§4. Mean-Field Networks.

1. Recall $f(x;W) = \frac{1}{m}\sum_{k=1}^{m} \phi(x; w_k)$ → activation e.g. $\phi(x;w) = \mathrm{ReLU}(w_0 + \langle x, w_{1:d}\rangle)$

Let $\mu = \frac{1}{m}\sum_{k=1}^{m} \delta_{w_k}$ ← the empirical distr. of the neurons.

$\Rightarrow f(x;W) = \int \phi(x;w)\, d\mu(w) =: f(x;\mu)$

Allow μ to be any (reasonably regular) distribution on $\mathbb{R}^{d'}$

$\Rightarrow$ mean-field networks.

· MSE: $\mathcal{L}(\mu) = \frac{1}{2}\mathbb{E}_{x,y}\left\{ (y - f(x;\mu))^2 \right\}$

↳ a functional over probability measures.

GF of $\mathcal{L}$?

2. The vector space structure

Notations. $\mathcal{M}_\pm(\mathbb{R}^d) = \{$ signed measures over $\mathbb{R}^d \}$

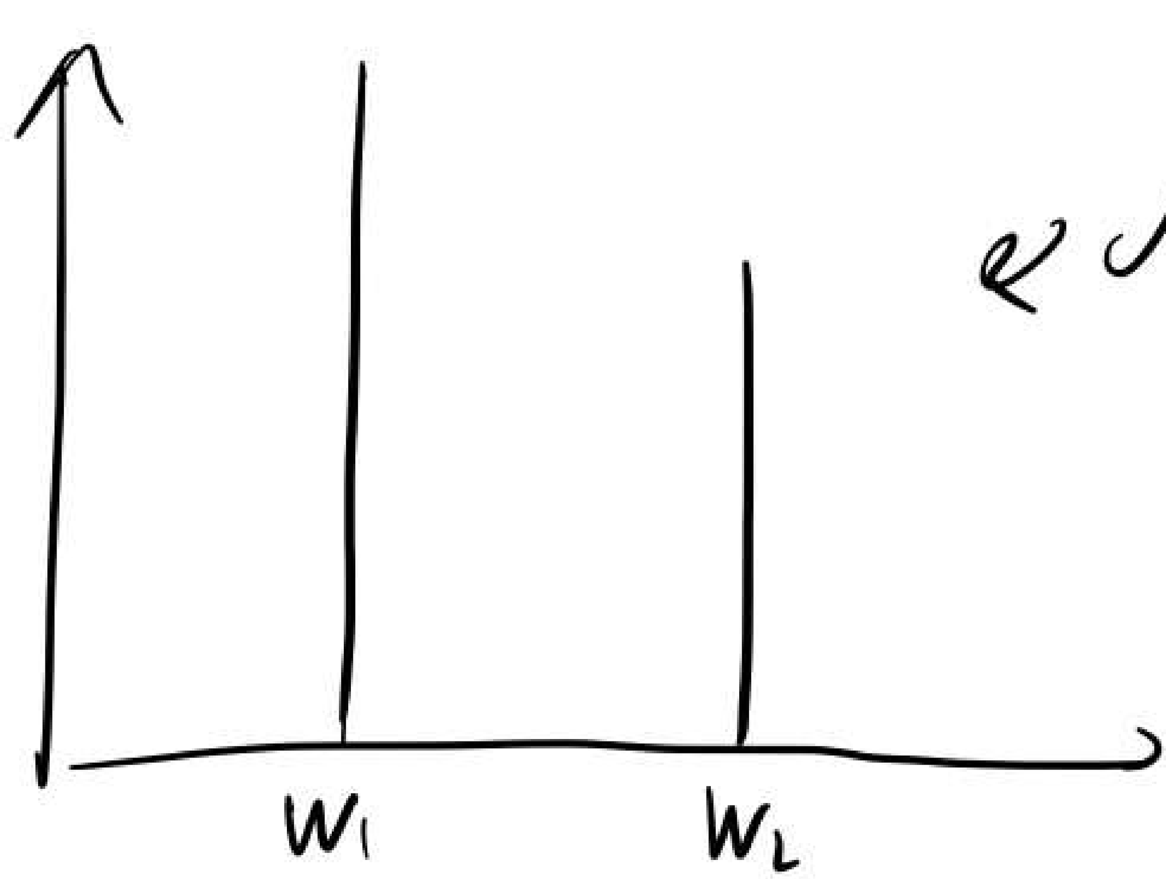
$P(\mathbb{R}^d) = \{$ probability measures $\sim \}$, $P_2(\mathbb{R}^d) = \{ \mu \in P(\mathbb{R}^d) : \int \|x\|^2 d\mu < \infty \}$

