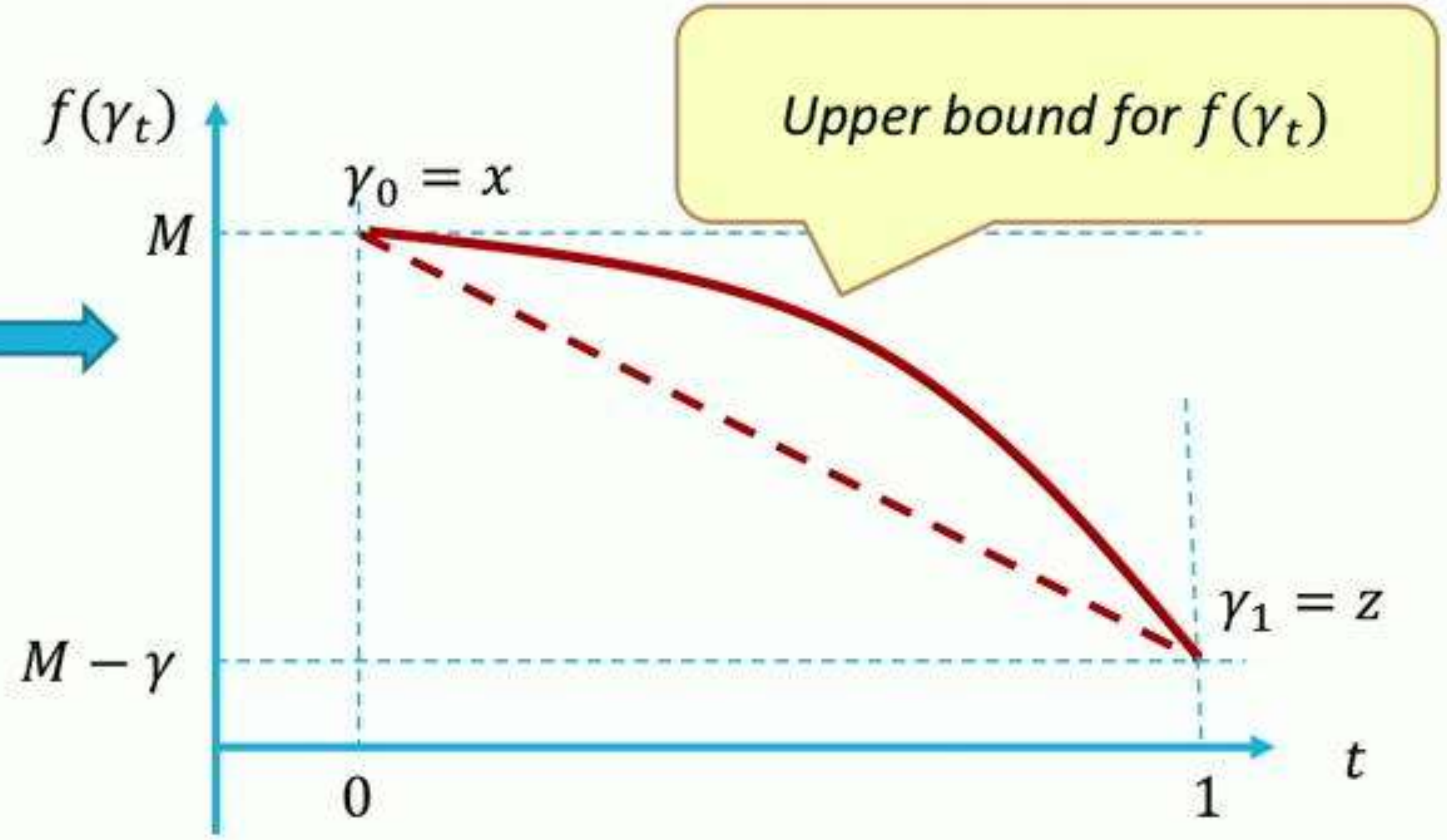
Then:

- $\forall\, M \geq \left(\min_{\mathcal{F}} f \right) + \lambda,$ the level sets $L(f, \mathcal{F}, \mathrm{M})$ are connected.

Basically, any local minimum is
at most γ above the global minimum.

Proof

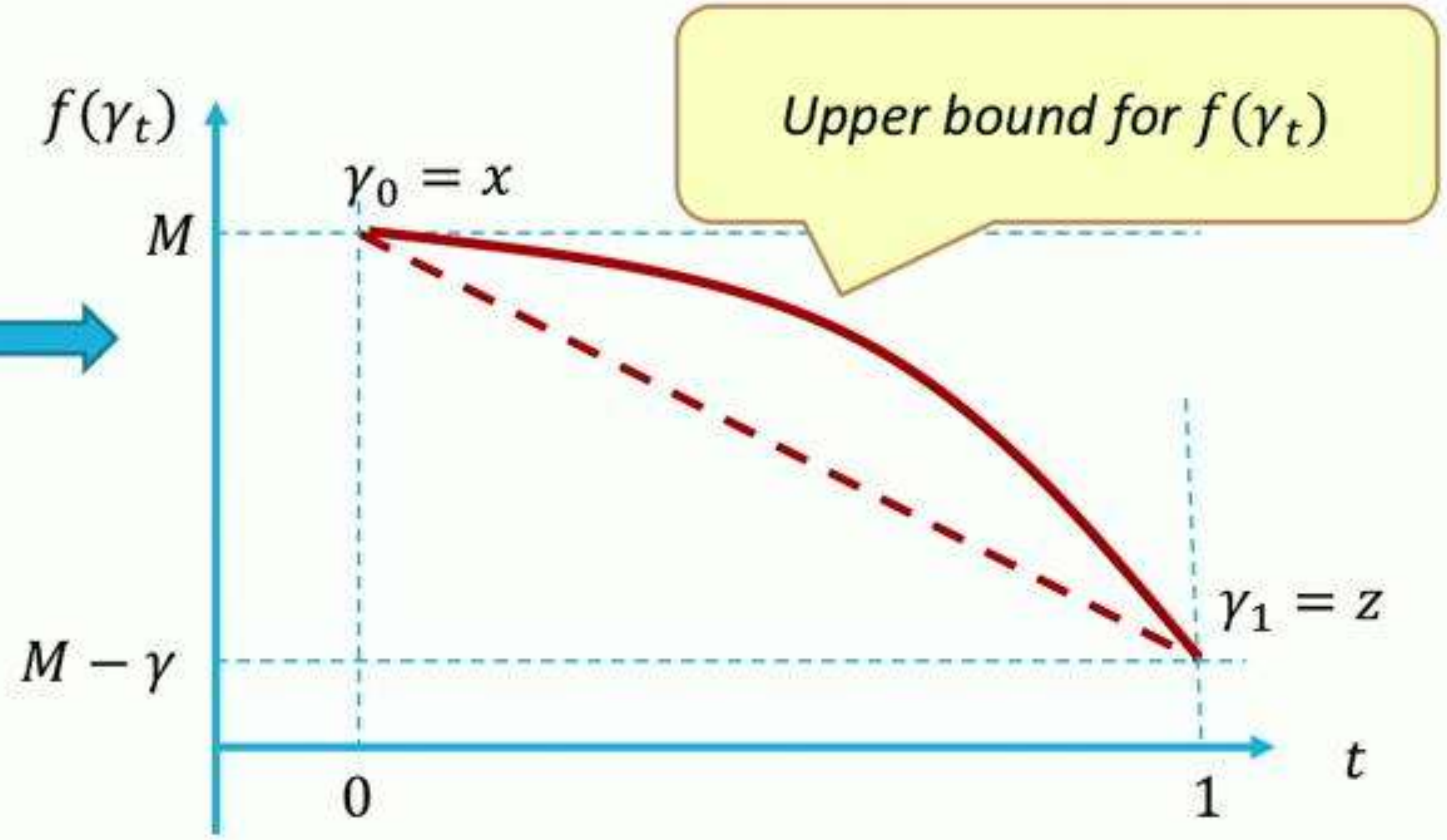
- Let $x, y \in L(f, \mathcal{F}, \mathrm{M})$ with $M \geq \left(\min_{\mathcal{F}} f\right) + \gamma$. We have $f(x) \leq M$ and $f(y) \leq M$.
- Pick $z \in L(f, \mathcal{F}, M)$ such that $f(z) \leq \mathrm{M} - \gamma$.
- Find a curve $\gamma \in \mathcal{C}$ connecting x to z such that $\forall t \in [0,1],\ \gamma_t \in \mathcal{F}$.
- Observe that $\gamma_t \in L(f, \mathcal{F}, M)$
- Similarly curve $\gamma' \in \mathcal{C}$ connecting z to y
- Concatenate curves γ and γ' to form a path that connects x to y without leaving $L(f, \mathcal{F}, M)$.



