Effective models in high dimensional unstable regimes

Some examples from machine learning and porous media

Antonin Chodron de Courcel

Ecole Normale Supérieure

June 2026

High dimension, instability, correlations

Kinetic theory

- $N \approx 10^{11}$ stars, $6 \times N$ variables
- Microscopic instabilities creating tractable correlations

Learning

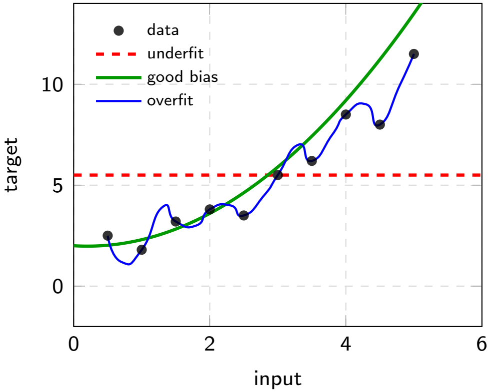
- high-dimensional correlated data and architectures
- unstable training procedures

What is the effect of the underlying instabilities on the effective model?

Overparametrisation and implicit regularization

$\#\text{parameters} \gg \#\text{data}$

- Easy interpolation of the **training set**
- **Generalization:** good prediction on **unseen data**
- Many interpolating predictors
- Why this one rather than another?



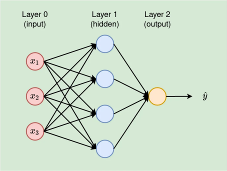
We need a **bias / regularization** selecting the right interpolant.
This is the mechanism we want to capture in the effective dynamics.

Wide two-layer networks: the model

- **One-hidden-layer model:**

$$f_N(x) = \frac{1}{N}\sum_{i=1}^{N} \phi(x,\theta_i) = \int \phi(x,\theta)\, d\rho_N(\theta), \qquad \rho_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta_i}$$

- θ_i: parameters of one neuron
- ρ_N: empirical distribution of parameters
- $\phi(x,\theta)$: response of one neuron to the input x



- **Training functional:**

$$\mathcal{F}(\rho) = \mathbb{E}_{(x,y)}\left[\ell(y, f_\rho(x))\right]$$

$$f_\rho(x) = \int \phi(x,\theta)\, d\rho(\theta)$$

$\ell(y, f_\rho(x))$ = prediction error against the label y

Overparametrization becomes a problem on probability measures.

Wide two-layer networks: the mean-field picture

- **Gradient descent on the neurons:**

$$\theta_i^{(k+1)} = \theta_i^{(k)} - \eta N \nabla_{\theta_i} \mathcal{F}(\rho_N^{(k)})$$

- **Small learning rate, many neurons:**

$$\eta \ll 1, \qquad t = k\eta, \qquad N \to \infty \qquad \Longrightarrow \qquad \partial_t \rho - \mathrm{div}_\theta \left(\rho \nabla_\theta \frac{\partial \mathcal{F}}{\partial \rho}(\rho) \right) = 0$$

- **Specific choice for $\mathcal{F}$:**

$$\mathcal{F}(\rho) = \frac{1}{2} \int_\Omega \mathrm{g} * (\rho - \mu)\, d(\rho - \mu) \qquad \text{Maximum mean discrepancy}$$

Small η: mean-field / Wasserstein gradient flow
Large η: unstable dynamics $\Rightarrow$ modified effective model

The stable case: small learning rate

collaboration with Matthew Rosenzweig

The framework: a mean-field equation

Consider the small learning-rate regime $\eta \to 0$.

- $t \geq 0$: training time, ρ: parameter density, $\mu \in \mathcal{P}(\mathbb{T}^d)$: target distribution.

$$\begin{cases} \partial_t \rho - \mathrm{div}(\rho \nabla \mathrm{g} * (\rho - \mu)) = 0, \\ \rho\big|_{t=0} = \rho_0. \end{cases} \tag{1}$$

- This is the Wasserstein gradient flow of the Coulomb MMD:

$$\mathrm{MMD}^2(\rho, \mu) = \frac{1}{2} \int_{\mathbb{T}^d} \mathrm{g} * (\rho - \mu) \, d(\rho - \mu).$$

The framework: a mean-field equation

Consider the small learning-rate regime $\eta \to 0$.

