Exact asymptotics and universality for gradient flows and other first order algorithms

Andrea Montanari

Stanford University

January 25, 2022

Michael Celentano

Chen Cheng

Yuchen Wu

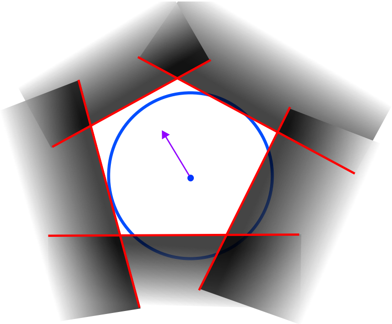
Toy problem

Find $\boldsymbol{\theta} \in \mathbb{R}^d$, such that $\|\boldsymbol{\theta}\|_2 = 1$, and

$$\begin{cases} \langle \boldsymbol{x}_1, \boldsymbol{\theta} \rangle \geq \boldsymbol{\kappa}\,, \\ \vdots \\ \langle \boldsymbol{x}_n, \boldsymbol{\theta} \rangle \geq \boldsymbol{\kappa}\,. \end{cases}$$

$(\boldsymbol{\kappa} < 0)$

Toy problem ('Negative perceptron')



$$(\boldsymbol{x}_i)_{i\le n} \sim \mathsf{N}(0, \boldsymbol{I}_d)\,,$$
$$\mathcal{E}_{n,d}(\kappa) := \{\boldsymbol{\theta} \in \mathbb{S}^{d-1} : \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle \ge \kappa \ \ \forall i \le n\}\,.$$

Franz, Parisi 2016 (Physics); El Alaoui, Sellke, 2020; Baldi, Vershynin, 2021 (Math)

Toy problem ('Negative perceptron')

$$
\begin{aligned}
&(\boldsymbol{x}_i)_{i\le n} \sim \mathsf{N}(0, \boldsymbol{I}_d)\,,\\
&\mathcal{E}_{n,d}(\boldsymbol{\kappa}) := \{\boldsymbol{\theta} \in \mathbb{S}^{d-1} : \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle \ge \boldsymbol{\kappa}\ \forall i \le n\}\,.
\end{aligned}
$$

Questions

- is $\mathcal{E}_{n,d}(\kappa)$ non-empty?
- What is its geometry?
- Can we find $\boldsymbol{\theta} \in \mathcal{E}_{n,d}(\kappa)$ in polynomial time?

Gradient flow

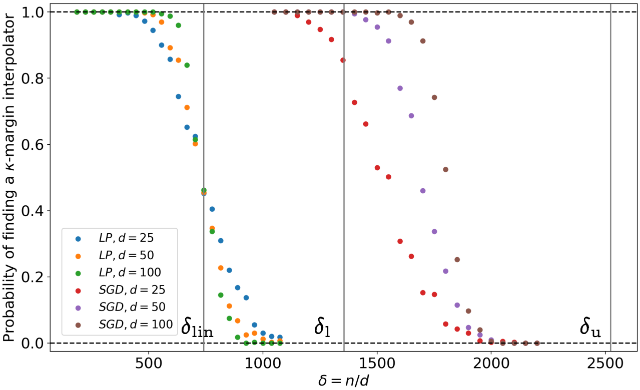
$$\begin{cases} \text{find} & \theta \in \mathbb{S}^{d-1}, \\ \text{such that} & \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle \geq \boldsymbol{\kappa} \quad \forall i \leq n \end{cases}$$

$$\dot{\boldsymbol{\theta}}^t = \sum_{i=1}^{n} \rho(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t \rangle - \boldsymbol{\kappa} \|\boldsymbol{\theta}^t\|_2) \left(\boldsymbol{x}_i - \boldsymbol{\kappa} \hat{\boldsymbol{\theta}}^t \right) ,$$

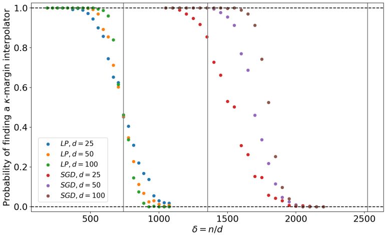
$$\hat{\boldsymbol{\theta}}^t = \frac{\boldsymbol{\theta}^t}{\|\boldsymbol{\theta}^t\|_2}, \qquad \rho(x) := \frac{1}{1 + e^x} .$$

If $\boldsymbol{\theta}^t \to \infty$, then $\hat{\boldsymbol{\theta}}^t \to \mathcal{E}_{n,d}(\kappa)$.

