Gradient flows on Graphons

Sewoong Oh[1], Soumik Pal[2], Raghav Somani[1] and Raghav Tripathi[2]

[1]UW CSE & [2]UW Math

March 18, 2022

Objective

Study large scale optimization problems that have permutation symmetries.

- Exploiting symmetries allow taking limits of the size of optimization problems.

Objective

Study large scale optimization problems that have permutation symmetries.

- Exploiting symmetries allow taking limits of the size of optimization problems.

For $n \in \mathbb{N}$, consider minimizing the following interaction energy $V_n \colon \mathbb{R}^n \to \mathbb{R}_+$

$$V_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2} (x_i - x_j)^2 .$$

Objective

Study large scale optimization problems that have permutation symmetries.

- Exploiting symmetries allow taking limits of the size of optimization problems.

For $n \in \mathbb{N}$, consider minimizing the following interaction energy $V_n \colon \mathbb{R}^n \to \mathbb{R}_+$

$$V_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2}(x_i - x_j)^2 \,.$$

- Starting from $\{X_{i,0}\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \rho_0$, one can perform a gradient flow:

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t})\,\mathrm{d}t \,, \qquad \forall\, i \in [n],\ t \geq 0 \,.$$

Objective

Study large scale optimization problems that have permutation symmetries.

- Exploiting symmetries allow taking limits of the size of optimization problems.

For $n \in \mathbb{N}$, consider minimizing the following interaction energy $V_n : \mathbb{R}^n \to \mathbb{R}_+$

$$V_n(x) \coloneqq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{2} (x_i - x_j)^2 .$$

- Starting from $\{X_{i,0}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \rho_0$, one can perform a gradient flow:

$$\mathrm{d} X_{i,t} = -\frac{1}{n} \sum_{j=1}^{n} (X_{i,t} - X_{j,t}) \, \mathrm{d}t , \qquad \forall \ i \in [n], \ t \geq 0 .$$

- Notice that V_n is essentially a function of the empirical measure of its inputs!

$$V_n(x) = \mathrm{Var}(\mathrm{Emp}_n(x)) .$$

Can we approximate this problem by lifting it over the space of measures?

Particle System to Measures

- If a function $V_n : \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.

Particle System to Measures

- If a function $V_n : \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var}\,.$$

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var}\,.$$

- As well solve latter using *Wasserstein gradient flows*!

Particle System to Measures

- If a function $V_n\colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V\colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var}\ .$$

- As well solve latter using *Wasserstein gradient flows*!
  - Consider the ODE

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t})\,\mathrm{d}t\ , \qquad \forall\ i \in [n],\ t \geq 0\ ,$$

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var} \ .$$

- As well solve latter using *Wasserstein gradient flows*!
  - Consider the ODE with an added diffusion process

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t})\,\mathrm{d}t \quad + \sqrt{2}\,\mathrm{d}B_{i,t}\ , \qquad \forall\ i \in [n],\ t \geq 0\ ,$$

where B_t is the standard Brownian motion on $\mathbb{R}^n$.

Particle System to Measures

- If a function $V_n : \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \mathrm{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \mathrm{Var} \, .$$

- As well solve latter using *Wasserstein gradient flows*!
  - Consider the ODE with an added diffusion process

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t})\,\mathrm{d}t \quad + \sqrt{2}\,\mathrm{d}B_{i,t} \, , \qquad \forall\, i \in [n],\ t \geq 0 \, ,$$

  where B_t is the standard Brownian motion on $\mathbb{R}^n$.
  - This SDE captures the Wasserstein gradient flow of $\mathrm{Var} + \mathrm{Ent} : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.