- $t \geq 0$: training time, ρ: parameter density, $\mu \in \mathcal{P}(\mathbb{T}^d)$: target distribution.

$$\begin{cases} \partial_t \rho - \mathrm{div}(\rho \nabla \mathrm{g} * (\rho - \mu)) = 0, \\ \rho|_{t=0} = \rho_0. \end{cases} \tag{1}$$

- This is the Wasserstein gradient flow of the Coulomb MMD:

$$\mathrm{MMD}^2(\rho, \mu) = \frac{1}{2} \int_{\mathbb{T}^d} \mathrm{g} * (\rho - \mu)\, d(\rho - \mu).$$

- g: the Coulomb kernel

$$-\Delta \mathrm{g} = \delta_0 - 1.$$

- **Stable regime:** instability is encoded in the singular kernel but does not modify the effective model.

Well-posedness and asymptotic behavior

- **Trade-off:** large-time behavior and stability of the flow.
- **Previous results:**[1] global well-posedness in Hölder spaces and convergence if

$$\inf_{\mathbb{T}^d} \rho_0 > 0, \qquad \inf_{\mathbb{T}^d} \mu > 0.$$

Well-posedness and asymptotic behavior

- **Trade-off:** large-time behavior and stability of the flow.
- **Previous results:**[1] global well-posedness in Hölder spaces and convergence if

$$\inf_{\mathbb{T}^d} \rho_0 > 0, \qquad \inf_{\mathbb{T}^d} \mu > 0.$$

Theorem (ACDC, M.Rosenzweig)

Let $\rho_0 \in \mathcal{P}(\mathbb{T}^d)$ and $\mu \in \mathcal{P} \cap L^\infty(\mathbb{T}^d)$ with

$$\lambda := \inf_{\mathbb{T}^d} \mu > 0.$$

Then the Coulomb MMD flow admits weak solutions that satisfy $\rho_t \in L^\infty(\mathbb{T}^d)$ for all $t > 0$ and

$$\mathrm{MMD}^2(\rho_t, \mu) \leq C_s e^{-2\lambda(t-s)} \, \mathrm{MMD}^2(\rho_s, \mu) \qquad (t \geq s > 0).$$

Well-posedness and asymptotic behavior

- **Trade-off:** large-time behavior and stability of the flow.
- **Previous results:**[1] global well-posedness in Hölder spaces and convergence if

$$\inf_{\mathbb{T}^d} \rho_0 > 0, \qquad \inf_{\mathbb{T}^d} \mu > 0.$$

Theorem (ACDC, M.Rosenzweig)

Let $\rho_0 \in \mathcal{P}(\mathbb{T}^d)$ and $\mu \in \mathcal{P} \cap L^\infty(\mathbb{T}^d)$ with

$$\lambda := \inf_{\mathbb{T}^d} \mu > 0.$$

Then the Coulomb MMD flow admits weak solutions that satisfy $\rho_t \in L^\infty(\mathbb{T}^d)$ for all $t > 0$ and

$$\mathrm{MMD}^2(\rho_t, \mu) \leq C_s e^{-2\lambda(t-s)} \, \mathrm{MMD}^2(\rho_s, \mu) \qquad (t \geq s > 0).$$

Use of a *defect* Polyak-Lojasiewicz inequality.

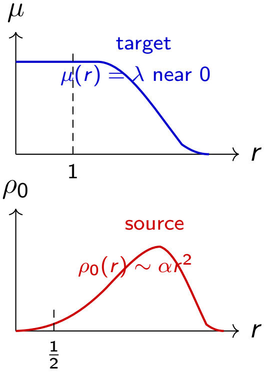
Regularity can deteriorate

- Hölder regularity is propagated, but the norm may grow.

Theorem (ACDC, M.Rosenzweig)

There exist smooth compactly supported radial densities ρ_0, μ on $\mathbb{R}^d$ such that, for every $\beta \in (0,1]$ and $t \geq 0$,

$$[\rho_t]_{C^{0,\beta}} \geq c_\beta e^{\beta\lambda \frac{d+2}{2d} t}, \qquad \lambda = \mu(0).$$



The unstable case: large learning rate

Stability condition for explicit Euler discretization

Gradient descent:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla E(\theta^{(k)}).$$

Explicit Euler scheme for the **gradient flow** at time $t = k\eta$:

$$\frac{d\theta}{dt} = -\nabla E(\theta), \qquad \theta(k\eta) \approx \theta^{(k)} \quad \text{as } \eta \ll 1.$$