2- Approximation properties, global minimization, and parametrization bias

Proof

- Let $x, y \in L(f, \mathcal{F}, \mathrm{M})$ with $M \geq \left(\min_{\mathcal{F}} f\right) + \gamma$. We have $f(x) \leq M$ and $f(y) \leq M$.
- Pick $z \in L(f, \mathcal{F}, M)$ such that $f(z) \leq \mathrm{M} - \gamma$.
- Find a curve $\gamma \in \mathcal{C}$ connecting x to z such that $\forall t \in [0,1],\ \gamma_t \in \mathcal{F}$.
- Observe that $\gamma_t \in L(f, \mathcal{F}, M)$
- Similarly curve $\gamma' \in \mathcal{C}$ connecting z to y
- Concatenate curves γ and γ' to form a path that connects x to y without leaving $L(f, \mathcal{F}, M)$.



Discussion (general)

- These results are independent from the parametrization of $\mathcal{F}$.
  They depend on whether any two points in $\mathcal{F}$ can be connected by a suitable curve that (a) either remains in $\mathcal{F}$, or (b) can be well approximated by elements of $\mathcal{F}$.
- In θ space, the level sets can be very nonconvex, and yet connected.
- However, because learning algorithms operate in θ space,
  the parametrization changes the implicit biases that affect
  - which global minimum is returned in overparametrized models, or,
  - which solution is returned after early stopping.

Discussion (mixture curves)

- When the family of functions $\mathcal{F}$ has strong enough approximations properties to closely represent linear mixtures of any two of its functions, any reasonable learning algorithm will eventually find a near-global minimum. *(must cite many recent work here)*
- We can say this because the learning algorithm has the possibility to overcome the parametrization bias and essentially function as it would for a kernel model.
- But the learning algorithm might find a good enough solution without exercising this possibility. This can improve generalization performance when the parametrization bias is sensible for the problem at hand…
- This is doomed to be problem-specific ☹

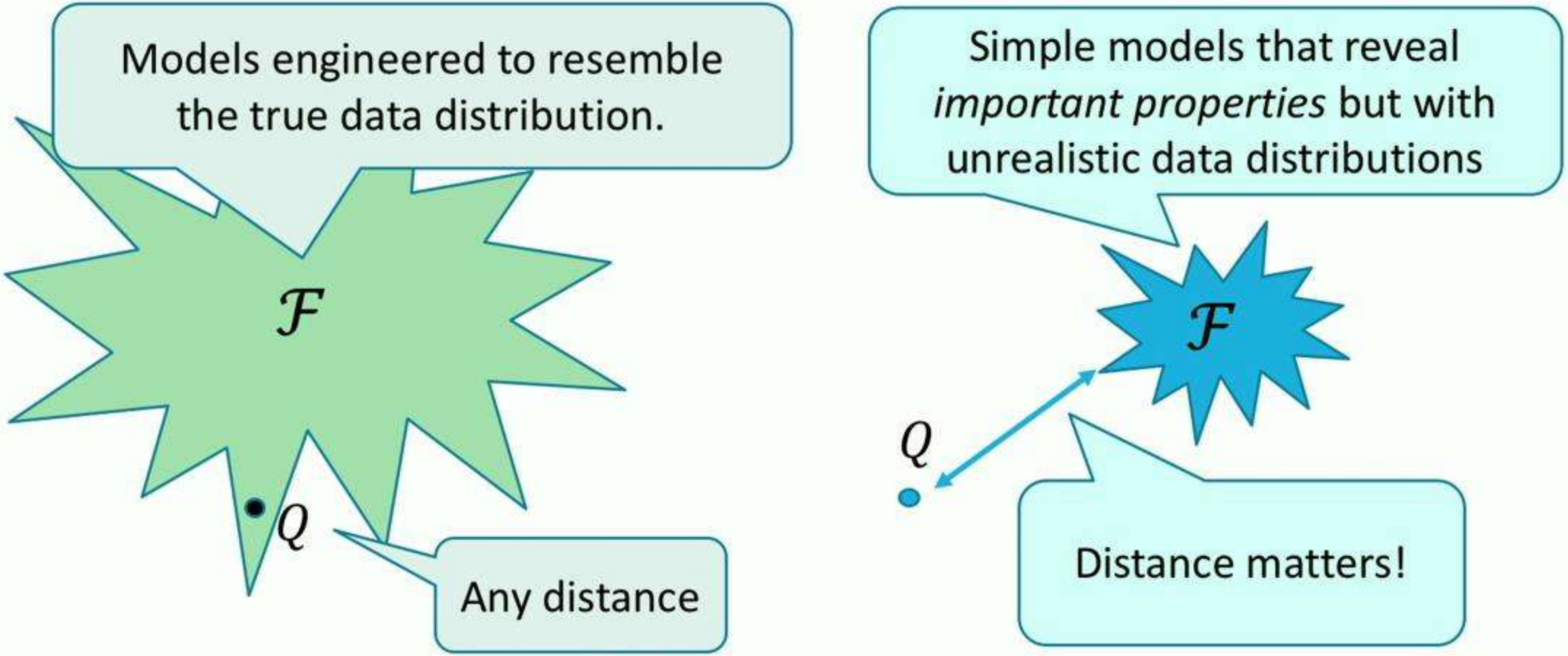
Discussion (mixture curves)

- When the family of functions $\mathcal{F}$ has strong enough approximations properties to closely represent lin… learning algorithm will eventually fi…
- We can say … vercome the parametriza… nodel.
- But the learn… exercising this possibility. This can improve generalization performance when the parametrization bias is sensible for the problem at hand…
- This is doomed to be problem-specific ☹

What about using other kinds of curves?

3- The case of implicit generative models.

Two learning approaches



Griefs about Maximum Likelihood

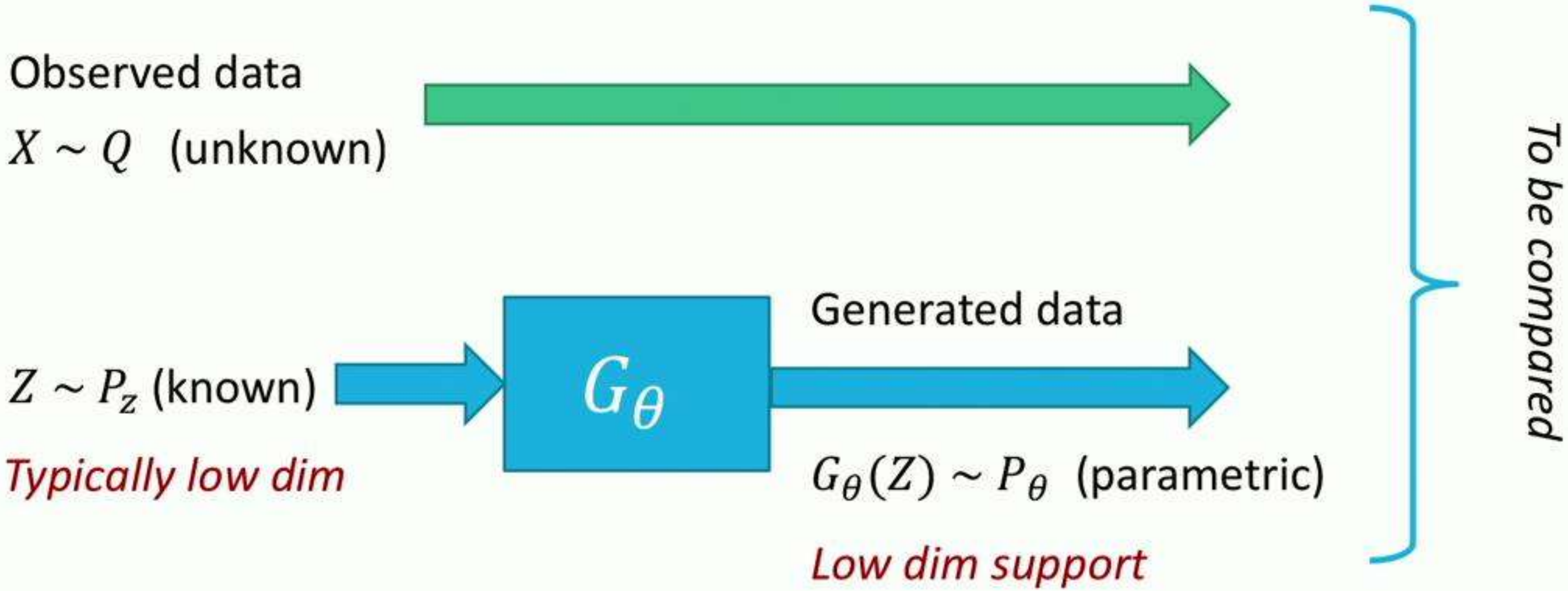
What is a simple model?