Particle System to Measures

- If a function $V_n \colon \mathbb{R}^n \to \mathbb{R}$ is invariant under permutations of its input, then it can be extended to a function $V \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.
- For the interaction energy V_n, we know that $V(\rho) = \operatorname{Var}(\rho)$ for $\rho \in \mathcal{P}(\mathbb{R})$.
- Notice that for all $n \in \mathbb{N}$:

$$\min_{\mathbb{R}^n} V_n = \min_{\mathcal{P}(\mathbb{R})} \operatorname{Var} .$$

- As well solve latter using *Wasserstein gradient flows*!
  - Consider the ODE with an added diffusion process

$$\mathrm{d}X_{i,t} = -\frac{1}{n}\sum_{j=1}^{n}(X_{i,t} - X_{j,t})\,\mathrm{d}t \quad + \sqrt{2}\,\mathrm{d}B_{i,t} \ , \qquad \forall\, i \in [n],\ t \geq 0 \ ,$$

    where B_t is the standard Brownian motion on $\mathbb{R}^n$.
  - This SDE captures the Wasserstein gradient flow of $\operatorname{Var} + \operatorname{Ent} \colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$.

Upshot

Allows approximability to finite dimensional version, under mild assumptions.

Optimization on Large Graphs

Q. What about optimization over dense unlabeled (weighted) graphs?

Optimization on Large Graphs

Q. What about optimization over dense unlabeled (weighted) graphs?

Triangle density

Let G be a finite simple graph with n vertices,

$$h_{\triangle}(G) = \frac{|\text{Number of triangles in } G|}{n^3} .$$

Scalar Entropy

For a graph G with adjacency matrix A, let $h(p) = p \log p + (1-p) \log(1-p)$,

$$E(G) = \frac{1}{n^2} \sum_{i,j=1}^{n} h(A_{i,j}) .$$

Optimization on Large Graphs

Q. What about optimization over dense unlabeled (weighted) graphs?

Triangle density

Let G be a finite simple graph with n vertices,

$$h_{\triangle}(G) = \frac{|\text{Number of triangles in } G|}{n^3} \, .$$

Scalar Entropy

For a graph G with adjacency matrix A, let $h(p) = p \log p + (1-p) \log(1-p)$,

$$E(G) = \frac{1}{n^2} \sum_{i,j=1}^{n} h(A_{i,j}) \, .$$

A Problem on Large Graphs

Consider minimizing $h_{\triangle} + E$ over the set of all graphs. (e.g. Chatterjee & Varadhan)

Is there a symmetry?

Is there a symmetry?

- Notice that unlabeled graphs have a symmetry under vertex relabeling.

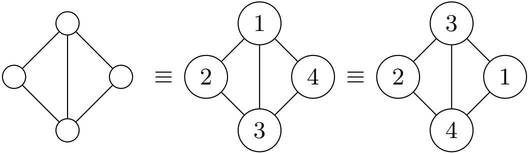


Figure: Symmetry in unlabeled graphs.

- I.e., for an unlabeled graph G with n vertices.
  If A is its adjacency matrix, so is $A_\pi = \left(A_{\pi(i),\pi(j)}\right)_{i,j}$.

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \equiv \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} = A_\pi \ .$$

- This makes functions over graphs *invariant* under this symmetry.

Neural Networks: Another Example

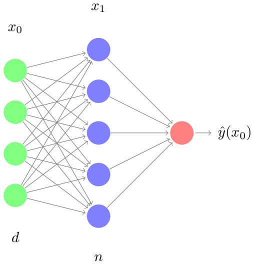


Figure: NN with 1 hidden layer.

$$\hat{y}(x_0) = \frac{1}{n}\sum_{i=1}^{d} \sigma(A_{i,j}x_{0,j}) , \quad A \in \mathbb{R}^{n\times d} ,$$

$$R_n(A) \coloneqq \mathbb{E}_{(X,Y)\sim\mu}[\ell(Y, \hat{y}(X))] .$$

A Mean Field View of the Landscape of Two-Layer Neural Networks - Mei, Montanari & Nguyen, 2018

On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport - Chizat & Bach, 2018

What we need?

A common set that contains all unlabeled graphs Embedding

What we need?

A common set that contains all unlabeled graphs	Embedding
A suitable notation of ‘graph convergence’	Topology

What we need?

A common set that contains all unlabeled graphs	Embedding
A suitable notation of 'graph convergence'	Topology
Contains all graph limits	Completion

What we need?

