A Mean-Field Optimal Control Formulation of Deep Learning

Jiequn Han

Department of Mathematics,
Princeton University
Joint work with Weinan E and Qianxiao Li

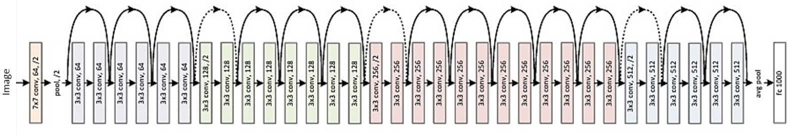
Mathematical Advances in Deep Learning, SIAM CSE19
Spokane, Washington, Feb 25, 2019

Outline

Table of Contents

Great Success of Deep Learning

- Deep learning has achieved remarkable success in many machine learning tasks
- Compositional structure is widely considered the essence of deep neural networks, but the mechanism stills remains mystery.
- Deep residual network (ResNet) and its variants make use of much deeper architectures and achieve the state-of-the-art in numerous computer vision applications.
- composition + skip connection $\rightarrow$ dynamic system



Dynamical System Viewpoint of ResNet

- Residual block

$$x_{l+1} = x_l + f(x_l, W_l)$$

- Closely connected with dynamic system in discrete time

$$x_{t+1} = x_t + f(x_t, W_t)\Delta t$$

Dynamical System Viewpoint of ResNet

- Residual block

$$x_{l+1} = x_l + f(x_l, W_l)$$

- Closely connected with dynamic system in discrete time

$$x_{t+1} = x_t + f(x_t, W_t)\Delta t$$

- Motivate us to consider a formulation in continuous time – independent of time resolution
- Allow us to study deep learning in a new framework that has intimate connections with differential equations, numerical analysis, and optimal control theory
- The compositional structure is explicitly taken into account as time evolution (total time ≈ network depth)

Related Work

- Early work: continuous-time analogs of deep neural networks (E, 2017, Haber and Ruthotto, 2017)

- Most work on the dynamical systems viewpoint of deep learning mainly focused on designing

  - new optimization algorithms: maximum principle based (Li et al., 2017, Li and Hao, 2018), neural ODE (Chen et al., 2018), layer-parallel training (Günther et al., 2018)

  - new network structures: stable structure (Haber and Ruthotto, 2017), multi-level structure (Lu et al., 2017, Chang et al., 2017), reversible structure (Chang et al., 2018)

  However, the mathematical aspects has not been explored yet

- Mean-field optimal control itself is still an active area of research

Mathematical Formulation

Given the data-label joint distribution $(x_0, y_0) \sim \mu$ on $\mathbb{R}^d \times \mathbb{R}^l$, we aim to solve the following population risk minimization problem (E, 2017)

$$\inf_{\boldsymbol{\theta} \in L^\infty([0,T],\Theta)} J(\boldsymbol{\theta}) := \mathbb{E}_\mu \left[\Phi(x_T, y_0) + \int_0^T L(x_t, \theta_t) dt \right],$$

$$\text{Subject to } \dot{x}_t = f(x_t, \theta_t).$$

$T > 0,$	time length (network "depth")
$f : \mathbb{R}^d \times \Theta \to \mathbb{R}^d,$	feed-forward dynamics
$\Phi : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R},$	terminal loss function
$L : \mathbb{R}^d \times \Theta \to \mathbb{R},$	regularizer

Two Sides of the Same Coin: Optimal Control

Maximum principle (Pontrayagin, 1950s): – local characterization of optimal solution in terms of ODEs of state and co-state variables, giving necessary condition

Dynamic programming (Bellman, 1950s) – global characterization of the value function in terms of PDE (HJB equation), giving necessary and sufficient condition / later made rigorous by the development of viscosity solution by Crandall and Lions (1980s)

Intimately connected through the method of characteristics in Hamiltonian mechanics

Table of Contents

Mean-Field Pontrayagin's Maximum Principle

We assume:

- (A1) The function f is bounded; f, L are continuous in θ; and f, L, Φ are continuously differentiable with respect to x.
- (A2) The distribution μ has bounded support in $\mathbb{R}^d \times \mathbb{R}^l$.

