جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Optimal control and machine learning

Diogo A. Gomes

www.kaust.edu.sa

Outline

The calculus of variations problem

Given $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, find a curve $\mathbf{x} : [0, T] \to \mathbb{R}^d$ with $\mathbf{x}(0) = x_I$, $\mathbf{x}(T) = x_T$ that minimizes

$$\int_0^T L(\mathbf{x}(s), \dot{\mathbf{x}}(s))ds.$$

Variational problems like this arise in multiple contexts

- Geodesics (shortest path)
- Brachistochrone problem (shortest time)
- Classical mechanics (minimal action principle)
- Elasticity...

Shortest path problem

The lenght of a curve $(\mathbf{x}, \mathbf{y})$ between two points in the plane, (x_I, y_I) and (x_T, y_T), is given by the integral

$$\int_0^T \sqrt{\dot{\mathbf{x}}(s)^2 + \dot{\mathbf{y}}(s)^2} ds.$$

Euler-Lagrange equation

Let $\mathbf{x} : [0, T] \to \mathbb{R}^d$ be a C^2 minimizer of the calculus of variations problem. Let $\mathbf{z}$ be a C^2 function with $\mathbf{z}(0) = \mathbf{z}(T) = 0$. Then

$$i(\epsilon) = \int_0^T L(\mathbf{x}(s) + \epsilon \mathbf{z}(s), \dot{\mathbf{x}}(s) + \epsilon \dot{\mathbf{z}}(s)) ds.$$

has a minimum when $\epsilon = 0$. Hence, $i'(0) = 0$.

Differentiating and setting $\epsilon = 0$, we get

$$\begin{aligned} 0 &= \int_0^T D_x L(\mathbf{x}(t), \dot{\mathbf{x}}(t))\mathbf{z}(t) + D_v L(\mathbf{x}(t), \dot{\mathbf{x}}(t))\dot{\mathbf{z}}(t) dt \\ &= \int_0^T \left[D_x L(\mathbf{x}(t), \dot{\mathbf{x}}(t)) - \frac{d}{dt} D_v L(\mathbf{x}(t), \dot{\mathbf{x}}(t)) \right] \mathbf{z}(t) ds \end{aligned}$$

Because $\mathbf{z}$ is arbitrary, we deduce the Euler-Lagrange equation

$$D_x L(\mathbf{x}(t), \dot{\mathbf{x}}(t)) - \frac{d}{dt} D_v L(\mathbf{x}(t), \dot{\mathbf{x}}(t)) = 0.$$

The shortest path may be a line

For the minimal length problem, if $\dot{\mathbf{x}}^2 + \dot{\mathbf{y}}^2 \neq 0$, the Euler-Lagrange equation reads

$$\begin{cases} \frac{d}{dt}\frac{\dot{\mathbf{x}}}{\sqrt{\dot{\mathbf{x}}^2+\dot{\mathbf{y}}^2}} = 0 \\ \frac{d}{dt}\frac{\dot{\mathbf{y}}}{\sqrt{\dot{\mathbf{x}}^2+\dot{\mathbf{y}}^2}} = 0. \end{cases}$$

Thus,

$$\frac{\dot{\mathbf{x}}}{\sqrt{\dot{\mathbf{x}}^2 + \dot{\mathbf{y}}^2}} = \cos\theta, \qquad \frac{\dot{\mathbf{y}}}{\sqrt{\dot{\mathbf{x}}^2 + \dot{\mathbf{y}}^2}} = \sin\theta,$$

for some constant angle θ. Hence, the velocity vector has constant directiont; that is, **if there is a C^2 trajectory of shortest length between two points**, it must be a straight line.

Questions

- is there in fact a C^2 trajectory of shortest length?
- is this trajectory unique?
- can we just prove directly that a straight line is the shortest lenght path

Convexity and optimality - real case

Consider a real-valued C^1 function, f.

- Any minimizer $\bar{x}$ of f is a critical point, $Df(\bar{x}) = 0$.
- If f is **convex** any critical point is a minimizer.