Stability condition for explicit Euler discretization

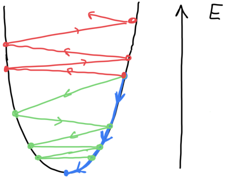
Gradient descent:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla E(\theta^{(k)}).$$

Explicit Euler scheme for the **gradient flow** at time $t = k\eta$:

$$\frac{d\theta}{dt} = -\nabla E(\theta), \qquad \theta(k\eta) \approx \theta^{(k)} \quad \text{as } \eta \ll 1.$$

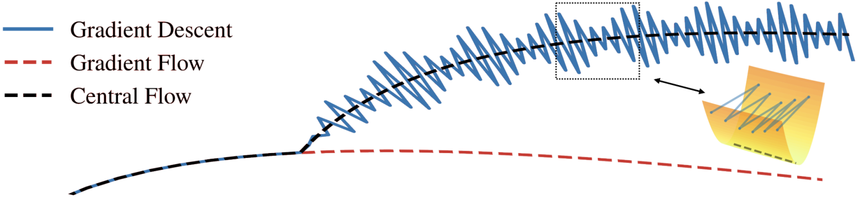
- **Issue:** large $\eta \implies$ instability.
- **Key quantities:** eigenvalues λ_i of $\nabla^2 E(\theta)$.



$$\eta\lambda_i \in (2, \infty) \Leftrightarrow \text{unstable oscillations along eigenvector of } \lambda_i,$$
$$\eta\lambda_i \in (1, 2) \Leftrightarrow \text{stable oscillations along eigenvector of } \lambda_i.$$

Related work

- **Goal:** modify the effective model.
- Cohen, Damian et al.: **Central Flow.**[2]



The central flow (black) models the time-averaged trajectory of gradient descent (blue).

[2] Jeremy Cohen et al. "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=jh-rTtvkGeM; Jeremy Cohen et al. "Understanding Optimization in Deep Learning with Central Flows". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=sIE2rI3ZPs.

Related work

Idea: ansatz $\theta^{(k)} = \bar{\theta}^{(k)} + \sqrt{\eta}\delta\theta^{(k)}$.

$$\frac{d\bar{\theta}}{dt} = -\nabla E(\bar{\theta}) - \frac{1}{2}\eta\nabla\nabla^2 E(\bar{\theta}) : \Sigma, \qquad \Sigma = \mathbb{E}(\delta\theta\delta\theta^T), \qquad \bar{\theta}(k\eta) \approx \theta^{(k)}.$$

- Stabilization : escape "too unstable" minima.
- **Central Flow:** structural assumptions on the covariance matrix Σ: **fast relaxation of oscillations**

$$\mathrm{Im}\,\Sigma \subset \ker(2/\eta - \nabla^2 E(\bar{\theta})).$$

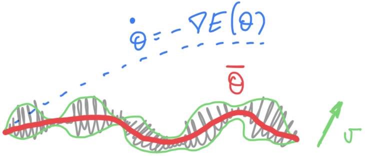
- Strength: strong empirical evidence.
- Limitation: strong assumptions on oscillations; hard to analyse rigorously.

A second-order model[3]

Unpredictable oscillations given $\bar{\theta}$: **second-order** model.

$$\frac{d\bar{\theta}}{dt} = -\eta \nabla E(\bar{\theta}) - \frac{1}{2}\eta^2 \nabla\nabla^2 E(\bar{\theta}) : \Sigma,$$
$$\frac{d\Sigma}{dt} = -\eta(\nabla^2 E(\bar{\theta})\Sigma + \Sigma\nabla^2 E(\bar{\theta})) + \eta^2 \nabla^2 E(\bar{\theta})\Sigma\nabla^2 E(\bar{\theta}).$$

- Σ: covariance envelope of the fast oscillations.
- Same macroscopic time scale for Σ and $\bar{\theta}$.
- No hard-coded relaxation constraint on Σ.



[3] Antonin Chodron de Courcel. *Gradient descent at the Edge of Stability: free energy model and kinetic description of the two-layer network*. 2026. arXiv: 2606.05326 [cs.LG]. URL: https://arxiv.org/abs/2606.05326.