Mean-Field Pontrayagin's Maximum Principle

We assume:

- (A1) The function f is bounded; f, L are continuous in θ; and f, L, Φ are continuously differentiable with respect to x.
- (A2) The distribution μ has bounded support in $\mathbb{R}^d \times \mathbb{R}^l$.

Theorem (Mean-field PMP)

Let (A1), (A2) be satisfied and $\boldsymbol{\theta}^* \in L^\infty([0,T],\Theta)$ be a solution of mean-field optimal control problem. Then, there exists absolutely continuous μ-a.s. stochastic processes $\boldsymbol{x}^*, \boldsymbol{p}^*$ such that

$$
\begin{aligned}
&\dot{x}_t^* = f(x_t^*, \theta_t^*), && x_t^* = x_0,\\
&\dot{p}_t^* = -\nabla_x H(x_t^*, p_t^*, \theta_t^*), && p_T^* = -\nabla_x \Phi(x_T^*, y_0),\\
&\mathbb{E}_\mu H(x_t^*, p_t^*, \theta_t^*) = \max_{\theta\in\Theta} \mathbb{E}_\mu H(x_t^*, p_t^*, \theta), && a.e.\, t \in [0,T],
\end{aligned}
$$

where the Hamiltonian function $H : \mathbb{R}^d \times \mathbb{R}^d \times \Theta \to \mathbb{R}$ is given by $H(x,p,\theta) = p \cdot f(x,\theta) - L(x,\theta)$.

Discussion of Mean-Field PMP

- It is a necessary condition for optimality
- What's new compared to classical PMP: the expectation over μ in the Hamiltonian maximization condition
- It includes, as a special case, the necessary conditions for the optimality of the sampled optimal control problem (by considering the empirical measure $\mu_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_0^i, y_0^i)}$)

$$\min_{\boldsymbol{\theta}\in L^\infty([0,T],\Theta)} J_N(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N}\left[\Phi(x_T^i, y_0^i) + \int_0^T L(x_t^i, \theta_t)dt\right],$$

subject to $\quad \dot{x}_t^i = f(x_t^i, \theta_t), \qquad i = 1, \ldots, N.$

Small-Time Uniqueness

Uniqueness + existence: necessary condition becomes sufficient

In the sequel, assume

- (A1′) f is bounded; f, L, Φ are twice continuously differentiable with respect to both x, θ, with bounded and Lipschitz partial derivatives.

Small-Time Uniqueness

Uniqueness + existence: necessary condition becomes sufficient

In the sequel, assume

- (A1′) f is bounded; f, L, Φ are twice continuously differentiable with respect to both x, θ, with bounded and Lipschitz partial derivatives.

Theorem (Small-time uniqueness)

Suppose that $H(x,p,\theta)$ is strongly concave in θ, uniformly in $x, p \in \mathbb{R}^d$, i.e. $H(x,p,\theta) + \lambda_0 I \preceq 0$ for some $\lambda_0 > 0$. Then, for sufficiently small T, the solution of the PMP is unique.

Remark

- *The strong concavity of the Hamiltonian does not imply that the loss function J is strongly convex, or even convex:* $f(x,\theta) = \theta\sigma(x), L(x) = \frac{1}{2}\lambda\|\theta\|^2$.
- *small $T \to$ low capacity model (the number of parameters is still infinite)*

From Mean-Field PMP to Sampled PMP

Goal:

population risk minimization	$\longleftrightarrow$	empirical risk minimization
mean-field PMP	$\longleftrightarrow$	sampled PMP

From Mean-Field PMP to Sampled PMP

Goal:

population risk minimization	$\longleftrightarrow$	empirical risk minimization
mean-field PMP	$\longleftrightarrow$	sampled PMP

Strategy: Denote

$$\begin{aligned} \dot{x}_t^{\boldsymbol{\theta}} &= f(x_t^{\boldsymbol{\theta}}, \theta_t), \qquad x_0^{\boldsymbol{\theta}} = x_0, \\ \dot{p}_t^{\boldsymbol{\theta}} &= -\nabla_x H(x_t^{\boldsymbol{\theta}}, p_t^{\boldsymbol{\theta}}, \theta_t), \qquad p_T^{\boldsymbol{\theta}} = -\nabla_x \Phi(x_T^{\boldsymbol{\theta}}, y_0). \end{aligned}$$