An experiment ($\kappa = -3$)



. . . and a theorem



Theorem (Zhong, Zhou, M, 2021)

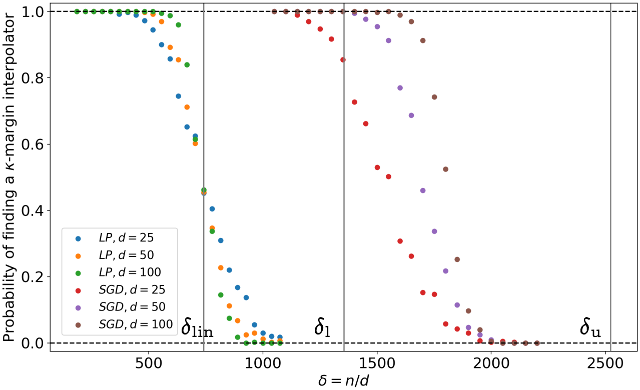
Assume $n, d \to \infty$ with $n, d \to \infty$ and define $\delta_s(\kappa)$ via

$$\delta < \delta_s(\kappa) \quad \Leftrightarrow \quad \liminf_{n\to\infty} \mathbb{P}_{n,n/\delta}(\mathcal{E}_{n,d}(\kappa) \neq \emptyset) > 0$$

Then

- $\delta_l(\kappa) \leq \delta_s(\kappa) \leq \delta_u(\kappa)$, *with* $\delta_l(\kappa), \delta_u(\kappa) = \Phi(\kappa)^{-1} \log|\kappa|(1 + o_\kappa(1))$.
- *Linear prog. succeeds whp if* $\delta < \delta_{\text{lin}}(\kappa) = \Phi(\kappa)^{-1}(1 + o_\kappa(1))$.

Questions



- Does gradient flow have a sharp threshold $\delta_{\mathrm{GF}}(\kappa)$?
- Is $\delta_{\mathrm{GF}}(\kappa) < \delta_{\mathrm{s}}(\kappa)$?
- ...

More ambitious (2-layer neural nets)

$\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$, $y_i \sim \mathsf{Unif}(\{+1, -1\})$, $a_\ell \in \{+1, -1\}$:

$$\begin{cases} \text{find} & \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k \in \mathbb{S}^{d-1}, \\ \text{such that} & y_i \sum_{\ell=1}^{k} a_\ell \sigma(\langle \boldsymbol{\theta}_\ell, \boldsymbol{x}_i \rangle) \geq \kappa \quad \forall i \leq n \end{cases}$$

- What is the interpolation threshold $\delta_{\rm s}(k, \kappa)$?
- Does gradient flow have a sharp threshold $\delta_{\rm GF}(k, \kappa)$?
- Is $\delta_{\rm GF}(k, \kappa) < \delta_{\rm s}(k, \kappa)$?
- ...

Franz, Hwang, Urbani,2019

General first order flows

General First Order Flows

Data

$$\boldsymbol{X} = \begin{bmatrix} \cdots & \boldsymbol{x}_1^\top & \cdots \\ \cdots & \boldsymbol{x}_2^\top & \cdots \\ & \vdots & \\ \cdots & \boldsymbol{x}_n^\top & \cdots \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Variable: $\boldsymbol{\theta} \in \mathbb{R}^{d \times k}$

Flow

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \ell_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y}),$$

General First Order Flows

Data

$$\boldsymbol{X} = \begin{bmatrix} \cdots & \boldsymbol{x}_1^\top & \cdots \\ \cdots & \boldsymbol{x}_2^\top & \cdots \\ & \vdots & \\ \cdots & \boldsymbol{x}_n^\top & \cdots \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Variable: $\boldsymbol{\theta} \in \mathbb{R}^{d \times k}$

Flow

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \ell_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y}),$$

General First Order Flows

Data

$$\boldsymbol{X} = \begin{bmatrix} \cdots & \boldsymbol{x}_1^\top & \cdots \\ \cdots & \boldsymbol{x}_2^\top & \cdots \\ & \vdots & \\ \cdots & \boldsymbol{x}_n^\top & \cdots \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Variable: $\boldsymbol{\theta} \in \mathbb{R}^{d \times k}$

Flow

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \ell_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y}),$$

General First Order Flows

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\mathsf{T} \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,,$$

$$\begin{aligned} \Lambda : \mathbb{R} &\to \mathbb{R}^{k\times k}\,, \\ t &\mapsto \Lambda_t \\ \boldsymbol{\ell}_t : \mathbb{R}^k \times \mathbb{R} &\mapsto \mathbb{R}^k, \\ (\boldsymbol{u}, y) &\mapsto \boldsymbol{\ell}_t(\boldsymbol{u}, y) \quad \text{(applies componentwise.)} \end{aligned}$$

- More general than gradient flow!
- Could add white noise term.