A common set that contains all unlabeled graphs	Embedding
A suitable notation of 'graph convergence'	Topology
Contains all graph limits	Completion
A notion of 'gradient flow' on this space	'Differentiable structure'

Kernels and Graphons

Kernels $\mathcal{W}$

A kernel is a measurable function $W\colon [0,1]^2 \to [-1,1]$ such that $W(x,y) = W(y,x)$.

Kernels and Graphons

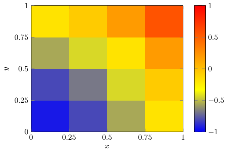
Kernels $\mathcal{W}$

A kernel is a measurable function $W : [0,1]^2 \to [-1,1]$ such that $W(x,y) = W(y,x)$.

- Symmetric matrices can be converted into a kernel.

$$\frac{1}{16}\begin{bmatrix} -16 & -15 & -12 & -14 \\ -15 & -14 & -11 & 1 \\ -12 & -11 & -6 & 4 \\ -7 & 1 & 4 & 9 \end{bmatrix}$$

Symmetric matrix A



Kernel representation of A

Kernels and Graphons

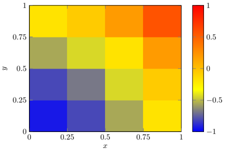
Kernels $\mathcal{W}$

A kernel is a measurable function $W : [0,1]^2 \to [-1,1]$ such that $W(x,y) = W(y,x)$.

- Symmetric matrices can be converted into a kernel.

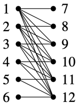
$$\frac{1}{16}\begin{bmatrix} -16 & -15 & -12 & -14 \\ -15 & -14 & -11 & 1 \\ -12 & -11 & -6 & 4 \\ -7 & 1 & 4 & 9 \end{bmatrix}$$

Symmetric matrix A



Kernel representation of A

- Therefore graphs can be made into kernel.



$$\begin{matrix} 0&0&0&0&0&0&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&1 \\ 1&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0 \\ 1&1&1&0&0&0&0&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0 \\ 1&1&1&1&1&1&0&0&0&0&0&0 \end{matrix}$$

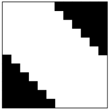
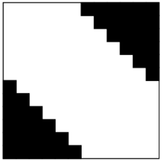
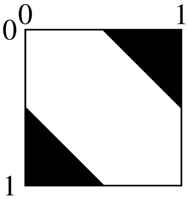


Figure: Example 4.1.6, Graph Theory and Additive Combinatorics, Yufei Zhao

Convergence of Graph(ons)



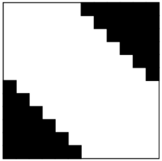
(a) Half Graph (Kernel)



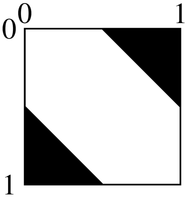
(b) Limit of Half Graph

Graph Theory and Additive Combinatorics, Yufei Zhao

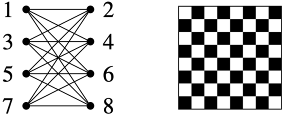
Convergence of Graph(ons)



(a) Half Graph (Kernel)



(b) Limit of Half Graph

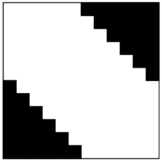


(a) Checkerboard

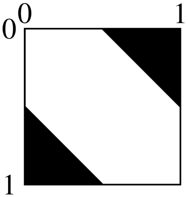
Q. Where does this sequence of kernels converge?

Graph Theory and Additive Combinatorics, Yufei Zhao

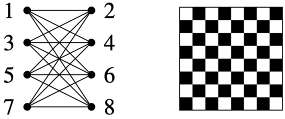
Convergence of Graph(ons)



(a) Half Graph (Kernel)

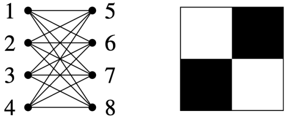


(b) Limit of Half Graph



(a) Checkerboard

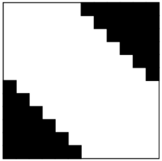
Q. Where does this sequence of kernels converge?