Assume the solution of mean-field PMP satisfies

$$\boldsymbol{F}(\boldsymbol{\theta}^*)_t := \mathbb{E}\nabla_\theta H(x_t^{\boldsymbol{\theta}^*}, p_t^{\boldsymbol{\theta}^*}, \theta_t^*) = 0.$$

We wish to find the solution of

$$\boldsymbol{F}_N(\boldsymbol{\theta}^N)_t := \frac{1}{N}\sum_{i=1}^N \nabla_\theta H(x_t^{\boldsymbol{\theta}^N, i}, p_t^{\boldsymbol{\theta}^N, i}, \theta_t^N) = 0.$$

This can be done through a contraction mapping

$$G_N(\boldsymbol{\theta}) := \boldsymbol{\theta} - DF_N(\boldsymbol{\theta}^*)^{-1} F_N(\boldsymbol{\theta}).$$

Error Estimates for Sampled PMP

Definition

For $\rho > 0$ and $x \in U$, define $S_\rho(x) := \{y \in U : \|x - y\| \leq \rho\}$. We say that the mapping F is stable on $S_\rho(x)$ if there exists a constant $K_\rho > 0$ such that for all $y, z \in S_\rho(x)$,

$$\|y - z\| \leq K_\rho \|F(y) - F(z)\|.$$

Error Estimates for Sampled PMP

Definition

For $\rho > 0$ and $x \in U$, define $S_\rho(x) := \{y \in U : \|x - y\| \leq \rho\}$. We say that the mapping F is stable on $S_\rho(x)$ if there exists a constant $K_\rho > 0$ such that for all $y, z \in S_\rho(x)$,

$$\|y - z\| \leq K_\rho \|F(y) - F(z)\|.$$

Theorem (Neighboring solution for sampled PMP)

Let $\boldsymbol{\theta}^*$ be a solution $\boldsymbol{F} = 0$, which is stable on $S_\rho(\boldsymbol{\theta}^*)$ for some $\rho > 0$. Then, there exists positive constants s_0, C, K_1, K_2 and $\rho_1 < \rho$ and a random variable $\boldsymbol{\theta}^N \in S_{\rho_1}(\boldsymbol{\theta}^*) \subset L^\infty([0,T], \Theta)$, such that

$$\mu[\|\boldsymbol{\theta} - \boldsymbol{\theta}^N\|_{L^\infty} \geq Cs] \leq 4 \exp\left(-\frac{Ns^2}{K_1 + K_2 s}\right), \qquad s \in (0, s_0],$$

$$\mu[\boldsymbol{F}_N(\boldsymbol{\theta}^N) \neq 0] \leq 4 \exp\left(-\frac{Ns_0^2}{K_1 + K_2 s_0}\right).$$

In particular, $\boldsymbol{\theta}^N \to \boldsymbol{\theta}^*$ and $\boldsymbol{F}_N(\boldsymbol{\theta}^N) \to 0$ in probability.

Error Estimates for Sampled PMP

Theorem

Let $\boldsymbol{\theta}^*$ be a solution of the mean-filed PMP such that there exists $\lambda_0 > 0$ satisfying that for a.e. $t \in [0,T]$, $\mathbb{E}\nabla^2_{\theta\theta} H(x_t^{\boldsymbol{\theta}^*}, p_t^{\boldsymbol{\theta}^*}, \theta_t^*) + \lambda_0 I \preceq 0$. Then the random variable $\boldsymbol{\theta}^N$ defined previously satisfies, with probability at least $1 - 6\exp\left[-(N\lambda_0^2)/(K_1 + K_2\lambda_0)\right]$, that θ_t^N is a strict local maximum of sampled Hamiltonian $\frac{1}{N}\sum_{i=1}^N H(x_t^{\boldsymbol{\theta}^N,i}, p_t^{\boldsymbol{\theta}^N,i}, \theta)$. In particular, if the finite-sampled Hamiltonian has a unique local maximizer, then $\boldsymbol{\theta}^N$ is a solution of the finite-sampled PMP with the same high probability.