General First Order Flows

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y}),$$

$$\begin{aligned} \Lambda : \mathbb{R} &\to \mathbb{R}^{k\times k}, \\ t &\mapsto \Lambda_t \\ \boldsymbol{\ell}_t : \mathbb{R}^k \times \mathbb{R} &\mapsto \mathbb{R}^k, \\ (\boldsymbol{u}, y) &\mapsto \boldsymbol{\ell}_t(\boldsymbol{u}, y) \quad \text{(applies componentwise.)} \end{aligned}$$

- More general than gradient flow!
- Could add white noise term.

General First Order Flows

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

Example #1: Negative perceptron

$$\dot{\boldsymbol{\theta}}^t = \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \kappa\hat{\boldsymbol{\theta}}^t\right) \rho\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle - \kappa\|\boldsymbol{\theta}^t\|_2\big)\,, \qquad \hat{\boldsymbol{\theta}}^t := \frac{\boldsymbol{\theta}^t}{\|\boldsymbol{\theta}^t\|_2}\,,$$

$$\dot{\boldsymbol{\theta}}^t = \boldsymbol{X}^\top \rho(\boldsymbol{X}\boldsymbol{\theta}^t - u_t\mathbf{1}) - \lambda_t \boldsymbol{\theta}_t\,,$$

$u_t, \lambda_t \in \mathbb{R}$ concentrate.

General First Order Flows

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\mathsf{T} \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

Example #1: Negative perceptron

$$\dot{\boldsymbol{\theta}}^t = \sum_{i=1}^n \left(\boldsymbol{x}_i - \kappa \hat{\boldsymbol{\theta}}^t\right) \rho\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t \rangle - \kappa \|\boldsymbol{\theta}^t\|_2\big)\,, \qquad \hat{\boldsymbol{\theta}}^t := \frac{\boldsymbol{\theta}^t}{\|\boldsymbol{\theta}^t\|_2}\,,$$

$$\dot{\boldsymbol{\theta}}^t = \boldsymbol{X}^\mathsf{T} \rho(\boldsymbol{X}\boldsymbol{\theta}^t - u_t \mathbf{1}) - \lambda_t \boldsymbol{\theta}_t\,,$$

$$u_t, \lambda_t \in \mathbb{R} \quad \text{concentrate.}$$

General First Order Flows

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y}) .$$

Example #2: 2-layer neural network

$$\dot{\boldsymbol{\theta}}_j^t = -\lambda_t \boldsymbol{\theta}_j^t - \nabla_{\boldsymbol{\theta}_j} \widehat{R}_n(\boldsymbol{\theta}^t) \qquad j \le k$$

$$\widehat{R}_n(\boldsymbol{\theta}^t) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \sum_{j \le k} a_j \sigma(\langle \boldsymbol{\theta}_j^t \rangle) \Big)^2 .$$

General First Order Flows: Assumptions

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t\Lambda^t - \boldsymbol{X}^\mathsf{T}\boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

- $n, d \to \infty$, $n/d \to \delta$, k fixed, $t = O(1)$
- $\{X_{ij}\}$ iid $\mathbb{E}\{X_{ij}\} = 0$, $\mathbb{E}\{X_{ij}^2\} = 1/d$, subgaussian.
- $\boldsymbol{\ell}$ differentiable, with bounded Lipschitz Jacobian.
- $y_i = f(\boldsymbol{\theta}_*^\mathsf{T}\boldsymbol{x}_i; z_i)$, $\boldsymbol{\theta}_* \in \mathbb{R}^{d\times k_0}$, z_i independent of $\boldsymbol{x}_i$.
- Empirical distribution of rows of $\boldsymbol{\theta}^0, \boldsymbol{\theta}_*$ converges.

Our approach

1. Discretize time (first order methods)
2. Reduction to Approximate Message Passing[1] (AMP)
3. Apply existing asymptotic characterizations of AMP[2]
4. Take the continuous time limit.

[1] Celentano, M, Wu, 2021
[2] Bayati, M, 2011; Chen, Lam, 2021

Our approach

1. Discretize time **(first order methods)**
2. Reduction to Approximate Message Passing[3] **(AMP)**
3. Apply existing asymptotic characterizations of AMP[4]
4. Take the continuous time limit.