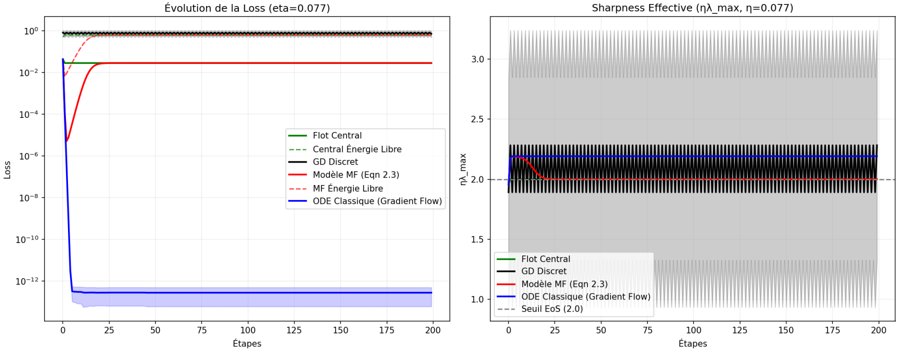
Free energy and low-curvature regime

$$H = \nabla^2 E(\bar\theta), \qquad F(\bar\theta, \Sigma) = E(\bar\theta) + \frac{\eta}{2} H : \Sigma.$$

- If $\|H\|_{L^\infty} \le 2/\eta$, then
$$\frac{d}{dt} F(\bar\theta_t, \Sigma_t) = -\eta \|\nabla_{\bar\theta} F(\bar\theta_t, \Sigma_t)\|^2 \le 0.$$
- Persistent oscillations are critical points of F:
$$\nabla_{\bar\theta} F = 0, \qquad H\Sigma + \Sigma H - \eta H \Sigma H = 0.$$
- **Low-curvature / flat landscape:** if $E \mapsto \lambda E$ with $\lambda \ll 1$ and $\tau = t/(\eta\lambda)$,
$$\frac{d\Sigma}{d\tau} = -(\Sigma H + H\Sigma).$$
- Rank is preserved, hence $\Sigma = v \otimes v$ and $\dfrac{dv}{d\tau} = -Hv$.

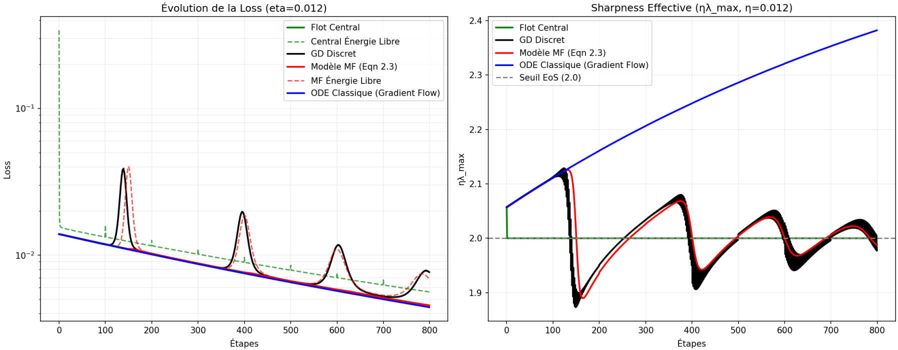
Numerical simulations

Matrix factorization



Matrix factorization, $\eta = 0.077$, 10 seeds, 200 steps.

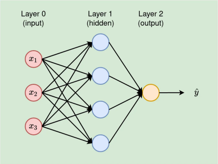
Numerical simulations
MLP on CIFAR-10



MLP on CIFAR-10, $n = 500$, MSE loss, $\eta = 0.012$, 800 steps.

Wide networks as mean-field models

- **One-hidden-layer model:**



$$f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \phi(x, \theta_i) = \int \phi(x, \theta)\, d\rho_N(\theta).$$

- **Gradient-flow training[4]:**

$$\mathcal{F}(\rho) = \mathbb{E}\left[\ell\left(y, \int \phi\, d\rho\right)\right], \quad \frac{\partial \rho}{\partial t} - \mathrm{div}_\theta \left(\rho \nabla_\theta \frac{\partial \mathcal{F}}{\partial \rho}(\rho)\right) = 0.$$

- **Special case:**

$$\mathcal{F}(\rho) = \frac{1}{2} \int \mathrm{g} * (\rho - \mu)\, d(\rho - \mu), \quad \frac{\partial \rho}{\partial t} - \mathrm{div}_\theta(\rho \nabla \mathrm{g} * (\rho - \mu)) = 0.$$

[4] Lénaïc Chizat and Francis Bach. "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport". In: *Advances in Neural Information Processing Systems 31 (NeurIPS)*. 2018, pp. 3040–3050. URL: http://papers.nips.cc/paper/7567-on-the-global-convergence-of-gradient-descent-for-over-parameterized-models-using-optimal-transport; Song Mei, Andrea Montanari, and Phan-Minh Nguyen. "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), E7665–E7671. DOI: 10.1073/pnas.1806579115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1806579115.