The vector space structure: $\mu, \nu \in \mathcal{M}_\pm(\mathbb{R}^d)$, $a, b \in \mathbb{R}$

measurable E $\quad (a\mu \oplus b\nu)(E) = a\mu(E) + b\nu(E)$.

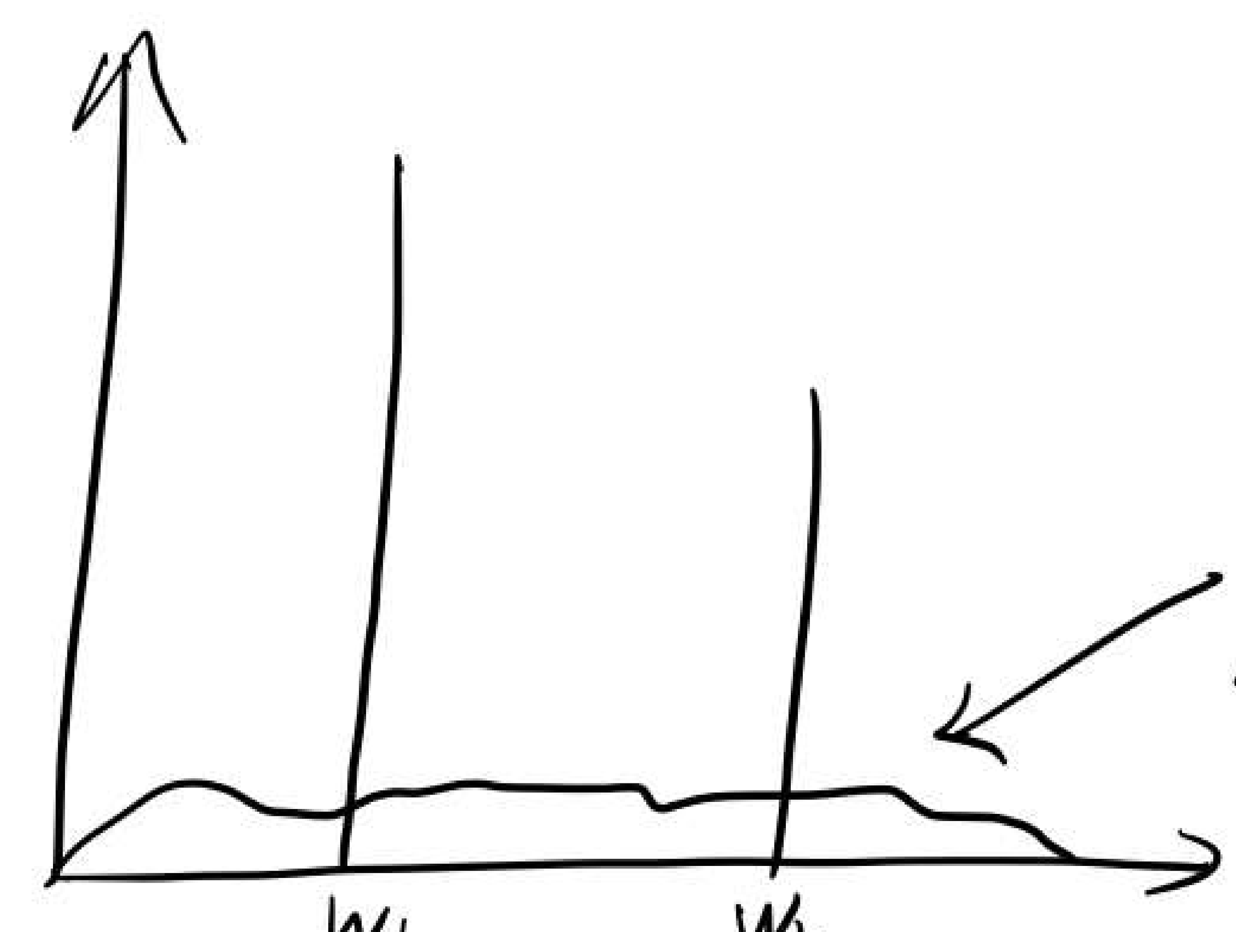
Let $P(\mathbb{R}^d)$ inherit this structure.