Error Estimates for Sampled PMP

Theorem

Let $\boldsymbol{\theta}^*$ be a solution of the mean-filed PMP such that there exists $\lambda_0 > 0$ satisfying that for a.e. $t \in [0, T]$, $\mathbb{E}\nabla^2_{\theta\theta} H(x_t^{\boldsymbol{\theta}^*}, p_t^{\boldsymbol{\theta}^*}, \theta_t^*) + \lambda_0 I \preceq 0$. Then the random variable $\boldsymbol{\theta}^N$ defined previously satisfies, with probability at least $1 - 6\exp\left[-(N\lambda_0^2)/(K_1 + K_2\lambda_0)\right]$, that θ_t^N is a strict local maximum of sampled Hamiltonian $\frac{1}{N}\sum_{i=1}^N H(x_t^{\boldsymbol{\theta}^N, i}, p_t^{\boldsymbol{\theta}^N, i}, \theta)$. In particular, if the finite-sampled Hamiltonian has a unique local maximizer, then $\boldsymbol{\theta}^N$ is a solution of the finite-sampled PMP with the same high probability.

Theorem

Let $\boldsymbol{\theta}^N$ be the random variable defined previously. Then there exist constants K_1, K_2 such that,

$$\mathbb{P}[|J(\boldsymbol{\theta}^N) - J(\boldsymbol{\theta}^*)| \geq s] \leq 4\exp\left(-\frac{Ns^2}{K_1 + K_2 s}\right), \qquad s \in (0, s_0].$$

Table of Contents

Mean-Field Dynamic Programming Principle

Key idea: take the joint distribution of (x_t, y_0) as state variable in Wasserstein space and consider the associated value function as solution of an infinite-dimensional Hamilton-Jacobi-Bellman (HJB) equation. Finally obtain uniqueness, regardless of time length.

Mean-Field Dynamic Programming Principle

Key idea: take the joint distribution of (x_t, y_0) as state variable in Wasserstein space and consider the associated value function as solution of an infinite-dimensional Hamilton-Jacobi-Bellman (HJB) equation. Finally obtain uniqueness, regardless of time length.

Notation:

w	concatenation of (x, y) as $(d+l)$-dimensional variable
$(\Omega, \mathcal{F}, \mathbb{P})$	fixed probability space, $\mathcal{F}$ is the Borel $\sigma-$algebra of $\mathbb{R}^{d+l}$
$L^2(\mathcal{F}; \mathbb{R}^{d+l})$	the space of square-integrable random variables with L^2 metric
$\mathcal{P}_2(\mathbb{R}^{d+l})$	the space of square-integrable measures with 2-Wasserstein metric

$$W \in L^2(\mathcal{F}; \mathbb{R}^{d+l}) \Longleftrightarrow \mathbb{P}_W \in \mathcal{P}_2(\mathbb{R}^{d+l})$$

We use $\bar{f}(w, \theta), \bar{L}(w, \theta), \bar{\Phi}(w)$ to denote corresponding functions in the extended $(d+l)$-dimensional space (e.g. $\bar{\Phi}(w) := \Phi(x, y)$).

Notation (cont.)

Given $\xi \in L^2(\mathcal{F}, \mathbb{R}^{d+l})$ and a control process $\boldsymbol{\theta} \in L^\infty([0,T], \Theta)$, we consider the following dynamic system for $t \le s \le T$:

$$W_s^{t,\xi,\boldsymbol{\theta}} = \xi + \int_t^s \bar{f}(W_\tau^{t,\xi,\boldsymbol{\theta}}, \theta_\tau)\, d\tau.$$

Let $\mu = \mathbb{P}_\xi \in \mathcal{P}_2(\mathbb{R}^{d+l})$, we denote the law of $W_s^{t,\xi,\boldsymbol{\theta}}$ for simplicity by

$$\mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} := \mathbb{P}_{W_s^{t,\xi,\boldsymbol{\theta}}}.$$

Notation (cont.)