- A model that only involves a couple observed or latent variables.
- A degenerate distribution supported by a low-dimensional manifold.
- It does not have a density --> no density estimation…

Ugly workaround

- Augment the simple model with a noise model, ... and …
  tweak the noise model to coerce MLE into producing the desired outcome.

Implicit modeling



Implicit modeling

Let z be a random variable with known distribution μ_z defined on a suitable probability space $\mathcal{Z}$ and let G_θ be a measurable function, called the *generator*, parametrized by $\theta \in \mathbb{R}^d$,

$$G_\theta : \ z \in \mathcal{Z} \ \mapsto \ G_\theta(z) \in \mathcal{X} \ .$$

The random variable $G_\theta(Z) \in \mathcal{X}$ follows the *push-forward* distribution[7]

$$G_\theta(z)_{\#}\mu_Z(z) : \ \ A \in \mathfrak{U} \ \mapsto \mu_z(G_\theta^{-1}(A)) \ .$$

By varying the parameter θ of the generator G_θ, we can change this push-forward distribution and hopefully make it close to the data distribution Q according to the criterion of interest.

Implicit modeling

Let z be a random variable with known distributi… …ned on a suitable probability space $\mathcal{Z}$ and let G_θ be a meas… …d the *generator*, parametrized by $\theta \in \mathbb{R}^d$,

Good for degenerate distributions

The randon… …*push-forward* distribution[7]

$$\ldots{}_{\#}\mu_Z(z) : \quad A \in \mathfrak{U} \ \mapsto \mu_z(G_\theta^{-1}(A)) \ .$$

By varying the parameter θ of the generator G_θ, we can change this push-forward distribution and hopefully make it close to the data distribution Q according to the criterion of interest.

Comparing distributions

- The *Total Variation* (TV) distance

$$\delta(Q,P) = \sup_{A \in \mathfrak{A}} |Q(A) - P(A)|$$

- The *Kullback-Leibler* (KL) divergence

$$KL(Q\|P) = \int \log\left(\frac{q(x)}{p(x)}\right) q(x) d\mu(x)$$

requires densities, asymmetric, possibly infinite

VAE

- The *Jensen-Shannon* (JS) divergence

$$JS(Q,P) = \tfrac{1}{2}KL(Q\|M) + \tfrac{1}{2}KL(P\|M) \quad \text{with} \quad M = \tfrac{1}{2}(P+Q)$$

symmetric, does not require densities, $0 \leq JS \leq \log(2)$

GAN_0

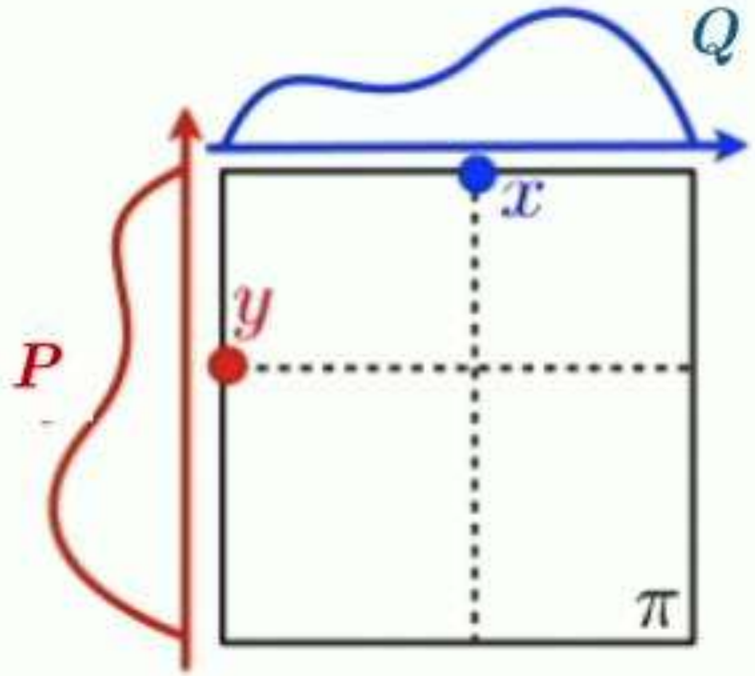
Comparing distributions

- The *Earth-Mover* (EM) distance or Wasserstein-1

$$W_1(Q,P) = \inf_{\pi \in \Gamma(Q,P)} \mathbb{E}_{(x,y)\sim\pi}[d(x,y)]$$
$$= \sup_{f \in \mathsf{Lip1}} \mathbb{E}_{x\sim Q}[f(x)] - \mathbb{E}_{y\sim P}[f(y)]$$

Always defined,
Involves metric on underlying space
Kantorovich duality.

WGAN



Comparing distributions

- The *Energy* (ED) distance $\equiv$ *Maximum Mean Discrepancy* (MMD)

$$\begin{aligned}\mathcal{E}(Q,P) &= 2\mathbb{E}_{x\sim Q\; y\sim P}[d(x,y)] - \mathbb{E}_{x,x'\sim Q}[d(x,x')] - \mathbb{E}_{y,y'\sim P}[d(y,y')] \\ &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_{x\sim Q}[f(x)] - \mathbb{E}_{y\sim P}[f(y)]\end{aligned}$$

Always defined when P and Q have first moments,
Needs a suitable metric/kernel on underlying space.

DiscoGANs

Readings: (Szekely 2002-; Gretton et al. 2012-) xGANS: (Bouchacourt et al., 2016 ; Bellemarre et al., 2017)

Mixtures

$$\forall t \in [0,1] \quad \mathrm{P_t} = (1-\mathrm{t})\ \mathrm{P_0} + \mathrm{t}\,\mathrm{P_1}$$

Let the set of distributions $\mathcal{F} = \{\, G_\theta \# \mu_z :\ \theta \in \mathbb{R}^d \,\}$ be mixture-convex.

→ For all $P_0, P_1 \in \mathcal{F}$ there is $t \mapsto \theta_t \in \mathbb{R}^d$ such that $P_t = G_{\theta_t} \# \mu_z$

Problem

If P_0 and P_1 have disjoint supports with nonzero margin, then either $t \mapsto \theta_t$ is discontinuous or $\theta \mapsto G_\theta$ is discontinuous.

Comparing distributions

- The *Energy* (ED) distance $\equiv$ *Maximum Mean Discrepancy* (MMD)

$$\begin{aligned}\mathcal{E}(Q,P) &= 2\mathbb{E}_{x\sim Q\ y\sim P}[d(x,y)] - \mathbb{E}_{x,x'\sim Q}[d(x,x')] - \mathbb{E}_{y,y'\sim P}[d(y,y')] \\ &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_{x\sim Q}[f(x)] - \mathbb{E}_{y\sim P}[f(y)]\end{aligned}$$

Always defined when P and Q have first moments,
Needs a suitable metric/kernel on underlying space.

DiscoGANs

Readings: (Szekely 2002-; Gretton et al. 2012-) xGANS: (Bouchacourt et al., 2016 ; Bellemarre et al., 2017)

Mixtures

$$\forall t \in [0,1] \quad \mathrm{P_t} = (1-\mathrm{t})\ \mathrm{P_0} + \mathrm{t}\,\mathrm{P_1}$$

Let the set of distributions $\mathcal{F} = \{\, G_\theta \# \mu_z :\ \theta \in \mathbb{R}^d \,\}$ be mixture-convex.

→ For all $P_0, P_1 \in \mathcal{F}$ there is $t \mapsto \theta_t \in \mathbb{R}^d$ such that $P_t = G_{\theta_t} \# \mu_z$

Problem

If P_0 and P_1 have disjoint supports with nonzero margin, then either $t \mapsto \theta_t$ is discontinuous or $\theta \mapsto G_\theta$ is discontinuous.

