Lecture 6: Gradient Dynamics of Shallow Neural Networks: From lazy to mean-field regime.

- **Lecture Objectives:**
  - → highlight transition from linear to non-linear learning
  - → limitations of linear models.
  - → Mean-field description for shallow neural networks, beyond the linearised regime.

Lazy Dynamics [Chizat, Oyallon, Bach, '19]

- Let $\phi(\theta)$ be a differentiable mapping $\Theta \xrightarrow{\phi} \mathcal{F}$

  parameter space   function space.

  θ_0: initial point of the learning algorithm.

- Linearised Tangent Model at initialisation:

$$\bar{\phi}(\theta) = \phi(\theta_0) + D\phi(\theta_0)\cdot(\theta - \theta_0)$$

fixed family of features

↳ $$\bar{\phi}(\theta)(x) = \phi(\theta_0; x) + \sum_j (\theta_j - (\theta_0)_j)\cdot \underbrace{\nabla_{\theta_j}\phi(\theta_0; x)}$$

→ observe that $\bar{\phi}$ is linear (or affine) w.r.t. θ (but generally non-linear wrt x!)

- Q: When is gradient-descent learning under these two models $(\phi, \bar{\phi})$ similar?

- Let $E(\theta) = L(\phi(\theta))$; e.g. $E(\theta) = \| \phi(\theta) - f^* \|^2$

  and assume θ_0 s.t. $E(\theta_0) > 0$.

- Consider a gradient step: $\theta_1 = \theta_0 - \eta \nabla E(\theta_0)$.

  ↳ Relative change in objective function:

  $\Delta_E = \frac{E(\theta_1) - E(\theta_0)}{\ldots} \approx \frac{\langle \nabla E(\theta_0), \theta_1 - \theta_0 \rangle}{\ldots} = \frac{\eta \|\nabla E(\theta_0)\|^2}{\ldots}$

$$\Delta E = \frac{E(\theta_0) - E(\theta_1)}{E(\theta_0)} \approx \frac{\langle \nabla E(\theta_0), \theta_1 - \theta_0 \rangle}{E(\theta_0)} = \frac{\eta \|\nabla E(\theta_0)\|^2}{E(\theta_0)}.$$

↳ Relative change in tangent model:

$$\Delta(D\phi) = \frac{\| D\phi(\theta_1) - D\phi(\theta_0) \|}{\| D\phi(\theta_0) \|} \overset{\text{1st order Taylor: } |f(a)-f(b)| \leq \sup|f'(t)||a-b|}{\leq} \frac{\eta \, \| D^2\phi(\theta_0) \| \, \| \nabla E(\theta_0) \|}{\| D\phi(\theta_0) \|}$$

→ We say that a differentiable model operates in the "lazy" regime when

$$(*) \quad \boxed{\Delta(D\phi) \ll \Delta(E)}$$

ie tangent model evolves much slower than the loss.

→ If we consider $L(f) = \| f - f^\circ \|^2$, then we verify that $(*)$ is equivalent to

$$\boxed{\underbrace{\| \phi(\theta) - f^\circ \| \frac{\| D^2\phi(\theta) \|}{\| D\phi(\theta) \|^2}}_{K_\phi(\theta) \text{ : relative scale of } \phi \text{ at } \theta} \ll 1}$$

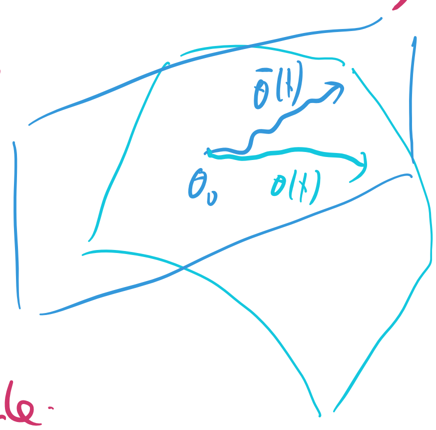
Theorem [COB, Thm 2.3 (Simplified)]: Assume ϕ, $D\phi$ are both Lipschitz in a neighborhood of θ_0. Let

$\theta(t)$: the (non-linear) Gradient flow that solves $\dot{\theta}(t) = -\nabla L(\phi(\theta(t)))$ [Non-linear]

$\bar{\theta}(t)$: the linearised gradient flow that solves $\dot{\bar{\theta}}(t) = -\nabla L(\bar{\phi}(\bar{\theta}(t)))$

Then $\exists \; C_\phi$ such that for $t \leq C_\phi$ it holds

$$\frac{\| \phi(\theta(t)) - \bar{\phi}(\bar{\theta}(t)) \|}{\| \phi(\theta_0) - f^\circ \|} \lesssim t^2 \underbrace{K_\phi(\theta_0)}_{\text{relative scale}}$$



Remark: Here we present a finite-time result, but under further assumptions, it can be extended to infinite-time.