Given $\xi \in L^2(\mathcal{F}, \mathbb{R}^{d+l})$ and a control process $\boldsymbol{\theta} \in L^\infty([0,T], \Theta)$, we consider the following dynamic system for $t \le s \le T$:

$$W_s^{t,\xi,\boldsymbol{\theta}} = \xi + \int_t^s \bar{f}(W_\tau^{t,\xi,\boldsymbol{\theta}}, \theta_\tau)\, d\tau.$$

Let $\mu = \mathbb{P}_\xi \in \mathcal{P}_2(\mathbb{R}^{d+l})$, we denote the law of $W_s^{t,\xi,\boldsymbol{\theta}}$ for simplicity by

$$\mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} := \mathbb{P}_{W_s^{t,\xi,\boldsymbol{\theta}}}.$$

In the sequel, we assume

- (A1″) f, L, Φ is bounded; f, L, Φ are Lipschitz continuous with respect to x, and the Lipschitz constants of f and L are independent of θ.
- (A2″) $\mu \in \mathcal{P}_2(\mathbb{R}^{d+l})$.

Continuity of Value Function and Mean-Field DPP

We rewrite the time-dependent objective functional and value function as

$$J(t,\mu,\boldsymbol{\theta}) = \langle \bar{\Phi}(.),\, \mathbb{P}_T^{t,\mu,\boldsymbol{\theta}} \rangle + \int_t^T \langle \bar{L}(.,\boldsymbol{\theta}_s),\, \mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} \rangle \, ds,$$

$$v^*(t,\mu) = \inf_{\boldsymbol{\theta} \in L^\infty([0,T],\Theta)} J(t,\mu,\boldsymbol{\theta}).$$

Continuity of Value Function and Mean-Field DPP

We rewrite the time-dependent objective functional and value function as

$$J(t,\mu,\boldsymbol{\theta}) = \langle \bar{\Phi}(.),\, \mathbb{P}_T^{t,\mu,\boldsymbol{\theta}} \rangle + \int_t^T \langle \bar{L}(.,\boldsymbol{\theta}_s),\, \mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} \rangle \, ds,$$

$$v^*(t,\mu) = \inf_{\boldsymbol{\theta} \in L^\infty([0,T],\Theta)} J(t,\mu,\boldsymbol{\theta}).$$

Theorem (Lipschitz continuity of value function)

The function $(t,\mu) \mapsto J(t,\mu,\boldsymbol{\theta})$ is Lipschitz continuous on $[0,T] \times \mathcal{P}_2(\mathbb{R}^{d+l})$, uniformly with respect to $\boldsymbol{\theta}$, and the value function $v^*(t,\mu)$ is Lipschitz continuous on $[0,T] \times \mathcal{P}_2(\mathbb{R}^{d+l})$.

Continuity of Value Function and Mean-Field DPP

We rewrite the time-dependent objective functional and value function as

$$J(t,\mu,\boldsymbol{\theta}) = \langle \bar{\Phi}(.),\, \mathbb{P}_T^{t,\mu,\boldsymbol{\theta}} \rangle + \int_t^T \langle \bar{L}(.,\boldsymbol{\theta}_s),\, \mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} \rangle \, ds,$$

$$v^*(t,\mu) = \inf_{\boldsymbol{\theta} \in L^\infty([0,T],\Theta)} J(t,\mu,\boldsymbol{\theta}).$$

Theorem (Lipschitz continuity of value function)

The function $(t,\mu) \mapsto J(t,\mu,\boldsymbol{\theta})$ is Lipschitz continuous on $[0,T] \times \mathcal{P}_2(\mathbb{R}^{d+l})$, uniformly with respect to $\boldsymbol{\theta}$, and the value function $v^*(t,\mu)$ is Lipschitz continuous on $[0,T] \times \mathcal{P}_2(\mathbb{R}^{d+l})$.