Connection with our model

$\bar\theta = (\bar\theta_1, \dots, \bar\theta_N) \in (\mathbb{R}^d)^N$ and $\Sigma = (\Sigma_{i,j})_{1\le i,j\le N} \in (\mathbb{R}^{d\times d})^{N\times N}$.

$$\frac{d\bar\theta_i}{dt} = -N\nabla_{\bar\theta_i}\mathcal{F}(\rho_N) - \frac{1}{2}\eta N \nabla_{\bar\theta_i}\nabla^2\mathcal{F}(\rho_N) : \Sigma,$$
$$\frac{d\Sigma}{dt} = -N\nabla^2\mathcal{F}(\rho_N)\Sigma - N\Sigma\nabla^2\mathcal{F}(\rho_N) + \underbrace{\eta N\nabla^2\mathcal{F}(\rho_N)\Sigma\nabla^2\mathcal{F}(\rho_N)}_{O(1/N)}.$$

Equation for $f(\bar\theta, v)$:

$$\frac{\partial f}{\partial t} - \mathrm{div}_{\bar\theta}(f\nabla \mathrm{g} * (\rho - \mu)) - \frac{1}{2}\eta\, \mathrm{div}_{\bar\theta}(f\nabla_{\bar\theta}(\nabla^2 \mathrm{g} : v\otimes v) * (f-\mu))$$
$$- \mathrm{div}_v(f(\nabla^2 \mathrm{g} v) * (f-\mu)) = 0,$$

where $\rho := \int f\, dv$ and $v \in \mathbb{R}^{d\times r}$ represents the r oscillation directions[5].

In the mean-field limit, the quadratic term in Σ disappears: we recover precisely the low-curvature / Bures regime.

[5] assumed $O(1)$.

Connection with our model

$$\frac{\partial f}{\partial t} - \mathrm{div}_{\bar\theta}(f \nabla \mathrm{g} * (\rho - \mu)) - \frac{1}{2}\eta\, \mathrm{div}_{\bar\theta}(f \nabla_{\bar\theta}(\nabla^2 \mathrm{g} : v \otimes v) * (f - \mu)) \\ - \mathrm{div}_v(f(\nabla^2 \mathrm{g} v) * (f - \mu)) = 0,$$

- Weighted Wasserstein-2 gradient flow, with metric $d\theta^2 + \eta\, dv^2$, for the free energy

$$\mathcal{F}(f) = \frac{1}{2}\int \mathrm{g} * (\rho - \mu)\, d(\rho - \mu) + \frac{\eta}{4}\int \nabla^2 \mathrm{g} : (v \otimes v) * (f - \mu)\, d(f - \mu).$$

- $\eta \propto N \Longrightarrow$ strong particle correlations : mean-field may be insufficient.

Theorem for the kinetic PDE[6]

$$\frac{\partial f}{\partial t} - \mathrm{div}_{\bar\theta}(f \nabla \mathrm{g} * (\rho - \mu)) - \frac{1}{2}\eta\, \mathrm{div}_{\bar\theta}(f \nabla_{\bar\theta}(\nabla^2 \mathrm{g} : v \otimes v) * (f - \mu)) - \mathrm{div}_v(f(\nabla^2 \mathrm{g} v) * (f - \mu)) = 0, \quad (2)$$

Theorem (ACDC)

Let $\mathrm{g} \in C_b^{3,1}$, $f_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d \times r})$ *compactly supported in the* v *variable and* $\mu \in \mathcal{P}(\mathbb{R}^d)$.
For all $T > 0$, *there exists a unique, compactly supported in* v, *solution* $f \in C([0, T], \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^{d \times r}))$ *to* (2), *which furthermore satisfies*

$$\frac{d}{dt}\mathcal{F}(f) + \int_{\mathbb{R}^d \times \mathbb{R}^{d \times r}} \left(\left| \nabla_\theta \frac{\delta \mathcal{F}}{\delta f} \right|^2 + \frac{1}{\eta} \left| \nabla_v \frac{\delta \mathcal{F}}{\delta f} \right|^2 \right) df \leq 0. \quad (3)$$

Finally, if g *is smooth enough, the regularity of* f_0 *is propagated in time.*

[6] Antonin Chodron de Courcel. *Gradient descent at the Edge of Stability: free energy model and kinetic description of the two-layer network.* 2026. arXiv: 2606.05326 [cs.LG]. URL: https://arxiv.org/abs/2606.05326.