Q: When does lazy regime happen?

↳ Scaled model: $\phi_\alpha(\theta) = \alpha \cdot \phi(\alpha)$

$$K_{\phi_\alpha}(\theta_0) = \frac{1}{\alpha} \| \alpha \phi(\theta_0) - f^* \| \cdot \frac{\|D^2 \phi(\theta_0)\|}{\|D\phi(\theta_0)\|^2}$$

$$\| \phi(\theta) - f^* \| \frac{\|D^2 \phi(\theta_0)\|}{\|D\phi(\theta_0)\|^2}$$

when $\phi(\theta_0) = 0$ (centering condition).

then we have $K_{\phi_\alpha}(\theta_0) = \frac{K_\phi(\theta_0)}{\alpha} \to 0$ as $\alpha \to \infty$.

↳ Homogeneous Model: $\phi(\lambda\theta) = \lambda^r \phi(\theta) \quad \forall \theta, \lambda > 0$.

Same as before $K_\phi(\lambda\theta_0) = \frac{1}{\lambda^r} K_\phi(\theta_0)$ (for centered init).

↳ Single hidden-layer NN: $\phi_m(\theta) = \alpha(m) \cdot \sum_{j=1}^{m} g(\theta_j)$, $\theta_j \sim \mu$ iid.

(ex, $g(\theta; x) = c \cdot \mathrm{Relu}(\langle x, a \rangle + b)$)

with $\mathbb{E}_\mu\, g(\theta) = 0$, and Dg is Lipschitz.

We want: $K_m = \mathbb{E}_\mu [K_{\phi_m}(\theta)]$.

↳ $\mathbb{E} \| \phi_m(\theta) \|^2 = m\, \alpha(m)^2\, \mathbb{E} \|g(\theta)\|^2$ (since θ_j are iid)

↳ $D\phi_m(\theta) = \alpha(m) [Dg(\theta_1), \dots, Dg(\theta_m)]$

$$\frac{D\phi_m(\theta) \cdot D\phi_m(\theta)^T}{m \cdot \alpha(m)^2} = \frac{1}{m} \sum_{j=1}^{m} Dg(\theta_j) \cdot Dg(\theta_j)^T$$

$$\xrightarrow[m\to\infty]{} \mathbb{E}\left[Dg(\theta) \cdot Dg(\theta)^T \right] \quad \text{(Law of Large Numbers)}$$

$$\Rightarrow \mathbb{E} \| D\phi_m(\theta)\|^2 = \mathbb{E} \| D\phi_m(\theta) \cdot D\phi_m(\theta)^T \|$$

$$\simeq m \cdot \alpha(m)^2 \, \underline{\mathbb{E} \| Dg(\theta)\|^2}$$

$$\hookrightarrow \| D^2 \phi_m(\theta) \| = \sup_{\substack{\|u\| \leq 1 \\ u \in \mathbb{R}^{d \times m}}} \alpha(m) \sum_{j=1}^{m} u_j^T D^2 g(\theta_j) u_j$$

$\sup \sum \leq \sum \sup$

$$\leq \alpha(m) \sup_{\theta_i} \| D^2 g(\theta_i) \| \overset{*}{\leq} \alpha(m) \cdot Lip(Dg)$$

$\hookrightarrow \sup_x \|\nabla f(x)\| \leq Lip(f)$.