Mixtures

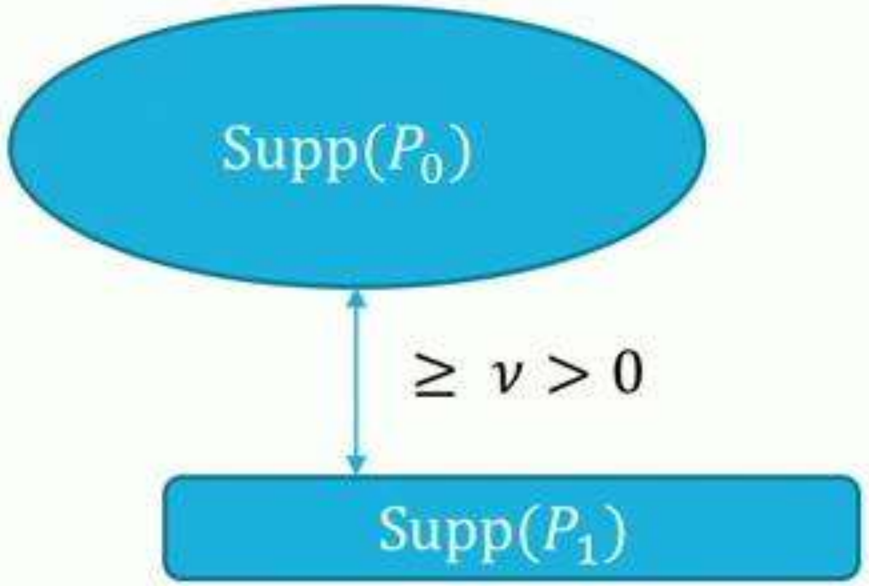
Proof:

Let P_0 and P_1 be two distributions whose supports are separated separated by a nonzero margin ν.

For all $\epsilon > 0$,

- $G_{\theta_0}(z) \in \mathrm{Supp}(P_0)$ with μ-probability one,
- For all $\epsilon > 0$, $G_{\theta_\epsilon}(z) \in \mathrm{Supp}(P_1)$ with μ-probability ϵ,

Therefore there is z such that $d\left(G_{\theta_0}(z), G_{\theta_\epsilon}(z)\right) \geq \nu > 0$



Mixtures

Proof:

Let P_0 and P_1 be two distributions whose supp…
are separated separated by a nonze…

For all $\epsilon > 0$,

- $G_{\theta_0}(z) \in \mathrm{Supp}($…
- For all $\epsilon > 0$, G_{θ_ϵ}…

$\mathrm{Supp}(P_1)$

Therefore there is z s… $) \geq \nu > 0$

Mixture curves do not match
the geometry of implicit models.

We need other kinds of curves!

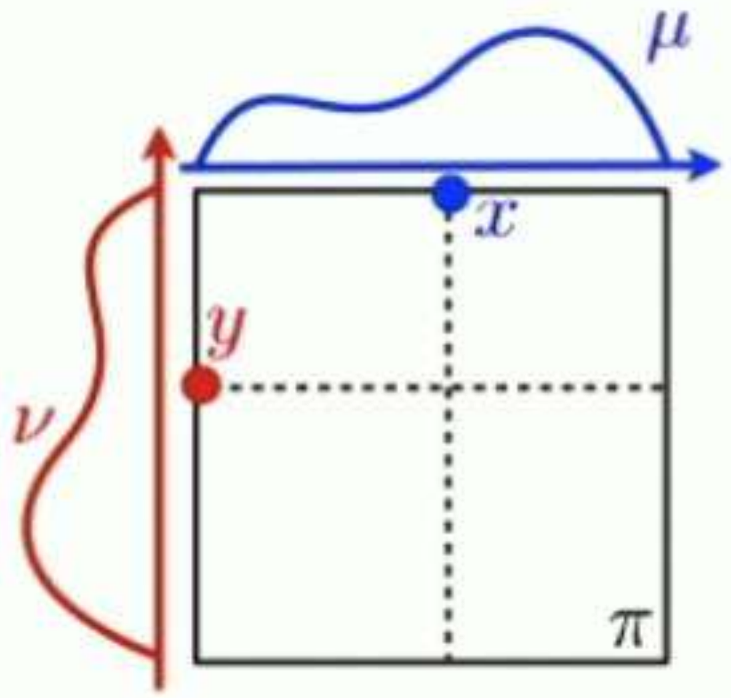
Displacement curves

Transportation plan from $\boldsymbol{P_0}$ to $\boldsymbol{P_1}$

- A joint distribution $\pi(x, y)$ whose marginals are P_0 and P_1
- Optimal when $\mathbb{E}_{(x,y)\sim\pi}[d(x, y)^p]$ is minimal

Exponent p as in Wasserstein(p) distance

Drawing stolen from Gabriel Peyré slides:
"An introduction to Optimal Transport"



Displacement curves (Euclidean)

Transportation plan from $\boldsymbol{P_0}$ to $\boldsymbol{P_1}$

- A joint distribution $\pi(x, y)$ whose marginals are P_0 and P_1
- Optimal when $\mathbb{E}_{(x,y)\sim\pi}[d(x, y)^p]$ is minimal

Displacement curve

$$P_t = \big((1-t)\,x + t\,y\big)\,\#\,\pi^*(x, y)$$

Displacement curves and implicit models

Let $P_0 = G_{\theta_0} \# \mu$ and $P_1 = G_{\theta_1} \# \mu$ be two elements of $\mathcal{F}$.

Transportation plan

$$\left(G_{\theta_0}, G_{\theta_1}\right) \# \mu$$

has displacement curves

$$P_t = \left((1-t)G_{\theta_0} + tG_{\theta_1} \right) \# \mu$$

If the family of G_θ functions has strong approximation properties,

this can be close to an optimal plan,

and this near optimal displacement curve is close to a $G_{\theta_t} \# \mu$.

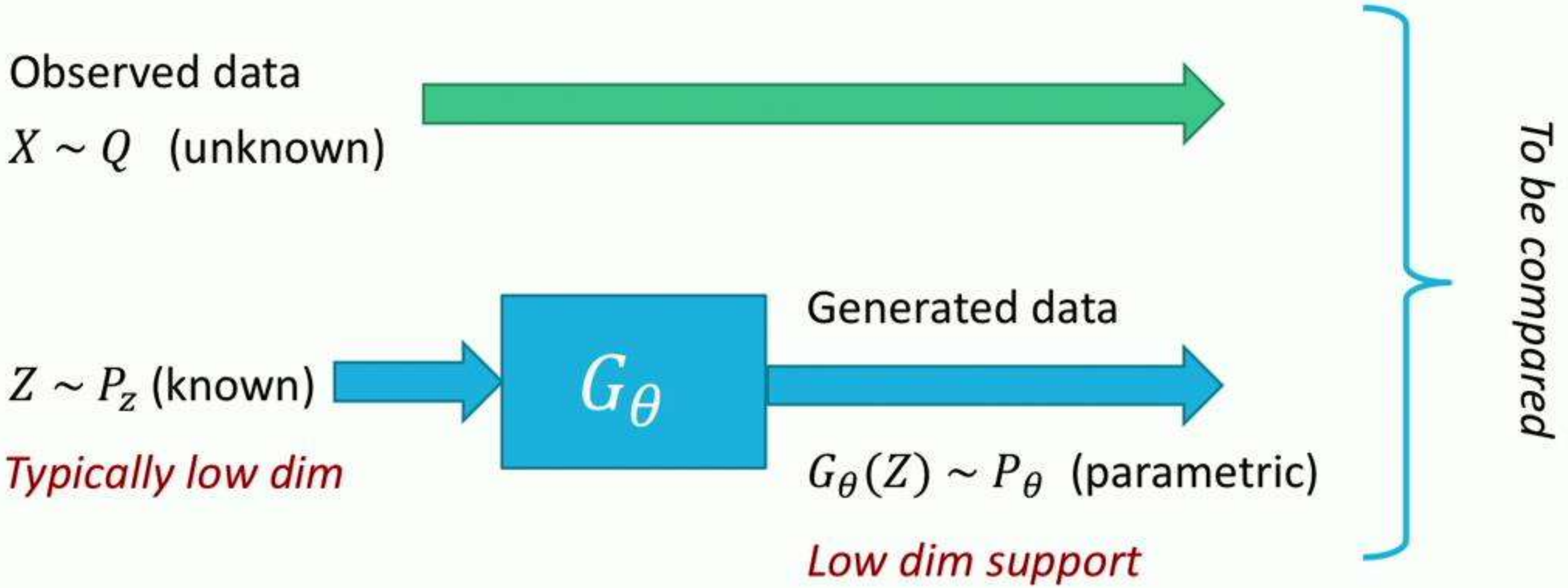
Displacement convexity and implicit models

- Displacement convexity
  is a natural notion of convexity for
  a family of distributions defined by an implicit model.

  Such families are typically not mixture-convex.