Proof: for any x and y, convexity gives

$$f(y) \geq f(x) + Df(x)(y - x).$$

If $x = \bar{x}$ is a critical point $Df(\bar{x}) = 0$ and so

$$f(y) \geq f(\bar{x}).$$

Convexity and optimality

Suppose L is convex. **Then any solution to the Euler-Lagrange equation is a minimizer.**

Convexity and optimality - proof

Let $\mathbf{x}$ solve the Euler-Lagrange equation, and $\mathbf{w}$ any trajectory with $\mathbf{x}(0) = \mathbf{w}(0)$, $\mathbf{x}(T) = \mathbf{w}(T)$. Then convexity gives

$$\int_0^T L(\mathbf{w}, \dot{\mathbf{w}}) \geq \int_0^T L(\mathbf{x}, \dot{\mathbf{x}}) + D_x L(\mathbf{x}, \dot{\mathbf{x}})(\mathbf{w} - \mathbf{x}) + D_v L(\mathbf{x}, \dot{\mathbf{x}})(\dot{\mathbf{w}} - \dot{\mathbf{x}}).$$

Because $\mathbf{x}$ solves the Euler-Lagrange equation, taking into account the boundary conditions and integrating by parts gives

$$\int_0^T (D_x L(\mathbf{x}, \dot{\mathbf{x}}) - \frac{d}{dt} D_v L(\mathbf{x}, \dot{\mathbf{x}}))(\mathbf{w} - \mathbf{x}) = 0.$$

But then

$$\int_0^T L(\mathbf{w}, \dot{\mathbf{w}}) \geq \int_0^T L(\mathbf{x}, \dot{\mathbf{x}}).$$

The shortest path is a line

- $L(\dot{\mathbf{x}}, \dot{\mathbf{y}}) = \sqrt{\dot{\mathbf{x}}^2 + \dot{\mathbf{y}}^2}$ is convex....

Key concepts and ideas

- First-order optimality conditions (Euler-Lagrange equations)
- Convexity and optimality

Key questions in calculus of variations

- Rigorous setting (functional spaces, technical conditions on L...)
- Existence of minimizers
- Necessary and sufficient conditions
- Solving Euler-Lagrange equations

Outline

The optimal control problem

The standard terminal value problem in control theory seeks to minimize

$$\int_0^T L(\mathbf{x}, \mathbf{a}) ds + \psi(\mathbf{x}(T))$$

under the constraint

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{a}).$$

The variable **x** is called the state and **a** is the control.

Linear-Quadratic control problem

A particular relevant problem is the linear quadratic problem, given matrices A, B, C, M, N, minimize

$$\int_0^T \mathbf{x}^T A\mathbf{x} + \mathbf{a}^T B\mathbf{a} ds + \mathbf{x}^T(T)C\mathbf{x}(T)$$

under the constraint

$$\dot{\mathbf{x}} = M\mathbf{x} + N\mathbf{a}.$$

Bolza vs Meyer problem

By applying a simple transformation is always possible to transform a control problem in either

- Bolza form: $\psi = 0$
- Meyer form $L = 0$.

Necessary optimality conditions

In optimal control, the Euler-Lagrange equations are replaced by the Pontryagin maximum principle. Let $(\mathbf{x}, \mathbf{a})$ be an optimal trajectory/control. Then

$$\begin{cases} \dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{a}) \\ \dot{\mathbf{p}} = -\mathbf{p} D_x f(\mathbf{x}, \mathbf{a}) - D_x L(\mathbf{x}, \mathbf{a}) \end{cases}$$

and $\mathbf{a}$ maximizes $-\mathbf{p} f(\mathbf{x}, \mathbf{a}) - L(\mathbf{x}, \mathbf{a})$. Moreover, $\mathbf{p}(T) = D_x \psi(\mathbf{x}(T))$.

Sufficiency of PMP

Suppose L is uniformly convex and f is affine and terminal cost vanishes.

- $L(\tilde{x}, \tilde{a}) \geq L(\mathbf{x}, \mathbf{a}) + D_x L(\mathbf{x}, \mathbf{a})(\tilde{x} - \mathbf{x}) + D_a L(\mathbf{x}, \mathbf{a})(\tilde{a} - \mathbf{a})$
- There exists a unique $\mathbf{a}$ maximizing $-\mathbf{p} f(\mathbf{x}, \mathbf{a}) - L(\mathbf{x}, \mathbf{a})$
- $D_a L(\mathbf{x}, \mathbf{a}) = -\mathbf{p} D_a f(\mathbf{x}, \mathbf{a})$
- $f(\mathbf{x}, \mathbf{a}) - f(\tilde{x}, \tilde{a}) = D_x f(\mathbf{x}, \mathbf{a})(\tilde{x} - \mathbf{x}) + D_a f(\mathbf{x}, \mathbf{a})(\tilde{a} - a)$
- $\mathbf{p}(T) = 0$