Singularities and saturation effects in porous media

Framework

Nonlinear mobility

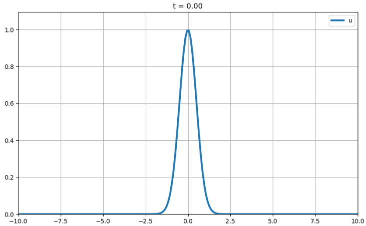
- Equations of the form:

$$\partial_t u - \mathrm{div}(u^m \nabla \mathrm{g} * u) = 0, \quad \mu(u) = u^{m-1} \quad \text{nonlinear mobility.}$$

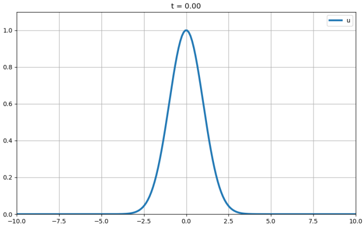
- Riesz interaction $\mathrm{g}(x) \approx |x|^{-s}$, $s \in (0, d)$:

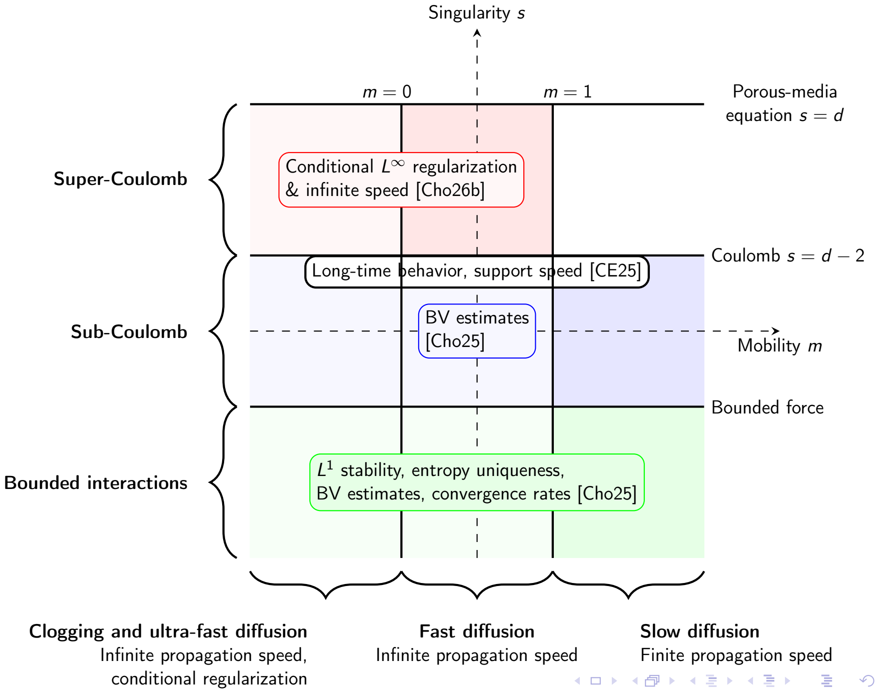
$$\partial_t u - m u^{m-1} \nabla \mathrm{g} * u \cdot \nabla u + u^m (-\Delta)^{\frac{s-d+2}{2}} u = 0.$$

Slow diffusion $m \geq 1$



Fast diffusion $m < 1$




Singularity s
$m = 0$
$m = 1$
Porous-media equation $s = d$
Super-Coulomb
Conditional L^∞ regularization & infinite speed [Cho26b]
Coulomb $s = d - 2$
Long-time behavior, support speed [CE25]
Sub-Coulomb
BV estimates [Cho25]
Mobility m
Bounded force
Bounded interactions
L^1 stability, entropy uniqueness, BV estimates, convergence rates [Cho25]
Clogging and ultra-fast diffusion
Infinite propagation speed, conditional regularization
Fast diffusion
Infinite propagation speed
Slow diffusion
Finite propagation speed