- Contrast with families defined by parametric density functions.
- Which cost functions are displacement convex, then?

Implicit modeling



How different are WD and MMD?

- Leaving aside the comparison criteria inducing a strong topology.

  (because they lead to discontinuous criteria
  when modeling distribution with disjoint supports).

- Two known criteria inducing a weak topology are

$$W_1(Q,P) = \sup_{f \in \mathrm{Lip1}} \mathbb{E}_Q[f(x)] - \mathbb{E}_P[f(x)] \ , \qquad \text{Wasserstein(1) distance}$$

$$\mathcal{E}_d(Q,P) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)] \ . \qquad \text{Energy distance, MMD}$$

Fact #1 – Minimal geodesics.

- When the space of distributions is equipped with the Energy distance $\mathcal{E}_d$ or the MMD distance $\mathcal{E}_{d_k}$, the shortest path between two distributions P_0 and P_1 is the mixture curve.
- When the space of distributions is equipped with the Wasserstein(p) distance W_p with $p > 1$, the shortest paths between two distributions P_0 and P_1 are the displacement curves.
- When the space of distributions is equipped with the Wasserstein(1) distance W_1, the shortest paths between two distributions P_0 and P_1 include the mixture curves, the displacement curves, and all kinds of hybrid curves.

Fact #2 – Statistical properties

Expected distance between a distribution Q and its empirical approximation $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$:

$$Q \in \mathcal{P}^1_{\mathcal{X}} \qquad \mathbb{E}_{x_1 \ldots x_n \sim Q}\left[\mathcal{E}_d(Q_n, Q)^2\right] = \frac{1}{n}\,\mathbb{E}_{x,x' \sim Q}[d(x,x')] = \mathcal{O}(n^{-1})\ .$$

$$Q \in \mathcal{P}^2_{R^d} \qquad \mathbb{E}_{x_1 \ldots x_n \sim Q}[\,W_1(Q_n, Q)\,] = \mathcal{O}(n^{-1/d})\ .$$

This is reached (Sanjeev's sphere)

Wasserstein seem hopeless

Fact #3 – In practice

Things look different in practice

- ED/MMD training of low dim implicit models works nicely.
- ED/MMD training of high dim implicit models often gets stuck.
- whereas "WD" training of the same high dim implicit models can give results.

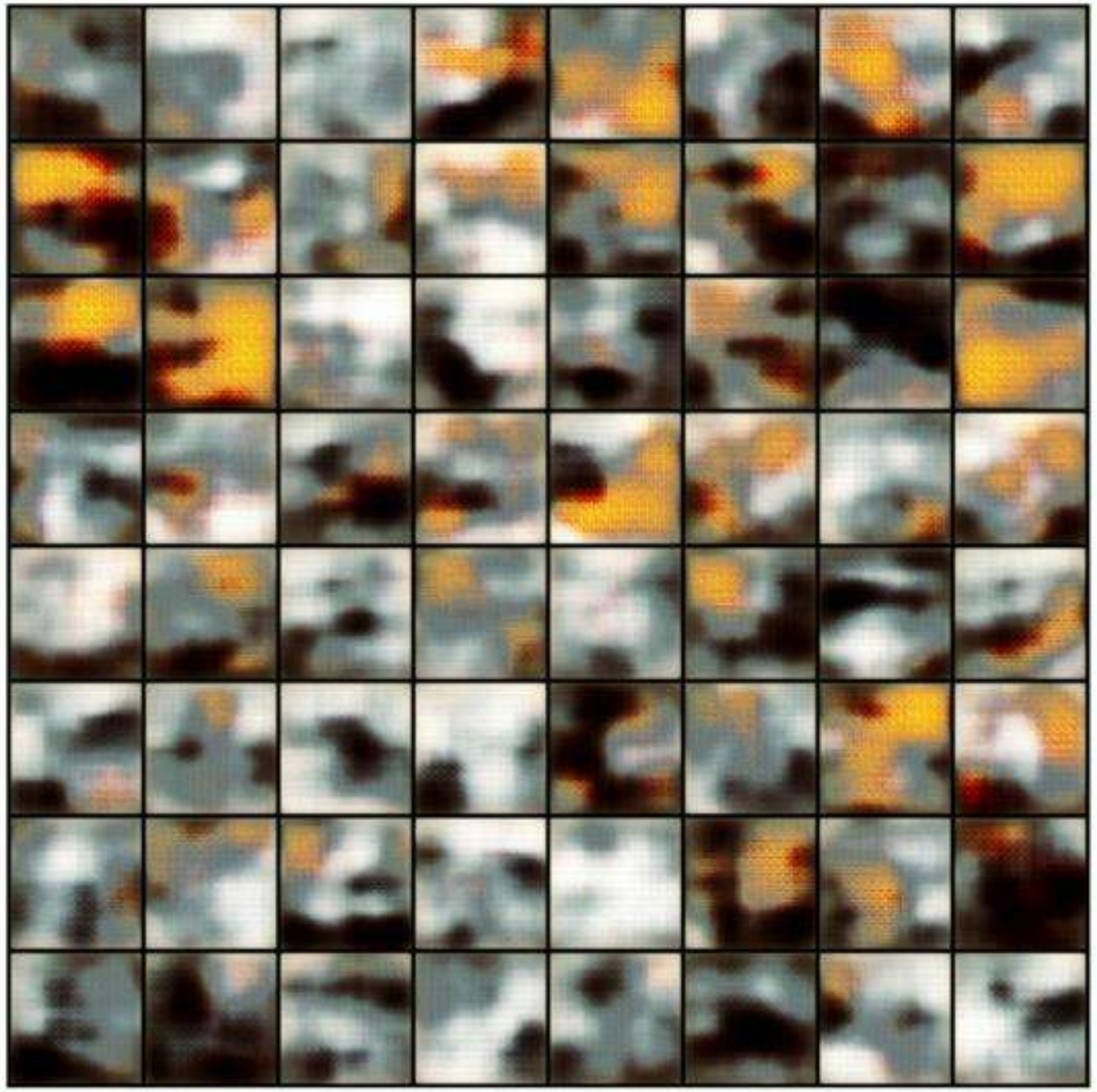
WD-like.

Just the opposite of what one would expect !