Sufficiency of PMP

(Suppose L is convex and f is affine and terminal cost vanishes.) Consider a triplet satisfying PMP, $(\mathbf{x}, \mathbf{a}, \mathbf{p})$, let $(\tilde{x}, \tilde{a})$ be a competing trajectory. Then

$$\int_0^T L(\tilde{x}, \tilde{a}) \geq \int_0^T L(\mathbf{x}, \mathbf{a}) + D_x L(\mathbf{x}, \mathbf{a})(\tilde{x} - \mathbf{x}) + D_a L(\mathbf{x}, \mathbf{a})(\tilde{a} - \mathbf{a})$$

$$\geq \int_0^T -(\dot{\mathbf{p}} + \mathbf{p} D_x f(\mathbf{x}, \mathbf{a}))(\tilde{x} - \mathbf{x}) - \mathbf{p} D_a f(\mathbf{x}, \mathbf{a})(\tilde{a} - a)$$

$$= \int_0^T \mathbf{p}(f(\mathbf{x}, \mathbf{a}) - f(\tilde{x}, \tilde{a})) - \mathbf{p} D_x f(\mathbf{x}, \mathbf{a})(\tilde{x} - \mathbf{x}) - \mathbf{p} D_a f(\mathbf{x}, \mathbf{a})(\tilde{a} - a)$$

$$= 0$$

Hamiltonian dynamics

If $f(x, v) = v$, then

$$H(x, p) = \sup_v -pv - L(x, v)$$

is called the Hamiltonian. Then, $D_p H = -v$ and $D_x H = -D_x L(x, v)$. Thus, the necessary optimality conditions become

$$\begin{cases} \dot{\mathbf{x}} = -D_p H(\mathbf{x}, \mathbf{p}) \\ \dot{\mathbf{p}} = D_x H(\mathbf{x}, \mathbf{p}), \end{cases}$$

the Hamiltonian dynamics, which is equivalent to the Euler-Lagrange equation (**exercise).**

LQ case

In the particular LQ case

$$L = \frac{|x|^2}{2} + \frac{|v|^2}{2} \qquad f(x, v) = v,$$

we have $H = \frac{|p|^2}{2} - \frac{|x|^2}{2}$

$$\begin{cases} \dot{\mathbf{x}} = -\mathbf{p} \\ \dot{\mathbf{p}} = -\mathbf{x}; \end{cases}$$

that is,

$$\ddot{\mathbf{x}} = \mathbf{x}.$$

Key concepts and ideas

- Pontryagin maximum principle
- LQ control problems

Key questions in optimal control

- Existence of optimal trajectories
- Computation of optimal trajectories
- Controllability issues

Outline

The value function

To simplify, consider the calculus of variations setting. The value function is

$$u(x,t) = \inf \int_t^T L(\mathbf{x}, \dot{\mathbf{x}}) ds + \psi(\mathbf{x}(T)),$$

where the infimum is taken over all C^1 trajectories $\mathbf{x} : [t, T] \to \mathbb{R}^d$ with $\mathbf{x}(t) = x$.

Dynamic programming principle

For $h > 0$ such that $t + h < T$, we have

$$u(x,t) = \inf \int_t^{t+h} L(\mathbf{x}, \dot{\mathbf{x}}) ds + u(\mathbf{x}(t+h), t+h).$$

The Hamilton–Jacobi equation (formal)

We have

$$\begin{aligned} u(x,t) &= \inf_{\mathbf{x}} \int_t^{t+h} L(\mathbf{x},\dot{\mathbf{x}})ds + u(\mathbf{x}(t+h),t+h) \\ &\simeq \inf_{\dot{\mathbf{x}}(t)} hL(x,\dot{\mathbf{x}})ds + u(\mathbf{x}(t),t) + hD_x u(x,t)\dot{\mathbf{x}}(t) + hu_t(x,t) \\ &= u(x,t) - hH(x,D_x u(x,t)) + hu_t(x,t) \end{aligned}$$

from which we deduce the Hamilton-Jacobi equation

$$-u_t + H(x, D_x u(x,t)) = 0.$$

Verification theorem

Suppose L is uniformly convex. Let u solve

$$-u_t + H(x, D_x u(x,t)) = 0$$

with $u(x,T) = \psi(x)$. Then, u is the value function. Moreover, the optimal dynamics is

$$\dot{\mathbf{x}} = -D_p H(\mathbf{x}, D_x u(\mathbf{x}(t), t)).$$

Riccati equation and the LQ problem

For our LQ example problem, $H = \frac{p^2}{2} - \frac{x^2}{2}$, the Hamilton-Jacobi equation becomes

$$-u_t + \frac{u_x^2}{2} - \frac{x^2}{2} = 0$$

which admits quadratic solutions of the form $u(x,t) = \alpha(t)x^2$ where α solves the Riccati equation

$$-\dot{\alpha} + 2\alpha^2 - \frac{1}{2} = 0.$$

Lack of smooth solutions for HJ equations

Unfortunately Hamilton-Jacobi may fail to admit solutions.
Consider the HJ equation

$$-u_t + \frac{u_x^2}{2} = 0.$$

Let $v = u_x$. Then

$$-v_t + v v_x = 0.$$

Consider the ODE

$$\dot{\mathbf{x}} = -v(\mathbf{x}, t).$$

Then, v is constant along $\mathbf{x}$ because

$$\frac{d}{dt} v(\mathbf{x}, t) = v_t(\mathbf{x}, t) + v_x(\mathbf{x}, t)\dot{\mathbf{x}} = 0.$$

But different initial conditions can cross...