Theorem (Mean-field DPP)

For all $0 \le t \le \hat{t} \le T$, $\mu \in \mathcal{P}_2(\mathbb{R}^{d+l})$, we have

$$v^*(t,\mu) = \inf_{\boldsymbol{\theta} \in L^\infty([0,T],\Theta)} \Big[\int_t^{\hat{t}} \langle \bar{L}(.,\boldsymbol{\theta}_s),\, \mathbb{P}_s^{t,\mu,\boldsymbol{\theta}} \rangle \, ds + v^*(\hat{t}, \mathbb{P}_{\hat{t}}^{t,\mu,\boldsymbol{\theta}}) \Big].$$

Derivative in Wasserstein Space

To define derivative w.r.t. measure, we lift function $u : \mathcal{P}_2(\mathbb{R}^{d+l}) \to \mathbb{R}$ into its "extension" $U : L^2(\mathcal{F}; \mathbb{R}^{d+l}) \to \mathbb{R}$ by

$$U[X] = u(\mathbb{P}_X), \quad \forall X \in L^2(\mathcal{F}; \mathbb{R}^{d+l}).$$

If U is Fréchet differentiable, we can define

$$\partial_\mu u(\mathbb{P}_X)(X) = DU(X),$$

for some function $\partial_\mu u(\mathbb{P}_X) : \mathbb{R}^{d+l} \to \mathbb{R}^{d+l}$.

Derivative in Wasserstein Space

To define derivative w.r.t. measure, we lift function $u : \mathcal{P}_2(\mathbb{R}^{d+l}) \to \mathbb{R}$ into its "extension" $U : L^2(\mathcal{F}; \mathbb{R}^{d+l}) \to \mathbb{R}$ by

$$U[X] = u(\mathbb{P}_X), \quad \forall X \in L^2(\mathcal{F}; \mathbb{R}^{d+l}).$$

If U is Fréchet differentiable, we can define

$$\partial_\mu u(\mathbb{P}_X)(X) = DU(X),$$

for some function $\partial_\mu u(\mathbb{P}_X) : \mathbb{R}^{d+l} \to \mathbb{R}^{d+l}$.

Given a smooth $u : \mathcal{P}_2(\mathbb{R}^{d+l}) \to \mathbb{R}$ and the following dynamic system,

$$W_t = \xi + \int_0^t \bar{f}(W_s)\, ds, \quad \xi \in L^2(\mathcal{F}; \mathbb{R}^{d+l}),$$

we have the chain rule

$$u(\mathbb{P}_{W_t}) = u(\mathbb{P}_{W_0}) + \int_0^t \langle \partial_\mu u(\mathbb{P}_{W_s})(.) \cdot \bar{f}(.), \, \mathbb{P}_{W_s} \rangle \, ds.$$

Infinite-Dimensional HJB Equation

Now we can write down the HJB equation, with $v(t,\mu)$ being the unknown solution,

$$\begin{cases} \dfrac{\partial v}{\partial t} + \inf\limits_{\theta_t \in \Theta} \Big\langle \partial_\mu v(t,\mu)(.) \cdot \bar{f}(.,\theta_t) + \bar{L}(.,\theta_t),\ \mu \Big\rangle = 0, & \text{on } [0,T) \times \mathcal{P}_2(\mathbb{R}^{d+l}), \\ v(T,\mu) = \langle \bar{\Phi}(.), \mu \rangle, & \text{on } \mathcal{P}_2(\mathbb{R}^{d+l}). \end{cases} \tag{1}$$

Infinite-Dimensional HJB Equation

Now we can write down the HJB equation, with $v(t,\mu)$ being the unknown solution,

$$\begin{cases} \dfrac{\partial v}{\partial t} + \inf\limits_{\theta_t \in \Theta} \Big\langle \partial_\mu v(t,\mu)(.) \cdot \bar{f}(.,\theta_t) + \bar{L}(.,\theta_t),\, \mu \Big\rangle = 0, & \text{on } [0,T) \times \mathcal{P}_2(\mathbb{R}^{d+l}), \\ v(T,\mu) = \langle \bar{\Phi}(.), \mu \rangle, & \text{on } \mathcal{P}_2(\mathbb{R}^{d+l}). \end{cases} \tag{1}$$

Theorem (Verification theorem)

Let v be function in $C^{1,1}([0,T] \times \mathcal{P}_2(\mathbb{R}^{d+l}))$. If v is a solution to (1) and there exists $\theta^*(t,\mu)$, a mapping $(t,\mu) \mapsto \theta$ attaining the infimum in (1), then $v(t,\mu) = v^*(t,\mu)$, and θ^* is the optimal feedback control.