$\hookrightarrow$ From triangle ineq: $\| \phi_m(\theta) - f^\circ \| \leq \| f^\circ \| + \underline{\| \phi_m(\theta) \|}$

We put everything together:

$$K_m \lesssim \left(C_1 + \sqrt{m} \cdot \alpha(m) \cdot C_2 \right) \cdot \frac{\alpha(m) \cdot C_3}{m \cdot \alpha(m)^2 \cdot C_4}$$

$$\boxed{K_m \lesssim \frac{\tilde{C}_1}{\sqrt{m}} + \frac{\tilde{C}_2}{m \cdot \alpha(m)}}$$

Conclusion: If $\underline{m \cdot \alpha(m) \to \infty}$ as m grows, then this model will become "lazy" in the overparametrized (m large) regime!

In particular, NTK considers $\alpha(m) = \frac{1}{\sqrt{m}}$, so $m\,\alpha(m) = \sqrt{m}$!

$\oplus$ It will guarantee global convergence; optimisation is easy!

$\ominus$ We are in essence giving up on the non-linear nature of the model: we are learning only a linear combination of fixed feature maps. ~~Ms~~

↳ No "representation" learning

↳ Associated functional space is the RKHS with Kernel generated by the NTK (features given by $\nabla\phi(\theta_0)$)

↳ even a single neuron $f^\circ(x) = g(\theta^\circ; x)$ is NOT in the RKHS [Bach '15].

Q: What happens when $m \cdot \alpha(m) = \Theta(1)$, ie $\alpha(m) = 1/m$?

Shallow Neural Networks and Particle Interaction Systems

→ Recall our shallow model $\phi(\theta_1 \cdots \theta_m; x) = \frac{1}{m}\sum_{j=1}^{m} g(\theta_j; x)$

$g(\theta; x) = c \cdot \sigma(\langle x, a\rangle + b)$ $\quad \theta = (a; b; c) \in \underbrace{\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}}_{\mathbb{R}^{d+2}} =: D$

↳ sigmoid, Relu, etc.

↳ output weight

↳ input weight

↳ bias

→ Consider a least-squares regression:

$$\min_{\substack{\theta_1 \cdots \theta_m \\ \vec{\theta}}} \mathcal{L}(\vec{\theta}) = \|\phi(\vec{\theta}) - f^\circ\|_\nu^2 + \lambda \cdot V(\vec{\theta}), \quad V(\vec{\theta}) = \frac{1}{m}\sum_j v(\theta_j)$$

$$\|f\|_\nu^2 = \mathbb{E}_{x\sim\nu}[f(x)]^2$$

→ By developing the square, we get

$$\mathcal{L}(\vec{\theta}) = \|f^\circ\|^2 - 2\langle \phi(\vec{\theta}), f^\circ\rangle + \|\phi(\vec{\theta})\|^2 + \lambda V(\vec{\theta})$$

→ Introduce the functions

$$\langle g, f\rangle := \mathbb{E}_\nu [f(x) \cdot g(x)]$$

$$\left(= \frac{1}{n}\sum_{i=1}^{n} f(x_i) g(x_i) \quad \text{empirical loss} \right)$$

$$F : D \to \mathbb{R}$$

$$\theta \longmapsto F(\theta) = \langle g(\theta), f^\circ\rangle - \frac{1}{2}\lambda \cdot v(\theta)$$

$$K : D \times D \to \mathbb{R}$$

$$(\theta, \theta') \longmapsto K(\theta, \theta') = \langle g(\theta), g(\theta') \rangle$$

↳ observe that K is symmetric and psd operator.

→ The loss becomes

$$\mathcal{L}(\theta_1 \cdots \theta_m) = \underbrace{C}_{\text{irrelevant.}} - \frac{2}{m} \langle \sum_{j=1}^{m} g(\theta_j), f^* \rangle + \| \sum_{j=1}^{m} g(\theta_j) \|^2 + \frac{\lambda}{m} \sum_{j=1}^{m} v(\theta_j)$$

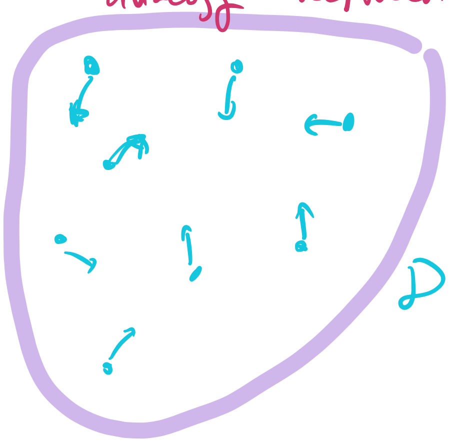
$$= C - \frac{2}{m} \sum_{j=1}^{m} F(\theta_j) + \frac{1}{m^2} \sum_{j,j'=1}^{m} K(\theta_j, \theta_{j'})$$

We have written an energy of a system of m interacting particles.