---

[3] Celentano, M, Wu, 2021

[4] Bayati, M, 2011; Chen, Lam, 2021

Time discretization

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

Euler scheme

$$\boldsymbol{\theta}_\varepsilon^{t+\varepsilon} = \boldsymbol{\theta}_\varepsilon^t - \varepsilon \boldsymbol{\theta}_\varepsilon^t \Lambda^t - \varepsilon \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}_\varepsilon^t; \boldsymbol{y})\,.$$

Claim

$$\lim_{\varepsilon \to 0} \limsup_{n,d\to\infty} \sup_{t \le T} \frac{1}{n} \|\theta^t - \theta_\varepsilon^t\|_2^2 = 0$$

Time discretization

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\mathsf{T} \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

Euler scheme

$$\boldsymbol{\theta}_\varepsilon^{t+\varepsilon} = \boldsymbol{\theta}_\varepsilon^t - \varepsilon \boldsymbol{\theta}_\varepsilon^t \Lambda^t - \varepsilon \boldsymbol{X}^\mathsf{T} \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}_\varepsilon^t; \boldsymbol{y})\,.$$

Claim

$$\lim_{\varepsilon\to 0} \limsup_{n,d\to\infty} \sup_{t\le T} \frac{1}{n} \|\boldsymbol{\theta}^t - \boldsymbol{\theta}_\varepsilon^t\|_2^2 = 0$$

Time discretization

$$\boldsymbol{\theta}_{\varepsilon}^{t+\varepsilon} = \boldsymbol{\theta}_{\varepsilon}^{t} - \boldsymbol{\varepsilon}\boldsymbol{\theta}_{\varepsilon}^{t}\Lambda^{t} - \boldsymbol{\varepsilon}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\ell}_{t}(\boldsymbol{X}\boldsymbol{\theta}_{\varepsilon}^{t};\boldsymbol{y})\,.$$

Will drop $\boldsymbol{y}$ hereafter

General First Order Methods

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t)\,, \\
\boldsymbol{v}^t &= \boldsymbol{X}^\top g_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,, \\
\hat{\boldsymbol{\theta}}^t &= h_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,.
\end{aligned}
$$

- $\boldsymbol{u}^t \in \mathbb{R}^{d \times k}$, $\boldsymbol{v}^t \in \mathbb{R}^{n \times k}$.
- Each iteration multiplication by $\boldsymbol{X}$ or $\boldsymbol{X}^\top$.
- f_t, g_t, h_t Lipschitz, act entry-wise[5].

Claim:

Euler scheme is a GFOM.

[5] More general assumptions, see M, Wu, 2022

General First Order Methods

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \ldots, \boldsymbol{v}^t)\,, \\
\boldsymbol{v}^{t} &= \boldsymbol{X}^{\mathsf{T}} g_t(\boldsymbol{u}^1, \ldots, \boldsymbol{u}^t)\,, \\
\hat{\boldsymbol{\theta}}^{t} &= h_t(\boldsymbol{u}^1, \ldots, \boldsymbol{u}^t)\,.
\end{aligned}
$$

- $\boldsymbol{u}^t \in \mathbb{R}^{d\times k}$, $\boldsymbol{v}^t \in \mathbb{R}^{n\times k}$.
- Each iteration multiplication by $\boldsymbol{X}$ or $\boldsymbol{X}^{\mathsf{T}}$.
- f_t, g_t, h_t Lipschitz, act entry-wise[5].

Claim:

Euler scheme is a GFOM.

[5] More general assumptions, see M, Wu, 2022

Approximate Message Passing (AMP)

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X}\psi_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t) - \sum_{s=1}^{t} \phi_s(\boldsymbol{u}^1, \dots, \boldsymbol{u}^s) \cdot a_{t,s}\,, \\
\boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}}\phi_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t) - \sum_{s=1}^{t-1} \psi_s(\boldsymbol{v}^1, \dots, \boldsymbol{v}^s) \cdot b_{t,s}\,, \\
\hat{\boldsymbol{\theta}}^t &= \vartheta_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,.
\end{aligned}
$$

- $\psi_t, \phi_t, \vartheta_t$ Lipschitz, act entry-wise[6].
- $a_{t,s}$, $b_{t,s} \in \mathbb{R}^{k\times k}$ deterministic, with explicit formulas.
- Key property (**state evolution**)

$$
\frac{1}{d}\sum_{i=1}^{d} \delta_{u_i^1,\dots u_i^t} \Rightarrow \mathsf{N}(0, \boldsymbol{\Sigma}_{\le t})\,, \quad + \text{recursion for } \boldsymbol{\Sigma}\,.
$$

