10/17/24

Lecture 12 : Mean-field limit

Derivation and meaning of continuity equation:

$$\partial_t \rho_t = \operatorname{div}(\rho_t \cdot \nabla \Psi_{\rho_t}) \qquad (*)$$

Holds in "weak sense", i.e. for any "nice" (e.g. bounded, differentiable, with bounded gradient) test function $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\int \varphi(\theta)\, \partial_t \rho_t(\theta)\, d\theta = \int \varphi(\theta) \cdot \operatorname{div}(\rho_t \cdot \nabla \Psi_{\rho_t})(\theta)\, d\theta \qquad (**)$$

(b/c differentiable solution to $(*)$ may not exist).

Note that for $\bar{\Theta}_t \sim \rho_t$,

$$\text{LHS } (*) = \frac{\partial}{\partial t} \mathbb{E}\left[\varphi(\bar{\Theta}_t)\right]$$

(diff. under integral)

$$= \mathbb{E}\left[\left\langle \nabla \varphi(\bar{\Theta}_t), \frac{d}{dt}\bar{\Theta}_t \right\rangle\right]$$

(gradient flow for $\bar{\Theta}_t$)

$$= \int \left\langle \nabla \varphi(\theta), -\nabla \Psi_{\rho_t}(\theta) \right\rangle d\rho_t(\theta) \quad =$$

(integration by parts)

$$\text{RHS } (*) = -\int \left\langle \nabla \varphi(\theta), \nabla \Psi_{\rho_t}(\theta) \right\rangle d\rho_t(\theta)$$

Non-asymptotic convergence to the mean-field limit:

"Propagation of chaos" [Kac '56], [McKean '69], [Sznitman '91]

want to compare:

- $\left(\Theta_i^{(k)}\right)_{k=0,1,2,\dots}$: GD iterates given by $\Theta_i^{(k+1)} \leftarrow \Theta_i^{(k)} - h\nabla L(\Theta^{(k)})$

- $\left(\bar{\Theta}_i^{t}\right)_{t \geq 0}$: mean-field iterates given by $d\bar{\Theta}_i^t = -\nabla L_{\rho_t}(\bar{\Theta}_i^t)\,dt$ where $\rho_t = \text{law}(\bar{\Theta}_i^t)$

Note:

$$\Theta_i^{(k)} = \Theta_i^{(0)} + 2h\sum_{l=0}^{k-1} F_i(\Theta^{(l)}; (x_{l+1}, y_{l+1}))$$

$$\bar{\Theta}_i^{t} = \Theta_i^{(0)} + 2\int_0^t G(\bar{\Theta}_i^s; \rho_s)\,ds$$

for $F_i(\Theta; (x,y)) \triangleq (y - f_\Theta(x)) \cdot \nabla_{\Theta_i}\sigma(x;\Theta_i)$

$G(\Theta, \rho) \triangleq -\nabla\Psi_\rho(\Theta)$

Our goal: upper bound $\left\|\bar{\Theta}_i^{kh} - \Theta_i^{(k)}\right\|$

To do so, will bound by a self-similar expression of the form

$$(\text{small terms}) + \int_0^{kh} \left\|\bar{\Theta}_i^s - \Theta_i^{(\lfloor s/h \rfloor)}\right\| ds$$

This will imply (by Grönwall's inequality), the desired bound

Let $[s] = h \cdot \lfloor s/h \rfloor$

$$\| \bar{\Theta}_i^{kh} - \Theta_i^{k} \|$$

$$= 2 \left\| \int_0^{kh} G(\bar{\Theta}_i^{s}; \rho_s)\, ds - h \sum_{\ell=0}^{k-1} F_i(\Theta^{(\ell)}, (x_{\ell+1}, y_{\ell+1})) \right\|$$

$$\leq 2 \left\| \int_0^{kh} \left[G(\bar{\Theta}_i^{s}; \rho_s) - G(\bar{\Theta}_i^{[s]}; \rho_{[s]}) \right] ds \right\| \quad ①$$

$$+$$

$$2 \left\| \int_0^{kh} \left[G(\bar{\Theta}_i^{[s]}; \rho_{[s]}) - G(\Theta_i^{(\lfloor s/h \rfloor)}; \rho_{[s]}) \right] ds \right\| \quad ②$$

$$+$$

$$2 \left\| h \sum_{\ell=0}^{k-1} \left[G(\Theta_i^{(\ell)}; \rho_{\ell h}) - F_i(\Theta^{(\ell)}; (x_{\ell+1}, y_{\ell+1})) \right] \right\| \quad ③$$