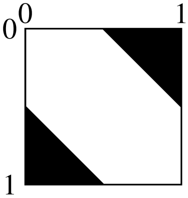
(b) Checkerboard after vertex relabeling

Graph Theory and Additive Combinatorics, Yufei Zhao

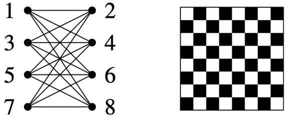
Convergence of Graph(ons)



(a) Half Graph (Kernel)

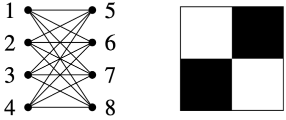


(b) Limit of Half Graph



(a) Checkerboard

Q. Where does this sequence of kernels converge?



(b) Checkerboard after vertex relabeling

A. The limit of a bipartite graph is **not** the 1/2.

Graph Theory and Additive Combinatorics, Yufei Zhao

Graphons

- We should identify two kernels if one can be obtained by 'permuting' the other.
- $W_1 \cong W_2$ if there is a measure preserving transform $\varphi\colon [0,1] \to [0,1]$ such that

$$W_1^{\varphi}(x,y) := W_1(\varphi(x), \varphi(y)) = W_2(x,y) \ .$$

Space of Graphons $\widehat{\mathcal{W}}$ (Lovász & Szegedy, 2006)

$$\widehat{\mathcal{W}} := \mathcal{W}/\cong \ .$$

A general recipe

Start with a norm $\|\cdot\|$ on $\mathcal{W}$. Define δ as

$$\delta(W_1, W_2) = \inf_{\varphi_1, \varphi_2} \|W_1^{\varphi_1} - W_2^{\varphi_2}\| \ ,$$

where $W^{\varphi}(x,y) = W(\varphi(x), \varphi(y))$.

Cut Metric: $\delta_\square$

$$\|W\|_\square := \sup_{S,T} \left| \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \right|.$$

[1] Lovász & Szegedy, 2006, using Szemerédi's regularity lemma
Frieze & Kannan, 1999

Cut Metric: $\delta_\square$

$$\|W\|_\square := \sup_{S,T} \left| \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \right|.$$

- Captures graph convergence.
  - $(G_n)_n$ converges in $\delta_\square$ if
    $$\lim_{n\to\infty} h_F(G_n)$$
    exists for all simple graphs $F \in \{-, \wedge, \triangle, \leftthreetimes, \sqcup, \square, \boxtimes, \ltimes, \bowtie, \ldots\}$.
- $(\widehat{\mathcal{W}}, \delta_\square)$ is compact.[1]

[1] Lovász & Szegedy, 2006, using Szemerédi's regularity lemma
Frieze & Kannan, 1999

Invariant L^2 metric δ_2

For $\|\cdot\| = \|\cdot\|_{L^2([0,1]^2)}$, we get the Invariant L^2 metric δ_2.

Borgs, Chayes, Lovász, Sós & Vesztergombi, 2008

Invariant L^2 metric δ_2

For $\|\cdot\| = \|\cdot\|_{L^2([0,1]^2)}$, we get the Invariant L^2 metric δ_2.

- Stronger than the cut metric (i.e., $\delta_\square \leq \delta_2$).
- Gromov-Wasserstein distance between the metric measure spaces $([0,1], \mathrm{Leb}, W_1)$ and $([0,1], \mathrm{Leb}, W_2)$.
- Provides geodesic metric structure on $\widehat{\mathcal{W}}$. Allows notion of (geodesic) convexity.

Borgs, Chayes, Lovász, Sós & Vesztergombi, 2008

What is a ‘gradient flow’?

On $\mathbb{R}^d$

The ‘gradient flow’ u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a 'gradient flow'?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a ‘gradient flow’?

On $\mathbb{R}^d$

The ‘gradient flow’ u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \,,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \,.$$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a 'gradient flow'?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \ .$$

A curve u is a gradient flow of F if

$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2.$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a 'gradient flow'?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \ .$$

A curve u is a gradient flow of F if

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2.$$

On $(\widehat{\mathcal{W}}, \delta_2)$

Consider a curve ω and a function F on $\widehat{\mathcal{W}}$.