[6]More general assumptions, see Berthier, M, Nguyen 2018; M, Wu, 2022

Approximate Message Passing (AMP)

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X}\psi_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t) - \sum_{s=1}^{t} \phi_s(\boldsymbol{u}^1, \dots, \boldsymbol{u}^s) \cdot a_{t,s}\,, \\
\boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}}\phi_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t) - \sum_{s=1}^{t-1} \psi_s(\boldsymbol{v}^1, \dots, \boldsymbol{v}^s) \cdot b_{t,s}\,, \\
\hat{\boldsymbol{\theta}}^t &= \vartheta_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,.
\end{aligned}
$$

Remarks

- Any AMP algorithm is a GFOM
- Any GFOM is equivalent to an AMP algorithm[7]
  (With suitable post-processing ϑ_t)

[7] Celentano, M, Wu, 2021

Approximate Message Passing (AMP)

$$
\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X}\psi_t(\boldsymbol{v}^1,\dots,\boldsymbol{v}^t) - \sum_{s=1}^{t}\phi_s(\boldsymbol{u}^1,\dots,\boldsymbol{u}^s)\cdot a_{t,s}\,,\\
\boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}}\phi_t(\boldsymbol{u}^1,\dots,\boldsymbol{u}^t) - \sum_{s=1}^{t-1}\psi_s(\boldsymbol{v}^1,\dots,\boldsymbol{v}^s)\cdot b_{t,s}\,,\\
\hat{\boldsymbol{\theta}}^t &= \vartheta_t(\boldsymbol{u}^1,\dots,\boldsymbol{u}^t)\,.
\end{aligned}
$$

Remarks

- Any AMP algorithm is a GFOM
- **Any GFOM is equivalent to an AMP algorithm**[7]
  (With suitable post-processing ϑ_t)

[7] Celentano, M, Wu, 2021

DMFT asymptotics

Dynamical Mean Field Theory

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t; \boldsymbol{y})\,.$$

General form: As $n, d \to \infty$,

$$(\theta_i^t)_{t \le T} \Rightarrow (\theta^t)_{t \le T}$$

where $(\theta^t)_{t \le T}$ is a diffusion process with memory.

Dynamical Mean Field Theory: History

Physics

- Sompolinskiy, Zippelius, 1981 (SK model)
- Crisanti, Horner, Sommers, 1993; Cugliandolo, Kurchan, 1993 (spherical spin glass)
- ...

Mathematics

- Ben Arous and Guionnet, 1995 (SK model)
- Ben Arous, Dembo, Guionnet 2006 (spherical spin glass)
- ... (Gaussian disorder)

Dynamical Mean Field Theory: Recent work

(Heuristic) Applications in high-dimensional statistics

- Agoritsas et al., 2018 (Perceptrons)
- Mannelli et al. 2020 (Tensor PCA)
- Mignacco et al. 2020 (Gaussian mixtures)
- ...

Universality

- Dembo, Lubetzky, Zeitouni, 2019 (Asymmetric interactions)
- Dembo, Gheissari, 2021 (Interacting diffusions)

Present work yields universality

Dynamical Mean Field Theory: Recent work

(Heuristic) Applications in high-dimensional statistics

- Agoritsas et al., 2018 (Perceptrons)
- Mannelli et al. 2020 (Tensor PCA)
- Mignacco et al. 2020 (Gaussian mixtures)
- ...

Universality

- Dembo, Lubetzky, Zeitouni, 2019 (Asymmetric interactions)
- Dembo, Gheissari, 2021 (Interacting diffusions)

Present work yields universality

Dynamical Mean Field Theory: Recent work

(Heuristic) Applications in high-dimensional statistics

- Agoritsas et al., 2018 (Perceptrons)
- Mannelli et al. 2020 (Tensor PCA)
- Mignacco et al. 2020 (Gaussian mixtures)
- ...