Q: Is this the "correct" structure?



$\mu = \frac{1}{2}\delta_{w_1} + \frac{1}{2}\delta_{w_2}$

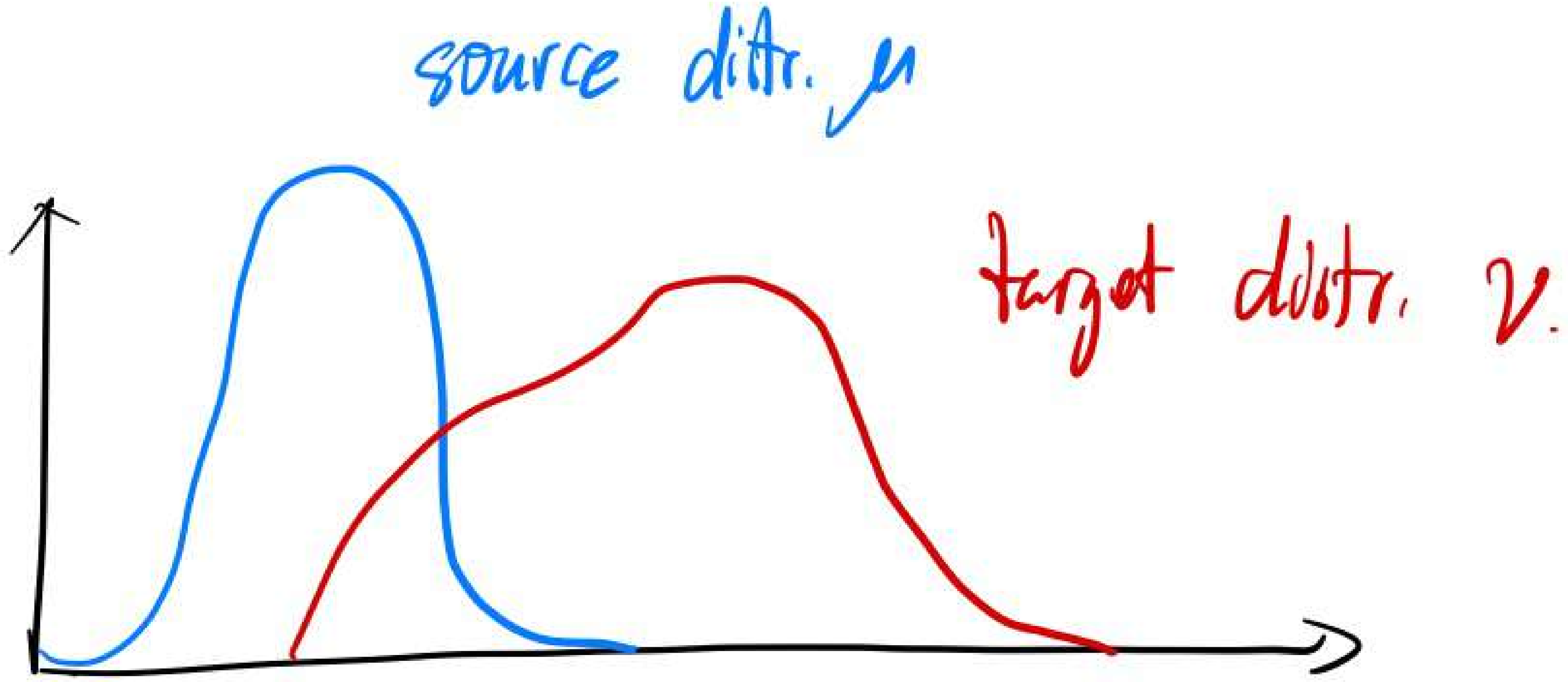
add a "small" perturbation ν



ν

3. Wasserstein-2 space $\mathbb{W}_2$

Sand moving



cost of moving 1 unit of sand from x to y: $\|x-y\|^2$

Goal: minimize the total cost.