- Speed of ω: Metric derivative $|\omega'|$

Metric Derivative of ω

$$|\omega'|(t) = \lim_{s \to t} \frac{\delta_2(\omega_t, \omega_s)}{|t-s|} \ .$$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a 'gradient flow'?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \ .$$

A curve u is a gradient flow of F if

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2.$$

On $(\widehat{\mathcal{W}}, \delta_2)$

Consider a curve ω and a function F on $\widehat{\mathcal{W}}$.

- Speed of ω: Metric derivative $|\omega'|$

Metric Derivative of ω

$$|\omega'|(t) = \lim_{s \to t} \frac{\delta_2(\omega_t, \omega_s)}{|t-s|} \ .$$

- Gradient of F: Fréchet-like derivative

Fréchet-like derivative of F: DF

Provides a local linear approximation of F.

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

What is a 'gradient flow'?

On $\mathbb{R}^d$

The 'gradient flow' u of a function $F\colon \mathbb{R}^d \to \mathbb{R}$ is given by solutions of

$$u'(t) = -\nabla F(u(t)) \ ,$$

$$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) = \langle u'(t), \nabla F(u(t)) \rangle$$

$$\geq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2 \ .$$

A curve u is a gradient flow of F if

$\frac{\mathrm{d}}{\mathrm{d}t} F(u(t)) \leq -\frac{1}{2}|u'|^2(t) - \frac{1}{2}|\nabla F(u(t))|^2.$

On $(\widehat{\mathcal{W}}, \delta_2)$

Consider a curve ω and a function F on $\widehat{\mathcal{W}}$.

- Speed of ω: Metric derivative $|\omega'|$

Metric Derivative of ω

$$|\omega'|(t) = \lim_{s \to t} \frac{\delta_2(\omega_t, \omega_s)}{|t - s|} \ .$$

- Gradient of F: Fréchet-like derivative

Fréchet-like derivative of F: DF

Provides a local linear approximation of F.

A curve u is a gradient flow of F if

$\frac{\mathrm{d}}{\mathrm{d}t} F(\omega(t)) \leq -\frac{1}{2}|\omega'|^2(t) - \frac{1}{2}|DF(\omega(t))|^2.$

Gradient Flows in Metric Spaces and in the Space of Probability Measures - Ambrosio, Gigli & Savare, 2005

Fréchet-like derivative and existence of gradient flow

Theorem [OPST '21]

If F

- has a Fréchet-like derivative,
- is geodesically semiconvex in δ_2,

then starting from any $W_0 \in \widehat{\mathcal{W}}$, the curve $(W_t)_{t\in\mathbb{R}_+}$ defined as

$$W_t := W_0 - \int_0^t DF(W_s)\,\mathrm{d}s\ ,$$

is a gradient flow of F.

Fréchet-like derivative and existence of gradient flow

Theorem [OPST '21]

If F

- has a Fréchet-like derivative,
- is geodesically semiconvex in δ_2,

then starting from any $W_0 \in \widehat{\mathcal{W}}$, the curve $(W_t)_{t\in\mathbb{R}_+}$ defined as

$$W_t := W_0 - \int_0^t DF(W_s)\,\mathrm{d}s\ ,$$

is a gradient flow of F.

- For the triangle density function $h_\triangle$,

$$(Dh_\triangle)(W)(x,y) = 3\int_0^1 W(x,z)W(z,y)\,\mathrm{d}z\ .$$

Fréchet-like derivative and existence of gradient flow

Theorem [OPST '21]

If F

- has a Fréchet-like derivative,
- is geodesically semiconvex in δ_2,

then starting from any $W_0 \in \widehat{\mathcal{W}}$, the curve $(W_t)_{t\in\mathbb{R}_+}$ defined as

$$W_t := W_0 - \int_0^t DF(W_s)\,\mathrm{d}s\,,$$

is a gradient flow of F.