Universality

- Dembo, Lubetzky, Zeitouni, 2019 (Asymmetric interactions)
- Dembo, Gheissari, 2021 (Interacting diffusions)

Present work yields universality

DMFT process

$$\frac{\mathrm{d}\boldsymbol{\theta}^t}{\mathrm{d}t} = -\boldsymbol{\theta}^t \Lambda^t - \frac{1}{\delta} \boldsymbol{X}^\top \boldsymbol{\ell}_t(\boldsymbol{X}\boldsymbol{\theta}^t), \quad \boldsymbol{\theta}^t \in \mathbb{R}^{d\times k}.$$

DMFT process $(\theta^t, r^t \in \mathbb{R}^k)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^t = -(\Lambda^t + \Gamma^t)\theta^t - \int_0^t R_\ell(t,s)\theta^s \mathrm{d}s + u^t, \quad u \sim \mathsf{GP}(0, C_\ell/\delta),$$

$$r^t = -\frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s)\mathrm{d}s + w^t, \quad w \sim \mathsf{GP}(0, C_\theta).$$

DMFT: Fixed point conditions

DMFT process $(\theta^t, r^t \in \mathbb{R}^k)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^t = -(\Lambda^t + \Gamma^t)\theta^t - \int_0^t R_\ell(t,s)\theta^s \mathrm{d}s + u^t\,, \qquad u \sim \mathsf{GP}(0, C_\ell/\delta)\,,$$

$$r^t = -\frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s)\mathrm{d}s + w^t\,, \qquad w \sim \mathsf{GP}(0, C_\theta)\,.$$

Fixed point equations

$$C_\theta(t,s) = \mathbb{E}\left[\theta^t {\theta^s}^{\mathsf{T}}\right]\,, \qquad R_\theta(t,s) = \mathbb{E}\left[\frac{\partial\theta^t}{\partial u^s}\right]\,,$$

$$C_\ell(t,s) = \mathbb{E}\left[\ell_t(r^t)\ell_s(r^t)^{\mathsf{T}}\right]\,, \qquad R_\ell(t,s) = \mathbb{E}\left[\frac{\partial\ell_t(r^t)}{\partial w^s}\right]\,,$$

$$\Gamma^t = \mathbb{E}\left[\nabla_r \ell_t(r^t; z)\right]\,.$$

DMFT: Fixed point conditions

DMFT process $(\theta^t, r^t \in \mathbb{R}^k)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta^t = -(\Lambda^t + \Gamma^t)\theta^t - \int_0^t R_\ell(t,s)\theta^s \mathrm{d}s + u^t\,, \qquad u \sim \mathsf{GP}(0, C_\ell/\delta)\,,$$

$$r^t = -\frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s)\mathrm{d}s + w^t\,, \qquad w \sim \mathsf{GP}(0, C_\theta)\,.$$

Fixed point equations

$$C_\theta(t,s) = \mathbb{E}\left[\theta^t \theta^{s\mathsf{T}}\right]\,, \qquad R_\theta(t,s) = \mathbb{E}\left[\frac{\partial \theta^t}{\partial u^s}\right]\,,$$

$$C_\ell(t,s) = \mathbb{E}\left[\ell_t(r^t)\ell_s(r^t)^\mathsf{T}\right]\,, \qquad R_\ell(t,s) = \mathbb{E}\left[\frac{\partial \ell_t(r^t)}{\partial w^s}\right]\,,$$

$$\Gamma^t = \mathbb{E}\left[\nabla_r \ell_t(r^t; z)\right]\,.$$

DMFT: Fixed point conditions

Fixed point equations

$$C_\theta(t,s) = \mathbb{E}\left[\theta^t \theta^{s\mathsf{T}}\right], \qquad R_\theta(t,s) = \mathbb{E}\left[\frac{\partial \theta^t}{\partial u^s}\right],$$

$$C_\ell(t,s) = \mathbb{E}\left[\ell_t(r^t)\ell_s(r^t)^\mathsf{T}\right], \qquad R_\ell(t,s) = \mathbb{E}\left[\frac{\partial \ell_t(r^t)}{\partial w^s}\right],$$

$$\Gamma^t = \mathbb{E}\left[\nabla_r \ell_t(r^t; z)\right].$$

$$(C_\theta, R_\theta) = \mathcal{T}(C_\ell, R_\ell, \Gamma), \qquad (C_\ell, R_\ell, \Gamma) = \hat{\mathcal{T}}(C_\theta, R_\theta)$$

1st theorem: Existence and uniqueness

Theorem (Celentano, Chen, M, 2021)

- *The solution* $(C_\theta, R_\theta, C_\ell, R_\ell, \Gamma)$ *of the fixed point conditions exists.*
- *It is unique among all tuples such that* (C_θ, R_θ) *are bounded in all compact sets in* $\mathbb{R}^2_{\geq 0}$*.*
- *The DMFT processes* $(\theta^t)_{t\geq 0}$, $(r^t)_{t\geq 0}$ *are well defined with continuous sample paths.*
- *The map* $\mathcal{T} \circ \hat{\mathcal{T}}$ *is a contraction.*