Lifted HJB Equation

For convenience, we define the Hamiltonian $H(\mu, p) : \mathcal{P}^2(\mathbb{R}^{d+l}) \times L^2_\mu(\mathbb{R}^{d+l}) \to \mathbb{R}$ as

$$H(\mu, p) \coloneqq \inf_{\theta \in \Theta} \left\langle p(.) \cdot \bar{f}(., \theta) + \bar{L}(., \theta),\, \mu \right\rangle .$$

Lifted HJB Equation

For convenience, we define the Hamiltonian $H(\mu, p) : \mathcal{P}^2(\mathbb{R}^{d+l}) \times L^2_\mu(\mathbb{R}^{d+l}) \to \mathbb{R}$ as

$$H(\mu, p) \coloneqq \inf_{\theta \in \Theta} \left\langle p(.) \cdot \bar{f}(., \theta) + \bar{L}(., \theta),\ \mu \right\rangle .$$

Then the original HJB can be rewritten as

$$\begin{cases} \dfrac{\partial v}{\partial t} + H(\mu, \partial_\mu v(t, \mu)) = 0, & \text{on} \ \ [0, T) \times \mathcal{P}_2(\mathbb{R}^{d+l}), \\ v(T, \mu) = \langle \bar{\Phi}(.), \mu \rangle, & \text{on} \ \ \mathcal{P}_2(\mathbb{R}^{d+l}). \end{cases}$$

The “lifted” Bellman equation is formally like above except that the state space is enlarged

$$\begin{cases} \dfrac{\partial V}{\partial t} + \mathcal{H}(\xi, DV(t, \xi)) = 0, & \text{on} \ \ [0, T) \times L^2(\mathcal{F}; \mathbb{R}^{d+l}), \\ V(T, \xi) = \mathbb{E}[\bar{\Phi}(\xi)], & \text{on} \ \ L^2(\mathcal{F}; \mathbb{R}^{d+l}). \end{cases}$$

Viscosity Solution: "Weak" Solution of PDE

Intuition: use monotonicity of the value function and sidestep non-differentiability through the test function

Definition

We say that a bounded, uniformly continuous function u is a viscosity subsolution (supersolution) to the original HJB equation (1) if the lifted function U defined by $U(t,\xi) = u(t, \mathbb{P}_\xi)$ is a viscosity subsolution (supersolution) to the lifted Bellman equation, that is

$$U(T,\xi) \leq (\geq) \mathbb{E}[\bar{\Phi}(\xi)],$$

and for any test function $\psi \in C^{1,1}([0,T] \times L^2(\mathcal{F};\mathbb{R}^{d+l}))$ such that the map $U - \psi$ has a local maximum (minimum) at $(t_0,\xi_0) \in [0,T) \times L^2(\mathcal{F};\mathbb{R}^{d+l})$, one has

$$\partial_t \psi(t_0,\xi_0) + \mathcal{H}(\xi_0, D\psi(t_0,\xi_0)) \geq (\leq) 0.$$

Existence and Uniqueness

Theorem (Existence)

The value function $v^*(t, \mu)$ is a viscosity solution to the HJB equation (1).

Existence and Uniqueness

Theorem (Existence)

The value function $v^*(t, \mu)$ is a viscosity solution to the HJB equation (1).

Theorem (Uniqueness)

Let u_1 and u_2 be viscosity subsolution and supersolution to (1) respectively. Then $u_1 \leq u_2$. Consequently, the value function $v^*(t, \mu)$ is the unique viscosity solution to the HJB equation (1). In particular, if the Hamiltonian $H(\mu, p)$ is defined on a unique minimizer θ^*, then the optimal control process $\boldsymbol{\theta}^*$ is also unique.

Table of Contents

Summary

1. We introduced the mathematical formulation of the population risk minimization problem of continuous-time deep learning in the context of mean-field optimal control.

2. Mean-field Pontrayagin's maximum principle and mean-field dynamic programming principle (HJB equation) provide us new perspectives towards theoretical understanding of deep learning: uniqueness, generalization estimates in finite-sample case with explicit rate, etc. More to be developed.

Thank you for your attention!