Example

A sample of 64 training examples

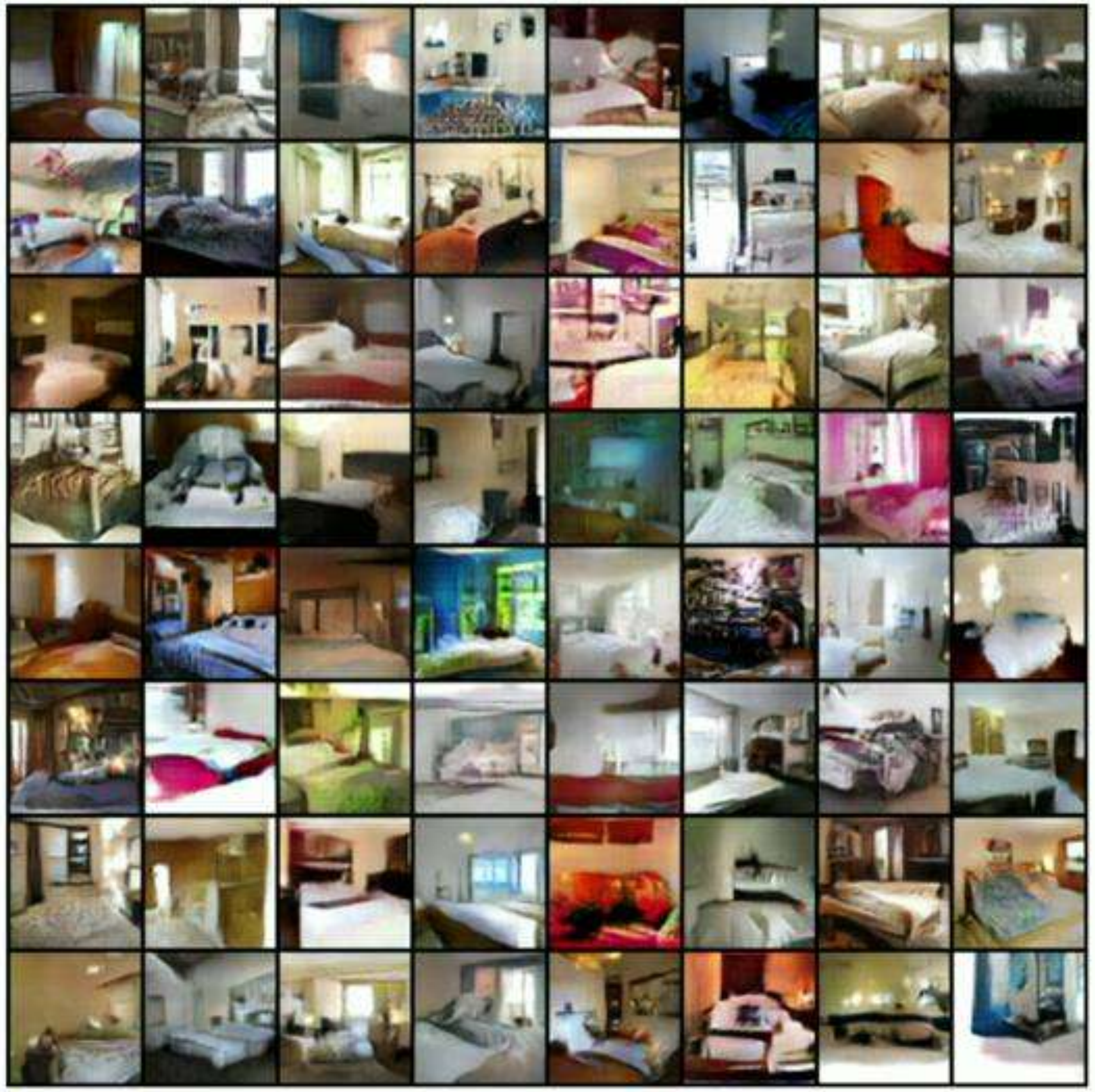
Example



Generated by the ED trained model

Example

WGAN



Generated by the WD trained model

How things can go wrong

Example 6.5 Let μ_z be the uniform distribution on $\{-1, +1\}$. Let the parameter θ be constrained to the square $[-1, 1]^2 \subset \mathbb{R}^2$ and let the generator function be

$$G_\theta : z \in \{-1, 1\} \mapsto G_\theta(z) = z\theta \ .$$

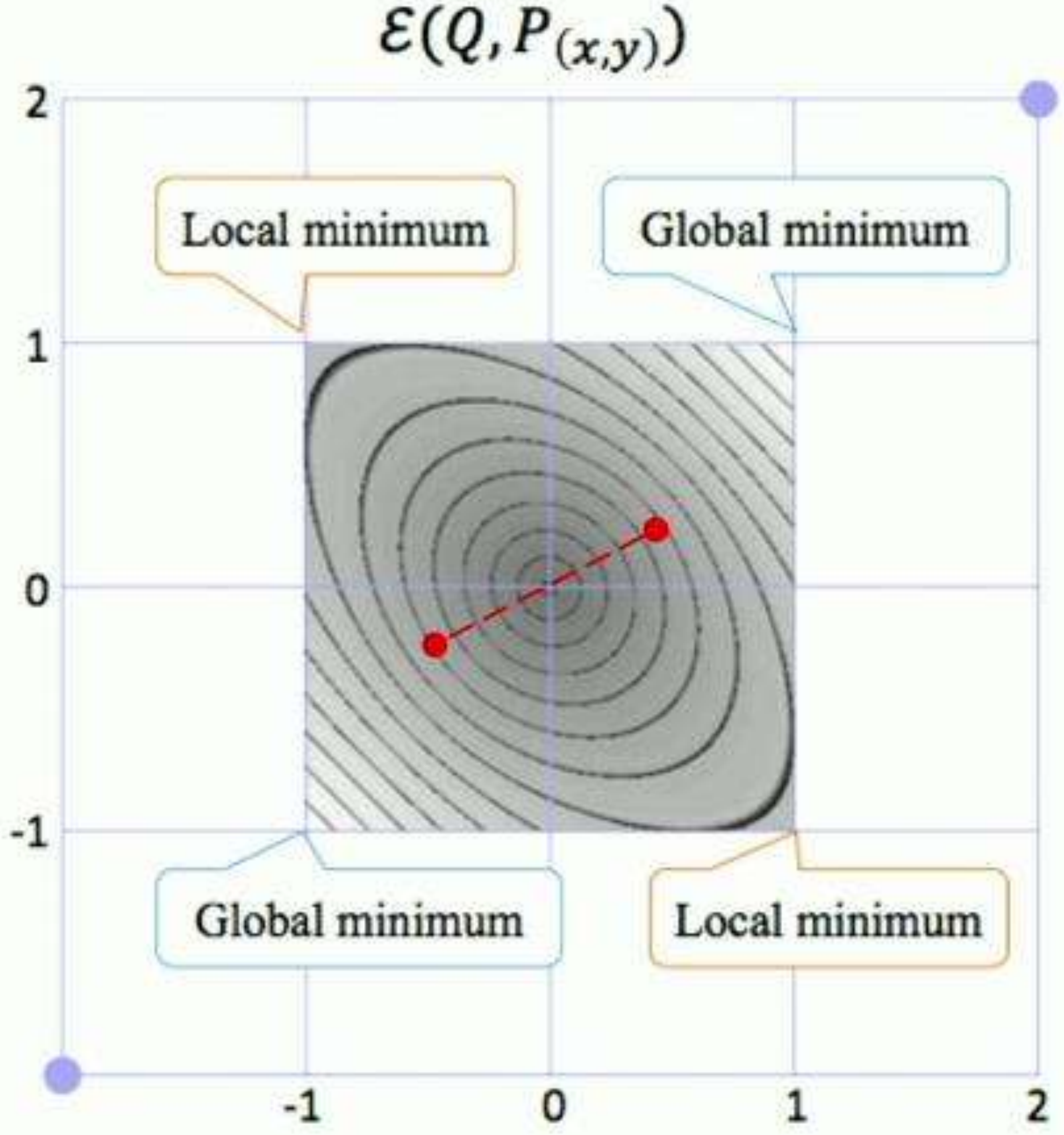
The corresponding model family is

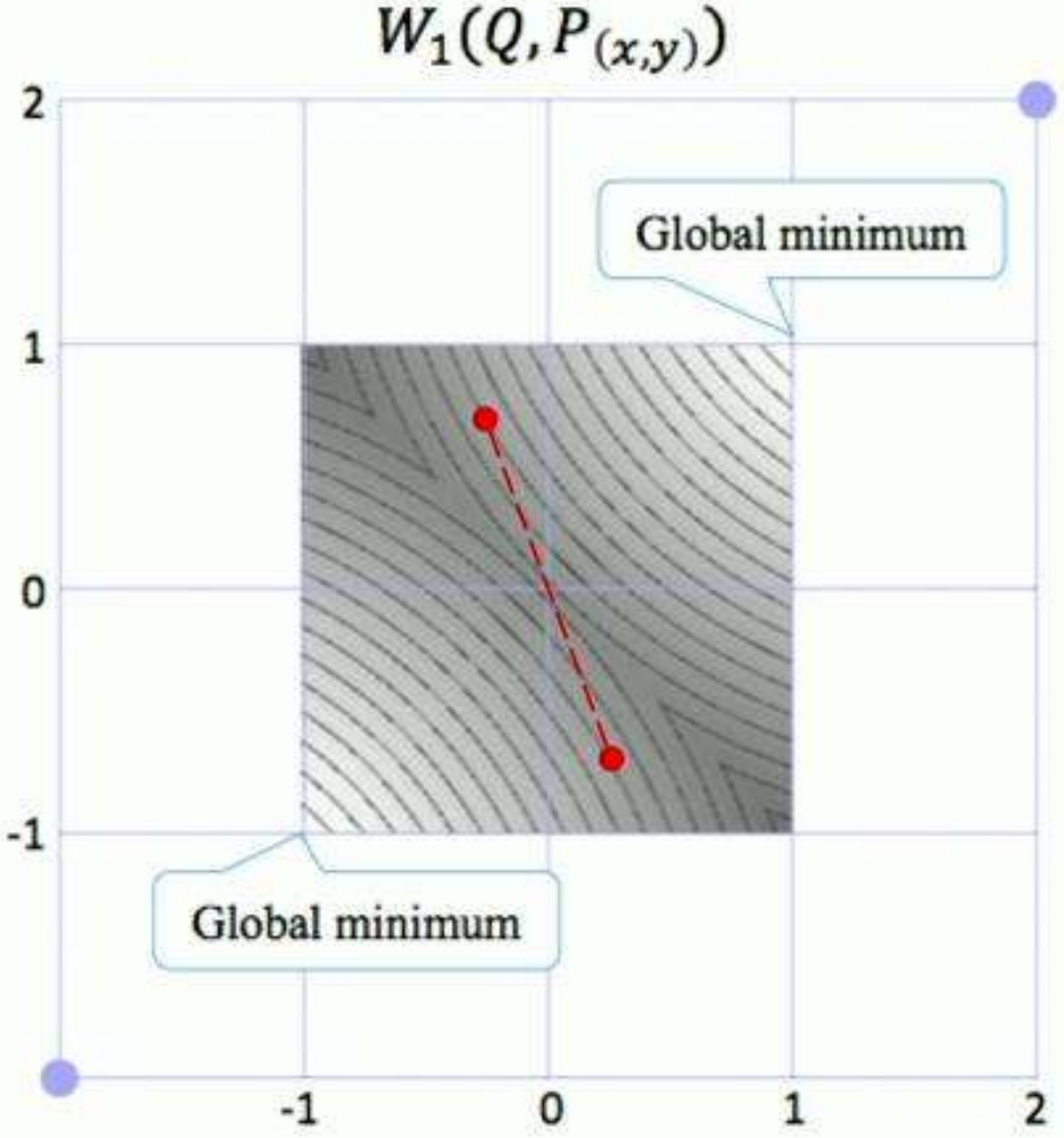
$$\mathcal{F} = \left\{ P_\theta = \tfrac{1}{2}(\delta_\theta + \delta_{-\theta}) : \theta \in [-1, 1] \times [-1, 1] \right\} \ .$$