2nd theorem: Convergence as n, $d \to \infty$

Theorem (Celentano, Chen, M, 2021)

Let $d_{\rm W}$ metrize weak convergence in $C([0,T],\mathbb{R}^k)$, and define $\boldsymbol{r}^t := \boldsymbol{X}\boldsymbol{\theta}^t \in \mathbb{R}^{n\times k}$.
Let $\mathrm{P}_{\theta_0^T}$, $\mathrm{P}_{r_0^T}$ be the laws of the DMFT processes.
Then, we have

$$\operatorname*{p-lim}_{n,d\to\infty} d_{\rm W}\Big(\frac{1}{d}\sum_{i=1}^{d}\delta_{(\theta_i)_0^T}, \mathrm{P}_{\theta_0^T}\Big) = 0\,,$$

$$\operatorname*{p-lim}_{n,d\to\infty} d_{\rm W}\Big(\frac{1}{n}\sum_{i=1}^{n}\delta_{(r_i)_0^T}, \mathrm{P}_{r_0^T}\Big) = 0\,.$$

Convergence as $n, d \to \infty$

Corollary

For any m, any $\psi : \mathbb{R}^m \to \mathbb{R}$ bounded continuous, and any $0 \le t_1 \le t_2 \le \cdots \le t_m$, we have

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(\theta_i^{t_1}, \dots, \theta_i^{t_m}) = \mathbb{E}\{\psi(\theta^{t_1}, \dots, \theta^{t_m})\}\,.$$

$$\underbrace{\frac{1}{d}\sum_{i=1}^{d}\psi(\theta_i^{t_1},\dots,\theta_i^{t_m})}_{\text{avg over coordinates}} \quad \xrightarrow{P} \quad \underbrace{\mathbb{E}\{\psi(\theta^{t_1},\dots,\theta^{t_m})\}}_{\text{expectation wrt DMFT}} \quad .$$

Epilogue: Statistically optimal methods

Reduction

General First Order Methods $\longrightarrow$ Approximate Message Passing

Q: Can we determine the optimal GFOM?

Statistical optimality

- Data

$$y_i = \varphi(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle, z_i)\,, \qquad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d/d) \perp z_i \sim \mathsf{Unif}([0,1])\,.$$

- General GFOM

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t)\,, \\ \boldsymbol{v}^t &= \boldsymbol{X}^\top g_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,, \\ \hat{\boldsymbol{\theta}}^t &= h_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,. \end{aligned}$$

$$\lim_{n,d\to\infty} \frac{1}{d} \inf_{\{f_s, g_s, h_s\} \in \mathscr{F}} \mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_*\|_2^2\} = ???$$

Statistical optimality

- Data

$$y_i = \varphi(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle, z_i)\,, \qquad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d/d) \;\perp\; z_i \sim \mathsf{Unif}([0,1])\,.$$

- General GFOM

$$\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t)\,,\\
\boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}} g_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,,\\
\hat{\boldsymbol{\theta}}^t &= h_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,.
\end{aligned}$$

$$\lim_{n,d\to\infty} \frac{1}{d} \inf_{\{f_s, g_s, h_s\}\in\mathscr{F}} \mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_*\|_2^2\} = ???$$

Statistical optimality

- Data

$$y_i = \varphi(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle, z_i)\,, \qquad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d/d) \;\perp\; z_i \sim \mathsf{Unif}([0,1])\,.$$

- General GFOM

$$\begin{aligned}
\boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t)\,, \\
\boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}} g_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,, \\
\hat{\boldsymbol{\theta}}^t &= h_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,.
\end{aligned}$$

$$\lim_{n,d\to\infty} \frac{1}{d} \inf_{\{f_s, g_s, h_s\} \in \mathscr{F}} \mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_*\|_2^2\} = ???$$

Statistical optimality

- Data

$$y_i = \varphi(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle; z_i)\,, \qquad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d/d) \perp z_i \sim \mathsf{Unif}([0,1])\,.$$

- General GFOM

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \boldsymbol{X} f_t(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t)\,, \\ \boldsymbol{v}^t &= \boldsymbol{X}^{\mathsf{T}} g_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,, \\ \hat{\boldsymbol{\theta}}^t &= h_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t)\,. \end{aligned}$$

$$\lim_{n,d\to\infty} \frac{1}{d} \inf_{\{f_s, g_s, h_s\} \in \mathscr{F}} \mathbb{E}\{\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_*\|_2^2\} = ???$$