---

① (easy):

Small because G is Lipschitz by assumption, and can show ρ varies smoothly over time so that ρ_s and $\rho_{[s]}$ are close

(2) :

Again by Lipschitzness of G,

$$\|G(\bar{\theta}_i^s ; \rho_{[s]}) - G(\theta_i^{(\lfloor s/h \rfloor)} ; \rho_{[s]})\| \lesssim \|\bar{\theta}_i^s - \theta_i^{(\lfloor s/h \rfloor)}\|,$$

so (2) is bounded by

$$\int_0^{kh} \underbrace{\|\bar{\theta}_i^s - \theta_i^{(\lfloor s/h \rfloor)}\|}_{\text{looks analogous to what we want to bound on LHS...}} ds$$

(3): $\sum_{\ell=0}^{k-1} \left[G(\theta_i^{(\ell)} ; \rho_{\ell h}) - F_i(\theta^{(\ell)} ; (x_{\ell+1}, y_{\ell+1})) \right]$

Key idea: this has expectation $G(\theta_i^{(\ell)} ; \hat{\rho}_\ell)$,

where $\hat{\rho}_\ell$ is empirical dist $\frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i^{(\ell)}}$

Over many steps ℓ, the total deviation between $F_i(\theta^{(\ell)} ; (x_{\ell+1}, y_{\ell+1}))$'s and $G(\theta_i^{(\ell)} ; \hat{\rho}_\ell)$'s is of order $h\sqrt{kp}$ by <u>martingale concentration</u>

Remains to bound

$$\sum_{l=0}^{k-1}\left[G(\theta_i^{(l)};\rho_{lh}) - G(\theta_i^{(l)};\hat{\rho}_l)\right]$$

$$= \frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\left[\mathbb{E}_{\bar{\theta}}\, U(\theta_i^{(l)}, \bar{\theta}_j^{lh}) - U(\theta_i^{(l)}, \theta_j^{(l)})\right]$$

again, by martingale concentration we can essentially replace $\mathbb{E}_{\bar{\theta}}\, U(\theta_i^{(l)}, \bar{\theta}_j^{lh})$ (deterministic)

with $U(\theta_i^{(l)}, \bar{\theta}_j^{lh})$ (random)

Then we use Lipschitzness of U to get

$$\frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\left\| U(\theta_i^{(l)}, \bar{\theta}_j^{lh}) - U(\theta_i^{(l)}, \theta_j^{(l)})\right\|$$

$$\leq \frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\| \underbrace{\bar{\theta}_j^{lh} - \theta_j^{(l)}}_{\text{once again, a term that looks similar to what we want to bound}} \|$$

When data distribution has symmetries, PDE simplifies considerably:

Suppose training data $\{(x_i, y_i)\}$ satisfy $x_i \sim N(0, I_d)$ and $y_i = \varphi(\Pi x)$ for Π a projection to a low-dim subspace V°.

Then joint dist over (x,y) invariant under rotations of x that preserve V°, i.e. $Rv \in V^\circ$ $\forall$ $v \in V^\circ$.

Observation: Let R be such a rotation. If ρ_0 and ρ_0' are two different initializations of the weights related by $\rho_0' = R_\# \rho_0$ (i.e. to sample (a', w') from ρ_0', sample (a, w) from ρ_0 and take $a' = a$, $w' = Rw$), then $\rho_t' = R_\# \rho_t$.

So if ρ_0 rotation-invariant, ρ_t is invariant to rotations preserving V°, for any $t \geq 0$!

ρ_t thus completely specified by distribution on

$$\Big(a, \underbrace{\Pi w}_{= s}, \underbrace{\|\Pi^\perp w\|_2}_{= r}\Big),$$

i.e. we get a $\boxed{\dim(V^{\phi})+2}$ -dimensional PDE!

Denote dist. on $(a, \vec{s}, r)$ by $\bar{\rho}_t$.

$$\partial_t \bar{\rho}_t = \operatorname{div}\left(\bar{\rho}_t \cdot \nabla_{\vec{s}} \Psi_{\bar{\rho}_t}\right) + \partial_a\left(\bar{\rho}_t \cdot \partial_a \Psi_{\bar{\rho}_t}\right) + \frac{1}{r}\partial_r\left(r \cdot \bar{\rho}_t \cdot \partial_r \Psi_{\bar{\rho}_t}\right).$$