Two Dirac distributions with mean zero in a square.

It is easy to see that this model family is displacement convex but not mixture convex. Figure 5 shows the level sets for both criteria $\mathcal{E}(Q, P_\theta)$ and $W_1(Q, P_\theta)$ for the target distribution $Q = P_{(2,2)} \notin \mathcal{F}$. Both criteria have the same global minima in $(1, 1)$ and $(-1, -1)$. However the energy distance has spurious local minima in $(-1, 1)$ and $(1, -1)$ with a relatively high value of the cost function.

How things can go wrong

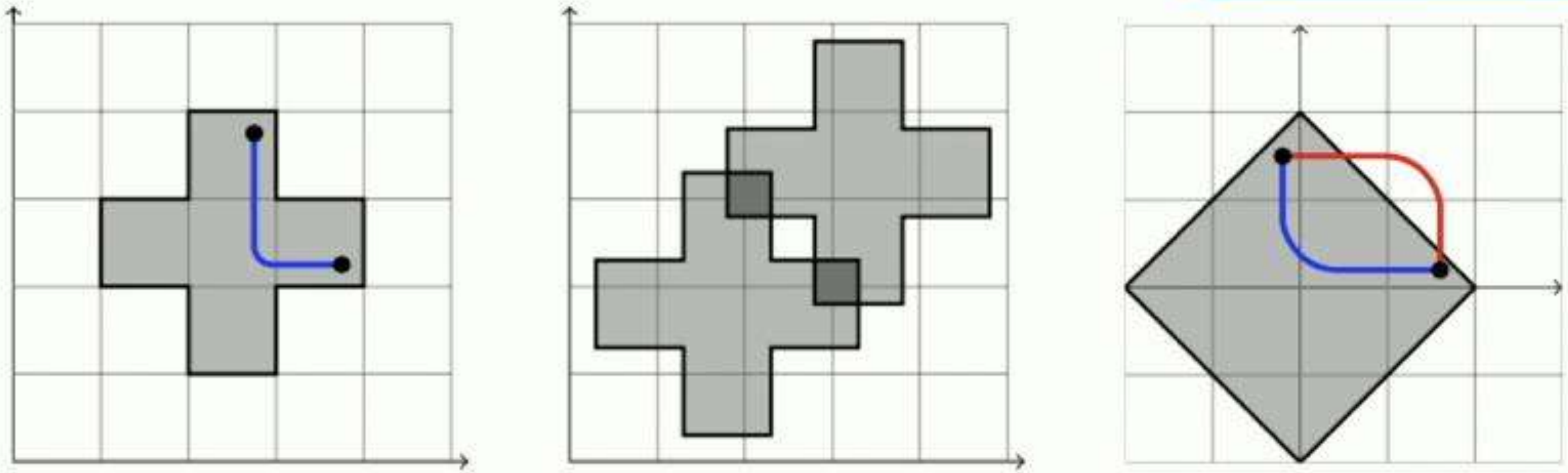




The convexity of distance functions

- Learn by minimizing $\min_{P_\theta \in \mathcal{F}} D(Q, P_\theta)$

When is the cost function
$P \mapsto D(Q,P)$
mixture-convex ?
displacement-convex?



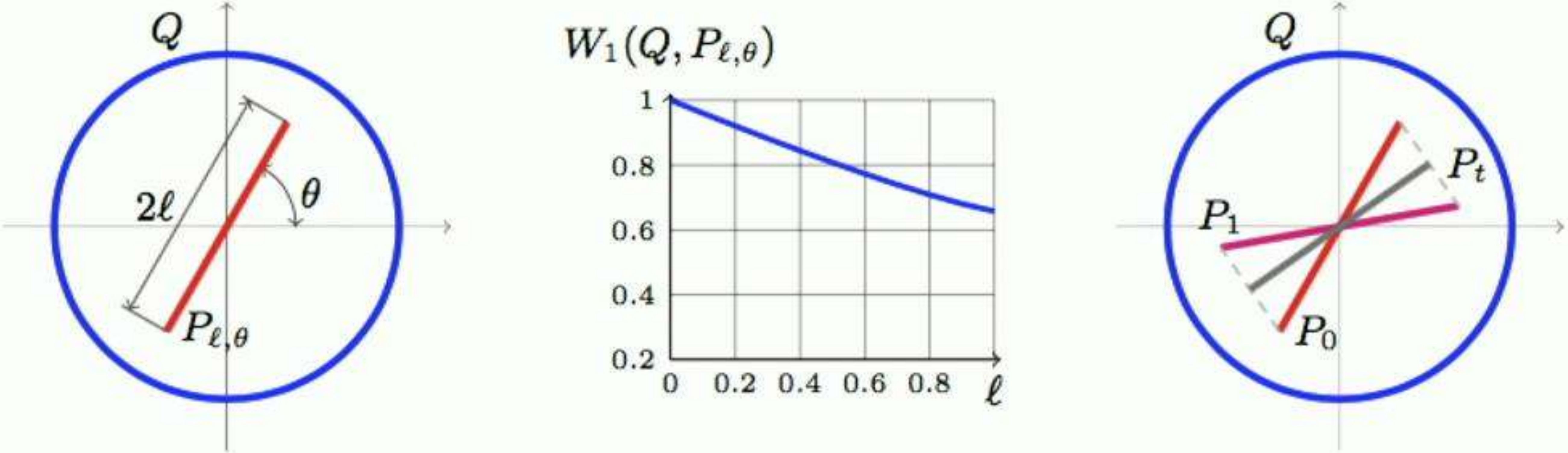
Mixture-convexity

Proposition 6.6. *Let $\mathcal{P}_{\mathcal{X}}$ be equipped with a distance D that belongs to the IPM family* (5). *Then D is mixture convex.*

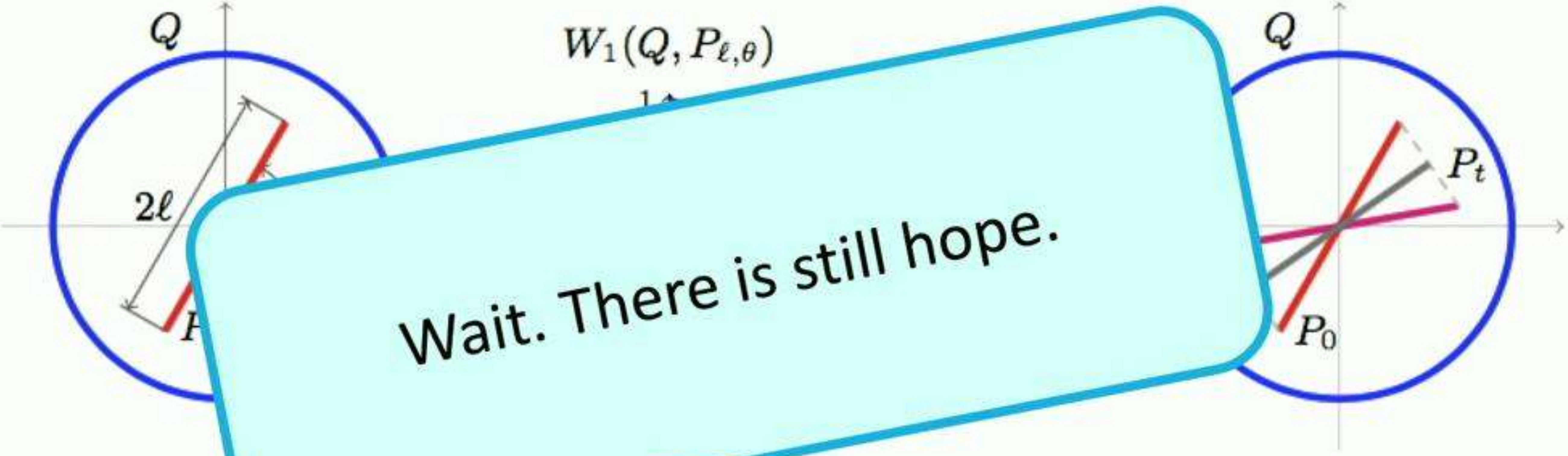
Therefore

- Cost function $P \mapsto \mathcal{E}_d(Q, P)$ is mixture convex.
- Cost function $P \mapsto W_1(Q, P)$ is mixture convex

The Wasserstein distance is not displacement convex



The Wasserstein distance is not displacement convex



Wait. There is still hope.