$\mathscr{F}$:= separable Lipschitz functions

Example: Noiseless phase retrieval

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle^2$$

Spectral initialization[8]

$$\begin{aligned}
\boldsymbol{D}_n &:= \sum_{i=1}^{n} \mathcal{T}(y_i) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}, \quad \mathcal{T}(y) = \frac{y-1}{y+\varepsilon}, \\
\boldsymbol{\theta}^0 &= c_{d,n} \cdot \boldsymbol{v}_1(\boldsymbol{D}_n).
\end{aligned}$$

[8] Mondelli, Montanari, 2018

First order methods in phase retrieval

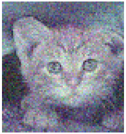
- Schniter, Rangan 2014
- Candés, Li, Soltanolkotabi, 2015
- Cai, Li, Ma, 2016
- Wang, Giannakis, Eldar, 2017
- Chen, Candés 2018
- Waldspurger, 2018
- Duchi, Ruan, 2019
- Maillard, Loureiro, Krzakala, Zdeborová, 2020
- Mondelli Venkatramanan, 2021
- **Review:** Fannjiang and Strohmer, 2020
- ...

Experiment with a real image: $d = 7560$

Original image.

Bayes AMP

$t = 2$



$t = 4$

$t = 8$

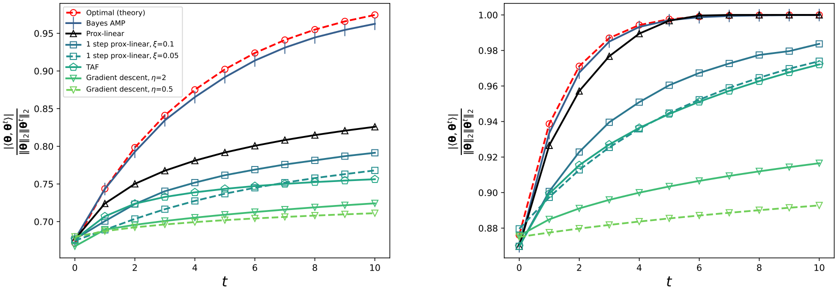
1 step prox-linear

$t = 2$

$t = 4$

$t = 8$

A simulation: Reconstruction accuracy $d = 400$



$n = 600$ $\qquad$ $n = 1000$

- Truncated Amplitude Flow (Wang, Giannakis, Eldar, 2017)
- Prox-linear[9] (Duchi, Ruan, 2019)
- Gradient descent.
- Bayes-optimal AMP.

[9]Not a GFOM.

Minimal error

- $\boldsymbol{y} = \varphi(\boldsymbol{X}\boldsymbol{\theta}_*, \boldsymbol{z})$
- Side information $\boldsymbol{v}$, $d^{-1}\sum_{i\le d}\delta_{\theta_{*,i},v_i} \Rightarrow P_{\Theta,V}$
- $n, d \to \infty$, $n/d \to \delta$ $\{X_{ij}\}$ bounded 4-th moment.

Define

$$\mathsf{mmse}_{\Theta,V}(\alpha) := \mathbb{E}[\Theta^2] - \mathbb{E}\{\mathbb{E}[\Theta \mid \alpha\Theta + Z, V]^2\}.$$

and recursively:

$$\beta_s^2 = \frac{1}{\sigma_s^2}\mathbb{E}[\mathbb{E}[Z_0 \mid \varphi(\sigma_s Z_0 + \tilde{\sigma}_s Z_1, W), U, Z_1]^2], \qquad \beta_s \ge 0,$$

$$\sigma_{s+1}^2 = \frac{1}{\delta}\mathsf{mmse}_{\Theta,V}(\beta_s), \qquad \tilde{\sigma}_{s+1}^2 = \frac{1}{\delta}(\mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,V}(\beta_s)).$$

Minimal error

Theorem (Celentano, M, Wu, 2020; M, Wu, 2022)

For $t \in \mathsf{NN}_{>0}$, *let* $\hat{\boldsymbol{\theta}}^t \in \mathbb{R}^d$ *be the output of any GFOM after* t *iterations, the following holds:*

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_*\|_2^2 \geq \mathsf{mmse}_{\Theta,V}(\beta_t).$$

Further, there exists a GFOM (Bayes AMP) which satisfies the above bound with equality.

Other settings in M, Wu, 2022

Conclusion

Conclusion

First order methods comprise the most used algorithms in ML

- GFOMs as a useful abstraction/class.
- Reduction GFOM $\to$ AMP
- Need to understand their behavior in high-dim/low SNR.
- New rigorous tool: DMFT (other availlable!)

Thanks!