$$\underset{T}{\text{minimize}} \int \|x - T(x)\|^2 \, d\mu(x) \quad \text{s.t.} \quad T\#\mu = \nu.$$

Def. Given $T: \mathbb{R}^d \to \mathbb{R}^d$ (transport map)

$$T\#\mu(E) := \mu(T^{-1}(E)).$$

Def. For two (sufficiently regular) probability measure $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the <u>Wasserstein-2 distance</u> between μ and ν is defined as.

$$W_2^2(\mu, \nu) = \inf \left\{ \int \|x - T(x)\|^2 \, d\mu(x) \; : \; T\#\mu = \nu \right\}$$

Remark. Use couplings instead of transport maps for less regular distributions.

Thm. $\mathbb{W}_2(\mathbb{R}^d) := (\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space

the constant speed geodesic between μ and ν is given by

$$t \mapsto \big((1-t)I_d + tT\big)_{\#}\mu, \quad t\in[0,1],$$

where T is the optimal transport map.

4. Wasserstein gradient flow

Remark. It's possible to define GF in general metric spaces. but we'll focus on $\mathbb{W}_2$.

Def. $F: \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$, $\mu \in \mathcal{P}(\mathbb{R}^d)$. We say $G: \mathbb{R}^d \to \mathbb{R}$ is the first variation of F at μ if $\forall\, \chi \in \mathcal{M}_{\pm}(\mathbb{R}^d)$ with $\mu + \varepsilon\chi \in \mathcal{P}(\mathbb{R}^d)$ $\forall \varepsilon > 0$, we have

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} F(\mu+\varepsilon\chi) = \int \underbrace{G(x)}_{=:\frac{\delta F}{\delta\mu}[\mu]}\, d\chi(x)$$

Remark. 1) directional derivative. 2) local linear approximation.

Example. 1) $F(\mu) = \int f(x)\, d\mu(x) \Rightarrow \frac{\delta F}{\delta \mu}[\mu](x) = f(x)$

2) $F(\mu) = \iint W(x,y)\, d\mu(x)\, d\mu(y)$. W symmetric $\Rightarrow \frac{\delta F}{\delta \mu}[\mu](x) = 2\int W(x,y)\, d\mu(y)$.

Q: How to minimize $F(\mu) = \int G(x)\, d\mu(x)$

$\longrightarrow$ cost of placing 1 unit of particles at x.

A: At each x, move the particles along $-\nabla G(x)$.

$$\Rightarrow \frac{d}{dt} X_t = -\nabla G(X_t) = -\nabla \frac{\delta F}{\delta \mu}[\mu](X_t). \quad \forall X_t \in \mathrm{supp}(\mu_t) \qquad (*)$$

$$\Leftrightarrow \partial \mu_t - \nabla \cdot \left(\mu_t \nabla \frac{\delta F}{\delta \mu}[\mu] \right) = 0. \qquad (**)$$

Def: $F: W_2 \to \mathbb{R}$. We say μ_t is the Wasserstein gradient flow of F if it satisfies the continuity equation $(*)$ or $(**)$.

Lemma (Descent lemma for WGF).

$$\frac{d}{dt} F(\mu_t) = \langle \nabla F, \dot{\mu}_t \rangle = \mathbb{E}_{v\sim\mu_t} \langle \nabla \frac{\delta F}{\delta \mu}[\mu](v), \dot{v} \rangle$$

formal

$$= - \mathbb{E}_{v\sim\mu_t} \| \nabla \frac{\delta F}{\delta \mu}[\mu](v) \|^2$$

3. Thm. (Chizat & Bach 2018. Thm 2.6).

$\mu_{m,0} = \frac{1}{m}\sum_{k=1}^{m} \delta_{w_k}$ - empirical distr. of the initialization of the finite width network with m neurons

$(\mu_{m,t})_t$. - obtained by running the classical GF on the finite-width network.