↳ external field/force.

↳ interaction kernel.

analogy between neurons ⟷ particles.



→ Scaled gradient Flow wrt $\theta_1 \cdots \theta_m$ gives the associated Lagrangian dynamics:

$$\dot{\theta}_j = -\frac{m}{2} \nabla_{\theta_j} \mathcal{L}(\theta_1 \cdots \theta_m) \qquad j = 1 \ldots m.$$

$$= + \nabla F(\theta_j) - \frac{1}{m} \sum_{j'=1}^{m} \nabla K(\theta_j, \theta_{j'})$$

↳ This Lagrangian description is increasingly complex (defined over $D^{\otimes m}$)

↳ It is exact at particle level — but the resulting optimisation landscape is very complex. (it has bad local minima).

↳ Can we instead obtain a simpler collective behavior?

→ Let's instead consider the Eulerian perspective

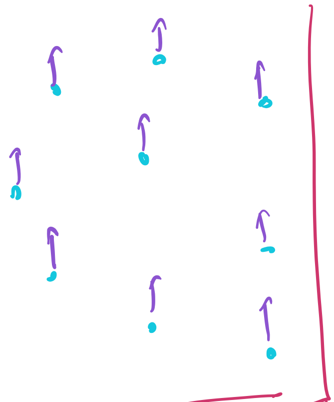
$\vec{\theta} = (\theta_1, \dots, \theta_m) \in D^{(\times m)} \longleftrightarrow \hat{\mu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{\theta_j} \in P(D)$

↳ space of probability measures over D.

Lagrangian view ⟵⟶ Eulerian view

D^m Finite-dimensional Euclidean space

Infinite-Dimensional Non-Euclidean.



$\mu = \delta_{\theta_0}$

$\int \chi(\theta)\, d\mu(\theta) = \chi(\theta_0)$

→ The model $\phi(\vec{\theta}; x) = \frac{1}{m}\sum_{j=1}^{m} g(\theta_j; x)$ becomes

$$\phi(\vec{\theta}; x) \stackrel{!}{=} \int_D g(\theta; x)\, \hat{\mu}_m(d\theta)$$

non linear dependency between ϕ and $\vec{\theta}$ ←

↳ ϕ is linear w.r.t. $\hat{\mu}_m$!!

→ The energy of the system expressed in terms of μ is

$\phi(x) = \int g(\theta, x)\, \mu(d\theta)$

$\mu_1, \mu_2 \; \mu = \frac{1}{2}(\mu_1 + \mu_2)$

$$- \frac{2}{m}\sum_{j=1}^{m} F(\theta_j) + \frac{1}{m^2}\sum_{j,j'=1}^{m} K(\theta_j, \theta_{j'})$$

$$\mathcal{L}[\mu] = -2\int_D F(\theta)\,\mu(d\theta) + \int_{D\times D} K(\theta, \theta')\,\mu(d\theta)\,\mu(d\theta')$$

$\phi(x) = \frac{1}{2}(\phi_1(x) + \phi_2(x))$

$\mathcal{L}[\mu]$ is now quadratic w.r.t. μ; in fact it is convex w.r.t. μ (recall that K is psd).

Q: Too good to be true?

A: Not so fast: $\mathcal{L}$ is convex with respect to the geometry of linear mixtures.

However, gradient descent dynamics correspond to a different geometry.

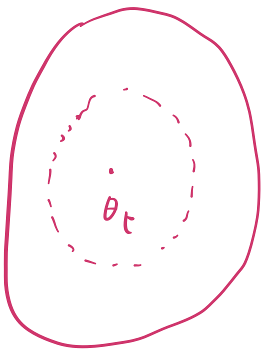
Q: What is the relationship between GD/GF and metric?