- For the triangle density function $h_\triangle$,

$$(Dh_\triangle)(W)(x,y) = 3\int_0^1 W(x,z)W(z,y)\,\mathrm{d}z\,.$$

- For the scalar entropy function E, if $0 < W < 1$, then

$$(DE)(W)(x,y) = \log\left(\frac{W(x,y)}{1-W(x,y)}\right).$$

Example

- Given Dh_F and DE, we can now perform a gradient flow to minimize $h_\triangle + E$ on the space of Graphons!
- Given initial conditions, one needs to solve for all $x, y \in [0,1]$,

$$W_t'(x,y) = -\left[3\int_0^1 W(x,z)W(z,y)\,\mathrm{d}z + \log\left(\frac{W(x,y)}{1-W(x,y)}\right)\right].$$

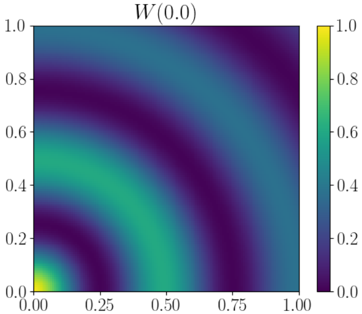


Figure: Gradient flow of $h_\triangle + 10^{-1}E$

Euclidean Gradient flow and Gradient flow on $\widehat{\mathcal{W}}$

Consider a function $F : \widehat{\mathcal{W}} \to \mathbb{R}$ that has following gradient flow

$$W(t) = W_0 - \int_0^t DF(W(s))\, \mathrm{d}s\ .$$

Euclidean Gradient flow and Gradient flow on $\widehat{\mathcal{W}}$

Consider a function $F : \widehat{\mathcal{W}} \to \mathbb{R}$ that has following gradient flow

$$W(t) = W_0 - \int_0^t DF(W(s))\,\mathrm{d}s\,.$$

- Note that the function F can be regarded as a function $F_n : \mathcal{M}_n \to R$. Suppose that F_n has a gradient flow. It is then given by

$$V^{(n)}(t) = V_0^{(n)} - \int_0^t \nabla_n F_n\Big(V^{(n)}(s)\Big)\,\mathrm{d}s\,.$$

Euclidean Gradient flow and Gradient flow on $\widehat{\mathcal{W}}$

Consider a function $F : \widehat{\mathcal{W}} \to \mathbb{R}$ that has following gradient flow

$$W(t) = W_0 - \int_0^t DF(W(s))\, \mathrm{d}s\,.$$

- Note that the function F can be regarded as a function $F_n : \mathcal{M}_n \to R$. Suppose that F_n has a gradient flow. It is then given by

$$V^{(n)}(t) = V_0^{(n)} - \int_0^t \nabla_n F_n\left(V^{(n)}(s)\right) \mathrm{d}s\,.$$

Question?

Are the curves $V^{(n)}$ and W close (if n is large)?

Euclidean Gradient and Fréchet-like derivative

Fréchet-like derivative

A symmetric measurable function $\phi \in L^\infty([0,1]^2)$ is said to be Fréchet-like derivative $DF(W)$ of F at $W \in \widehat{\mathcal{W}}$ if

$$\lim_{\substack{U \in \mathcal{W}, \\ \|U-W\|_2 \to 0}} \frac{F(U) - F(W) - \langle \phi, U - W \rangle_{L^2([0,1]^2)}}{\|U - W\|_2} = 0 \, .$$

- Recall that $F \colon \widehat{\mathcal{W}} \to \mathbb{R}$ can be regarded as a function $F_n \colon \mathcal{M}_n \to \mathbb{R}$.
- Let $\nabla_n F_n$ be Euclidean derivative of $F_n \colon \mathcal{M}_n \to \mathbb{R}$.