Cost function $\mathrm{P} \mapsto W_1(Q, P)$ is almost displacement-convex

Proposition 6.8. *Let $\mathcal{X}$ be a strictly intrinsic Polish space equipped with a geodesically convex distance d and let $\mathcal{P}^1_{\mathcal{X}}$ be equipped with the 1-Wasserstein distance W_1. For all $Q \in \mathcal{P}_{\mathcal{X}}$ and all displacement geodesics $t \in [0,1] \mapsto P_t$,*

$$\forall t \in [0,1] \quad W_1(Q, P_t) \leq (1-t)\, W_1(Q, P_0) + t\, W_1(Q, P_1) + 2t(1-t) K(Q, P_0, P_1)$$

with $K(Q, P_0, P_1) \leq 2 \min_{u_0 \in \mathcal{X}} \mathbb{E}_{u \sim Q}[d(u, u_0)]$.

GROSS BOUND

BOUND THE CONVEXITY VIOLATION

Cost function $\mathrm{P} \mapsto W_1(Q, P)$ is almost displacement-convex

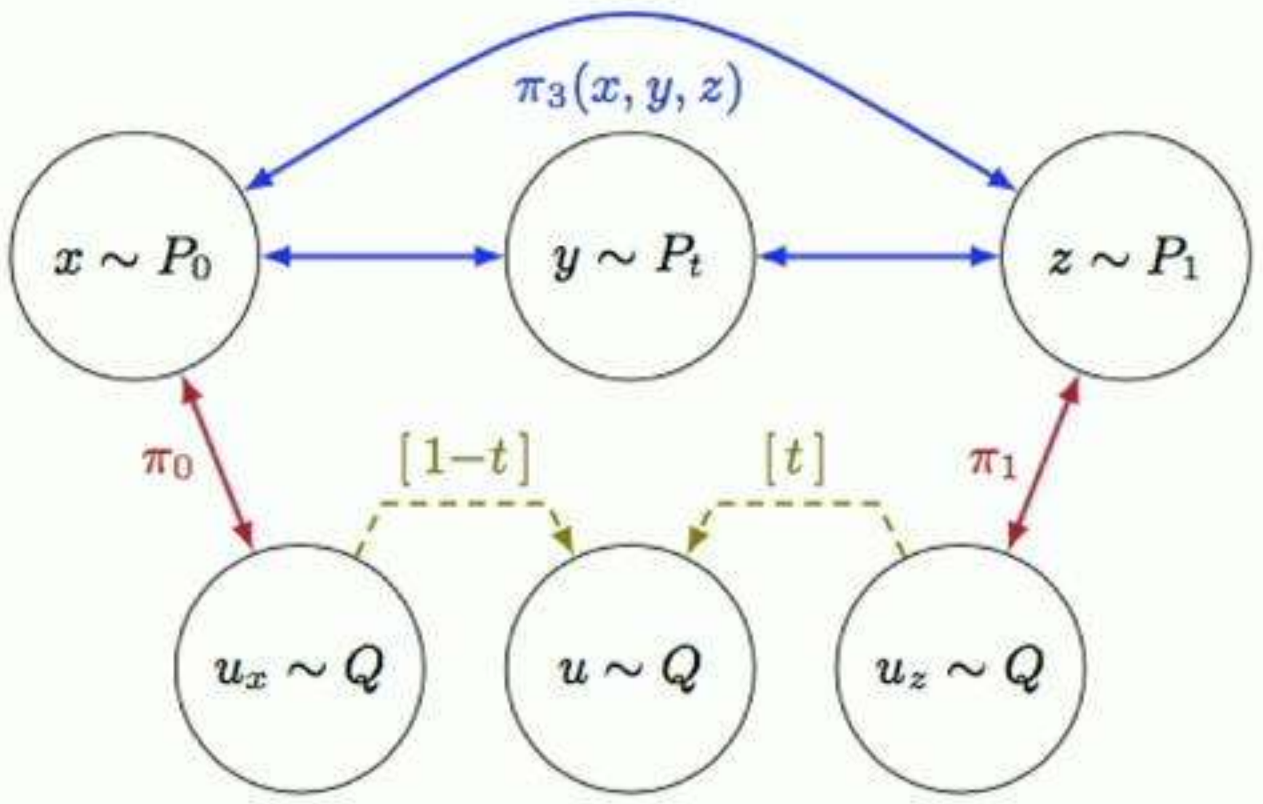


Fig. 8. The construction of $\pi \in \mathcal{P}_{\mathcal{X}^6}$ in the proof of Proposition 7.8.

Cost function $\mathrm{P} \mapsto W_1(Q, P)$ is almost displacement-convex

We can therefore apply the almost-convex-optimization-a-la-carte-theorem
and conclude guarantee that optimizing an implicit model with WD
has only local minima whose value is "near" that of the global minimum.

Although I am not very happy
with this bound (too gross).

Conclusion

- Convexity with respect to mixture curves makes clear that optimizing a regression model with strong approximation properties with a descent algorithm yields a near global minimum.
- This property is independent of the exact parametrization.
  → It says nothing about the implicit biases induced by the parametrization
- In implicit generative models, convexity with respect to displacement curves seems more interesting than convexity with respect to mixture curves.
  → Is there potential here? Displacement in images versus mixtures of images.

Discussion (general)

- These results are independent from the parametrization of $\mathcal{F}$.
  They depend on whether any two points in $\mathcal{F}$ can be connected by a suitable curve that (a) either remains in $\mathcal{F}$, or (b) can be well approximated by elements of $\mathcal{F}$.
- In θ space, the level sets can be very nonconvex, and yet connected.
- However, because learning algorithms operate in θ space,
  the parametrization changes the implicit biases that affect
  - which global minimum is returned in overparametrized models, or,
  - which solution is returned after early stopping.

Convex optimization « à la carte »

Theorem

Let $\mathcal{F} \subset \mathfrak{X}$ be convex with respect to $\mathcal{C}$.

Let the cost function $f : \mathfrak{X} \to \mathbb{R}$ be endpoints-convex with respect to $\mathcal{C}$.

Then:

- $\forall\, M \geq \min_{\mathcal{F}} f,$ the level sets $L(f, \mathcal{F}, \mathrm{M}) = \{x \in \mathcal{F} \;\; s.t.\; f(x) \leq M\}$ are connected.
- If $\mathcal{C}$ only contains bounded speed curves, all local minima of f in $\mathcal{F}$ are global.

Proof (2)

- A point $x \in \mathcal{F}$ is a local minimum of f in $\mathcal{F}$ iff there is $\epsilon > 0$ such that, for all $x' \in \mathcal{F}$, $d(x, x') < \epsilon \implies f(x') \geq f(x)$.
- Reasoning by contradiction, assume there is $y \in \mathcal{F}$ such that $f(y) < f(x)$.
- Let $\gamma \in \mathcal{C}$ be a bounded speed curve contained $\mathcal{F}$ and connecting x to y :
$$\forall \;\; 0 \leq t \leq t' \leq 1 \quad d(\gamma_t, \gamma_{t'}) \; \leq K\,(t' - t)$$
- Therefore $f\left(\gamma_{\epsilon/2K}\right) \geq f(\gamma_0) = f(x)$
- But endpoints convexity means $f\left(\gamma_{\epsilon/2K}\right) \leq \left(1 - \frac{\epsilon}{2K}\right) f(x) + \frac{\epsilon}{2K} f(y) < f(x)$!!!