Proximal Viewpoint of GD

The standard (Euclidean) GD step satisfies

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \underbrace{f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle}_{\text{linear approx of } f \text{ at } \theta_t} + \underbrace{\frac{1}{2\eta} \|\theta - \theta_t\|_2^2}_{\text{proximity term.}} \right\}$$



- How do we measure proximity? L_2 metric is one choice, but not always the "good" choice!
- The generalisation of GD on more general metrics is generally called <u>Mirror Descent</u> [Nemirovski, Yudin '83].

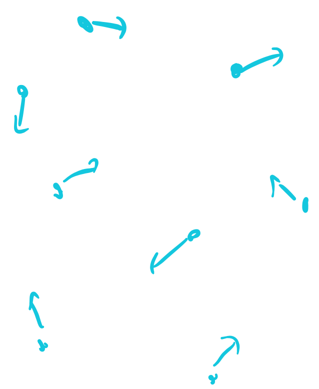
↳ In essence, $\theta_{t+1} = \underset{\theta}{\operatorname{argmin}} \left\{ f(\theta) + \frac{1}{2\eta} \underbrace{D(\theta; \theta_t)}_{\text{Divergence (Bregman) function measures proximity}} \right\}$

between θ_t and θ.

· Back at our measure space:

$$\begin{tikzcd}[column sep=huge]
\underline{\text{Lagrangian View}} & \underline{\text{Eulerian View}} \\
\vec{\theta}_t \arrow[d, "\text{Euclidean GD}"'] \arrow[r, leftrightarrow] & \mu_t = \frac{1}{m}\sum_{j=1}^{m} \delta_{\theta_j(t)} \arrow[d, bend left, "?"'] \\
\vec{\theta}_{t+1} \arrow[r, leftrightarrow] & \mu_{t+1}
\end{tikzcd}$$

what kind of descent does this update correspond to?



· Recall that $\dot{\theta}_j = +\nabla F(\theta_j) - \overbrace{\frac{1}{m}\sum_{j'=1}^{m} \nabla K(\theta_j, \theta_{j'})}$

· Using $\mu_t = \frac{1}{m}\sum_j \delta_{\theta_j(t)}$, we obtain

$$\boxed{\dot{\theta}_j(t) = \nabla F(\theta_j(t)) - \int_D \nabla K(\theta_j(t), \theta)\, \mu_t(d\theta)}$$

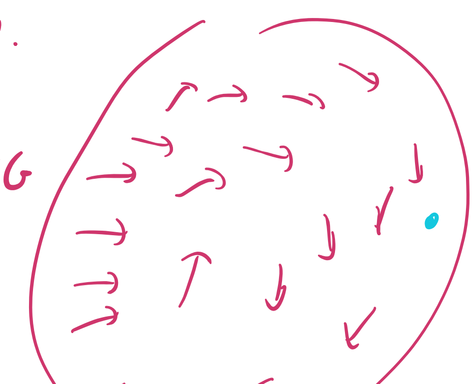
$$= \boxed{-\nabla G(\theta_j(t); \mu_t)}, \quad \text{with}$$

$$\underset{\theta \in D}{G(\theta; \mu)} = -F(\theta) + \int_D K(\theta, \theta')\, \mu(d\theta')$$

In physics, G is called the <u>instantaneous potential</u> of the system. Its gradient $\nabla G(\theta; \mu_t)$ defines a velocity field over D that any particle "feels".

→ Consider now a test function (smooth) $\chi: D \to \mathbb{R}$ and

∇G



$$\int_{\mathcal{D}} \chi(\theta)\, \mu_t(d\theta) = \frac{1}{m} \sum_{j=1} \chi(\theta_j(t))$$

$\downarrow$ Time derivatives.

$$\int_{\mathcal{D}} \chi(\theta)\, \partial_t \mu_t(d\theta) = \frac{1}{m} \sum_{j=1}^{m} \nabla \chi(\theta_j(t)) \cdot \dot{\theta}_j(t)$$

$$= -\frac{1}{m} \sum_{j=1}^{m} \left\langle \nabla \chi(\theta_j(t)), \nabla G(\theta_j(t); \mu_t) \right\rangle$$

$$= - \int_{\mathcal{D}} \left\langle \underline{\nabla \chi(\theta)}, \nabla G(\theta; \mu_t) \right\rangle \mu_t(d\theta)$$

→ This is known as a continuity/transport equation. (Mass is conserved).