μ_0 - the initialization distribution

$(\mu_t)_t$. — WGF of L started from μ_0

Under some regularity conditions, $\mu_{m,t} \to \mu_t$.

Caveat. In general, we need $\underset{\sim\sim\sim}{m = \exp(d)}$ for the difference between $\mu_{m,t}$ and μ_t to be sufficiently small.

6. First-variation of $\mathcal{L}(\mu) = \frac{1}{2}\mathbb{E}_x\{(f_*(x) - f(x;\mu))^2\}$

$\varepsilon > 0$, perturbation ν.

$$\mathcal{L}(\mu+\varepsilon\nu) = \frac{1}{2}\mathbb{E}_x\{(f_*(x) - f(x;\mu+\varepsilon\nu))^2\}$$

$$= \frac{1}{2}\mathbb{E}_x f_*^2(x) + \frac{1}{2}\mathbb{E}_x f^2(x;\mu+\varepsilon\nu) - \mathbb{E}_x f_*(x) f(x;\mu+\varepsilon\nu)$$

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathbb{E}_x f_*(x) f(x;\mu+\varepsilon\nu) = \mathbb{E}_x\Big\{f_*(x) \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \int \phi(x;w)\, d(\mu+\varepsilon\nu)(w)\Big\} = \mathbb{E}_x\Big\{f_*(x)\int\phi(x;w)\, d\nu(w)\Big\}$$

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \frac{1}{2}\mathbb{E}_x f^2(x;\mu+\varepsilon\nu) = \mathbb{E}_x\Big\{f(x;\mu)\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} f(x;\mu+\varepsilon\nu)\Big\} = \mathbb{E}_x\Big\{f(x;\mu)\int\phi(x;w)\, d\nu(w)\Big\}$$

$$\Rightarrow \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathcal{L}(\mu+\varepsilon\nu) = \int \mathbb{E}_x\{(f_*(x) - f(x;\mu))\,\phi(x;w)\}\, d\nu(w)$$

$$\Rightarrow \frac{\delta \mathcal{L}}{\delta\mu}[\mu](v) = \mathbb{E}_x\{(f_*(x) - f(x;\mu))\,\phi(x;v)\} = \langle f_* - f(\cdot;\mu), \phi(\cdot;v)\rangle_{L^2}.$$

7. Global convergence

Def. We say $\{G(\cdot;v)\}_{v\in\mathbb{R}^d}$ satisfies the <u>universal approximation property</u> if its span is dense in L^2.

Thm (CB18, Thm 3.3; PN21, Thm 8)

$(\mu_t)_t$ - WGF of L from μ_0

$\operatorname{supp}\mu_0 = \mathbb{R}^d$. $\mu_t \to \mu_\infty$

G universal approximation. $\phi(\cdot;w) = w_0\,\sigma(\cdot;w_{1:d})$

$\Rightarrow$ Under some regularity conditions,

μ_∞ is a global minimizer of L.

<u>proof idea</u>. Assume $\operatorname{supp}\mu_\infty = \mathbb{R}^d$.

(This is a strong assumption. Can be avoided using some algebraic topological argument).

Descent lemma for WGF, $\phi(\cdot;0)\equiv 0$.

$\Rightarrow$ a.e. $v\in\operatorname{supp}\mu_\infty = \mathbb{R}^d$.

$$\frac{\delta L}{\delta\mu}[\mu_\infty] = \langle f(\cdot;\mu_\infty) - f_*, \phi(\cdot;v)\rangle_{L^2} \equiv 0$$

Universal approximation.

$$\Rightarrow \exists\, g_m = \sum_{k=1}^m \phi(\cdot;v_k) \to f - f_*$$

$$\Rightarrow 0 = \sum_{k=1}^m \langle f - f_*, \phi(\cdot;v_k)\rangle_{L^2}$$

$$= \langle f - f_*, g_m\rangle_{L^2} \to \|f - f_*\|_{L^2}^2$$