Euclidean Gradient and Fréchet-like derivative

Fréchet-like derivative

A symmetric measurable function $\phi \in L^\infty([0,1]^2)$ is said to be Fréchet-like derivative $DF(W)$ of F at $W \in \widehat{\mathcal{W}}$ if

$$\lim_{\substack{U \in \mathcal{W}, \\ \|U-W\|_2 \to 0}} \frac{F(U) - F(W) - \langle \phi, U - W \rangle_{L^2([0,1]^2)}}{\|U - W\|_2} = 0 \,.$$

- Recall that $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ can be regarded as a function $F_n\colon \mathcal{M}_n \to \mathbb{R}$.
- Let $\nabla_n F_n$ be Euclidean derivative of $F_n\colon \mathcal{M}_n \to \mathbb{R}$.

The graphon corresponding to $n^2 \nabla_n F_n(W)$ equals $DF(W)$.

Euclidean gradient flow and gradient flow on Graphons

Gradient flow on $\widehat{\mathcal{W}}$

$$\frac{\mathrm{d}}{\mathrm{d}t}W(t) = -DF(W(t)) = -n^2\nabla_n F(W(t))$$

Gradient flow on $\mathcal{M}_n$

$$\frac{\mathrm{d}}{\mathrm{d}t}V(t) = -\nabla_n F(V(t))$$

Euclidean gradient flow and gradient flow on Graphons

Gradient flow on $\widehat{\mathcal{W}}$

$$\frac{\mathrm{d}}{\mathrm{d}t}W(t) = -DF(W(t))$$
$$= -n^2\nabla_n F(W(t))$$

Gradient flow on $\mathcal{M}_n$

$$\frac{\mathrm{d}}{\mathrm{d}t}V(t) = -\nabla_n F(V(t))$$

- The curve $\tilde{W}(t) := V(n^2 t)$ satisfies

Euclidean gradient flow and gradient flow on Graphons

Gradient flow on $\widehat{\mathcal{W}}$

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t) = -DF(W(t))$$
$$= -n^2 \nabla_n F(W(t))$$

Gradient flow on $\mathcal{M}_n$

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t) = -\nabla_n F(V(t))$$

- The curve $\tilde{W}(t) := V(n^2 t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} \tilde{W}(t) = -n^2 \nabla_n F(\tilde{W}(t)) = -DF(\tilde{W}(t)) \,.$$

- That is, it is reasonable to expect that the gradient flow on Graphons can be obtained a scaling limit of Euclidean gradient flows.

Convergence of Euclidean Gradient Flow

Theorem [OPST '21]

- Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ be a function with gradient flow

$$W(t) := W_0 - \int_0^t D_{\widehat{\mathcal{W}}} F(W)\,\mathrm{d}s\ .$$

- Consider the Euclidean gradient flow of $F_n\colon \mathcal{M}_n \to \mathbb{R}$ starting at $V_0^{(n)}$, i.e.,

$$V^{(n)}(t) := V^{(n)}(0) - \int_0^t \nabla_n F_n\Big(V^{(n)}(s)\Big)\,\mathrm{d}s\ .$$

- Set $W^{(n)}(t) = V^{(n)}(n^2 t)$.

Convergence of Euclidean Gradient Flow

Theorem [OPST '21]

- Let $F\colon \widehat{\mathcal{W}} \to \mathbb{R}$ be a function with gradient flow

$$W(t) := W_0 - \int_0^t D_{\widehat{\mathcal{W}}} F(W)\,\mathrm{d}s\ .$$

- Consider the Euclidean gradient flow of $F_n\colon \mathcal{M}_n \to \mathbb{R}$ starting at $V_0^{(n)}$, i.e.,

$$V^{(n)}(t) := V^{(n)}(0) - \int_0^t \nabla_n F_n\Big(V^{(n)}(s)\Big)\,\mathrm{d}s\ .$$

- Set $W^{(n)}(t) = V^{(n)}(n^2 t)$.

If $W_0^{(n)} \xrightarrow{\delta_\square} W_0$, then

$$W^{(n)} \overset{\delta_\square}{\rightrightarrows} W \qquad \text{as } n \to \infty\ ,$$

over compact time intervals.

Ongoing and Future directions

- Study convergence of stochastic gradient descent with and without added noise.
- Specialize the theory on optimization over multiple layer NNs.

Thank you!

- ArXiv version: https://arxiv.org/abs/2111.09459