→ The associated PDE is written as

$$\partial_t \mu_t = \operatorname{div}\left(\nabla G(\theta; \mu_t)\, \mu_t \right) \quad \boxed{\text{Liouville equation}}.$$

Q: Do these dynamics correspond also to a gradient flow?

A: Yes! The proximal interpretation is given in terms of a so-called Wasserstein Gradient Flow:

$$\mu_{t+1} = \underset{\mu \in P(\mathcal{D})}{\operatorname{argmin}} \left\{ \mathcal{L}[\mu] + \frac{1}{2\eta} W_2^2(\mu, \mu_t) \right\}$$

[JKO Scheme].
Jordan, Klein, Otto

Remarks:

→ This description is exact.

→ Add noise to the gradient updates (GF → Langevin dynam) translates into another PDE in the space of measures (Liouville eq → McKean-Vlasov eq).

This connection between Wasserstein gradient flow and training of shallow NN was made by [Chizat, Bach] [Rotskoff & Vanden-Eijnden] [Mei, Montanari, Nguyen] [Sirignano & Spiliopoulos] - all in 2018..

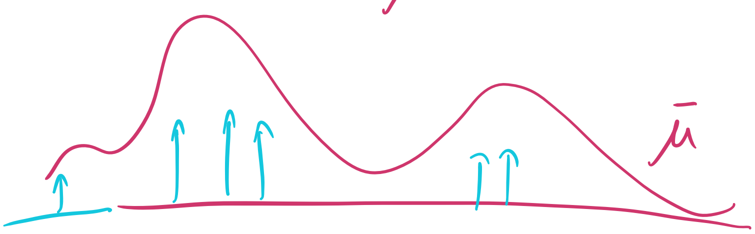
Mean-field regime

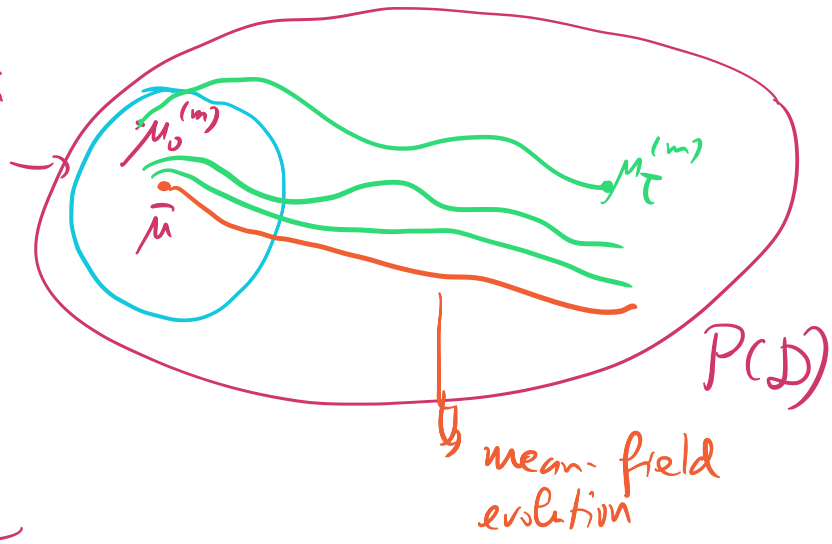
- Consider the evolution of the system as $m \to \infty$.

$\mu_t^{(m)}$: state after time t, with $\theta_i(0) \underset{iid}{\sim} \bar{\mu}$

$\mu_0^{(m)} = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j(0)}$, $\theta_j(0) \sim \bar{\mu}$

is the empirical measure associated with $\bar{\mu}$.





μ_t solves

$\partial \mu_t = \mathrm{div}(\nabla G(\theta; \mu_t)\mu_t)$

with initial condition

$\mu_0 = \bar{\mu}$

Theorem [CB, REVE, MMN, SS]

For any fixed $T > 0$, $\mu_T^{(m)}$ converges weakly to μ_T as $T \to \infty$, where μ_t solves.

In other words:

Dynamics and sampling commute in the limit of $m \to \infty$.

Two main questions

(i) Under what conditions does this PDE converge to the global minimum of $\mathcal{L}$? (convergence in time).

(ii) How are the dynamics affected in terms of overparam. (convergence/fluctuations in m).

[Existing positive results for global convergence, but they are all qualitative in m and also in t (no rates!).