The Coulomb case

The radial case

- **Key issues:**
  - shocks and rarefaction waves.
  - support propagation.
- **Radial case**[7]**:** link with Hamilton-Jacobi equations:

$$k(t,s) := \int_0^s u(t,r)\,dr$$

solves

$$\partial_t k + (\partial_s k)^m k = 0.$$

[7] José A. Carrillo, David Gómez-Castro, and Juan Luis Vázquez. "A fast regularisation of a Newtonian vortex equation". In: *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire* 39.3 (2022), pp. 705–747. DOI: 10.4171/AIHPC/17; José A. Carrillo, David Gómez-Castro, and Juan Luis Vázquez. "Vortex formation for a non-local interaction model with Newtonian repulsion and superlinear mobility". In: *Advances in Nonlinear Analysis* 11.1 (2022), pp. 937–967. DOI: doi:10.1515/anona-2021-0231. URL: https://doi.org/10.1515/anona-2021-0231.

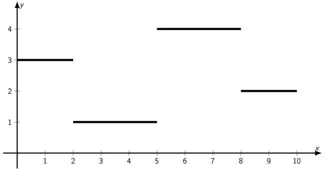
The non-radial Coulomb case[8]

- Compactness.
- Support analysis through decreasing rearrangement:

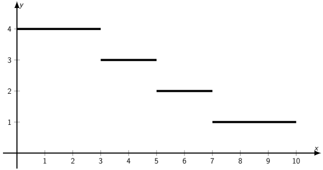
$$k(t,s) := \int_0^s u_*(t,\sigma)\,d\sigma,$$

where $u_*(t,\sigma) := \inf\{\xi : |\{u(t) > \xi\}| \le \sigma\}$.

Original function $u(x)$

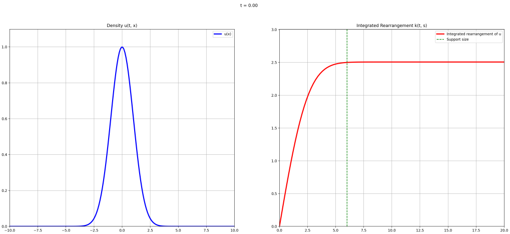


Decreasing rearrangement $u_*(x)$



- k is a viscosity subsolution of $\partial_t k + (\partial_s k)^m k = 0$.

[8] Antonin Chodron de Courcel and Charles Elbar. "On a repulsion model with Coulomb interaction and nonlinear mobility". In: *arXiv:2510.16894* (2025).



Theorem (ACDC, C. Elbar)

Let $u_0 \in L^\infty(\mathbb{T}^d)$ be nonnegative and let $m \geq 1$. There exists an entropy solution u whose support grows instantly if the low-density layer of u_0 is sufficiently thin, i.e.

$$\frac{k(0,S_0) - k(0,s)}{(S_0 - s)^{\frac{m}{m-1}}} \gg 1 \text{ as } s \to S_0, \quad S_0 := |\operatorname{supp} u_0|.$$

Conversely, if $u_0^{m-1} \in C^{0,1}$, then the support size remains constant for a positive waiting time.

References

[CE25] Antonin Chodron de Courcel and Charles Elbar. "On a repulsion model with Coulomb interaction and nonlinear mobility". In: *arXiv:2510.16894* (2025).

[Cho25] Antonin Chodron de Courcel. "Well-posedness of multidimensional nonlocal conservation laws with nonlinear mobility and bounded force". In: *arXiv:2512.13535* (2025).

[Cho26a] Antonin Chodron de Courcel. *Gradient descent at the Edge of Stability: free energy model and kinetic description of the two-layer network*. 2026. arXiv: 2606.05326 [cs.LG]. URL: https://arxiv.org/abs/2606.05326.

[Cho26b] Antonin Chodron de Courcel. "On Clogged and Fast Diffusions in Porous Media with Fractional Pressure". In: *SIAM J. Math. Anal.* 58.2 (2026), pp. 1367–1390.

[CR26] Antonin Chodron de Courcel and Matthew Rosenzweig. *About the Wasserstein gradient flow of the Coulomb discrepancy on $\mathbb{R}^d$ and $\mathbb{T}^d$*. 2026.

Thank you for your attention!