Key concepts and ideas

- Value function and dynamic programming principle
- Hamilton–Jacobi equation

Key questions in optimal control

- Solution of the Hamilton–Jacobi equation
- Extended notions of solution (viscosity solutions)

Outline

The (deterministic) reinforcement learning setting

- Agents have a state s and a possible action a
- In state s, the action a corresponds to a reward $r(s, a)$
- Actions change the state to a new state according to the discrete dynamics $s_{n+1} = f(s_n, a_n)$
- Rewards are discounted in time (now is better than later) by a parameter $0 < \theta < 1$
- Agents want to maximize the long-term reward

$$\sum_{i=1}^{\infty} \theta^i r(s_i, a_i)$$

This is a control problem!

Applications of reinforcement learning

- Solving sames (chess, go, ...)
- Finding the shortest path
- Traffic light control
- Bidding, advertising, personalized recommendations....
- Theorem proving

Value function

We define the value function

$$Q(s) = \inf_{a_i} \sum_{i=1}^{\infty} \theta^i r(s_i, a_i)$$

Dynamic programming/discrete HJ equation

We have the

$$Q(s) = \sup_a \left[r(s,a) + \theta Q(f(s,a)) \right].$$

Learning the value function

- The key problem in reinforcement learning is to approximate the value function, usually by iterative methods
- A popular method is the Q-learning algorithm

Q-learning

Let $0 < \gamma < 1$ (learning rate). Given a an approximation Q^n, choose a state s and let

$$Q^{n+1}(s) = Q^n(s) + \gamma \sup_a \left[r(s,a) + \theta Q^n(f(s,a)) - Q^n(s) \right].$$

The continuous analog

A continuous analog to Q learning is

$$u_t = \alpha u + H(x, Du).$$

The convergence of Q learning is replaced by the convergence as $t \to \infty$ of $u(x, t)$.

Approximation of the value function

- If the state space is very large (think all possible positions in chess), the function may not be representable in a computer.
- In this case, the value function must to be approximated. For example as a linear combination of feature maps

$$V(s) = \sum w_i \phi_i(s)$$

- This is similar to what chess players players use, queen=10, rook=5, bishop and kingt=3, pawn =1...
- Alternatively deep neural networks can be used to approximate the value function.

Key concepts and ideas

- Dynamic programming principle and Belmann equation
- Q-learning algorithm

Key questions in reinforcement learning

- Approximation of value function: features vs deep NN
- Training algorithms

Outline

Monge problem

- "How best to move a pile of soil or rubble to an excavation or fill with the least amount of work."

Probability measures

- If you know what a measure is, you know what a probability measure is.
- If you don't know measure theory, today, a probability measure is a non-negative function that integrates to 1.

Push-forward and transport of measures

Given a map $T : \mathbb{R}^d \to \mathbb{R}^d$ and a probability measure m in $\mathbb{R}^d$, the pushforward of m by T, $T\sharp m$, is the probability measure given by

$$\int f(y)(T\sharp m)(y)dy = \int f(Tx)m(x)dx.$$

Optimal transport

The Monge problem can be formulated as follows. Given two probability measures m_0 and m_1, find a map T that

- $T\sharp m_0 = m_1$
- T minimizes $\int |x - T(x)|^2 dm$ among all possible maps that satisfy the preceding condition.

- $T\sharp m_0 = m_1$ means for all f

$$\int f(T(x))m_0(x) = \int f(y)m_1(y)$$

- By the change of variables formula

$$\int f(y)m_1(y) = \int f(T(x))m_1(T(x)) \det T$$

- Accordingly

$$m_0(x) = m_1(T(x)) \det T.$$

Monge-Ampère equation

- It turns out that if m_0 and m_1 are positive smooth functions then $T = Du(x)$ for some function u
- By the change of variables formula, u satisfies the Monge-Ampere equation

$$m_1(\nabla u) \det D^2 u = m_0(x).$$

Lack of solutions

- The Monge problem may not have a solution for singular measures.
- For example, there is no map T that transports δ_0 into $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$.
- The Kantorowich problem is a relaxation of Monge's problem to address this lack of solutions.

Given a map that transports m_0 into m_1, we build the probability measure π in $\mathbb{R}^{2d}$ as follows

$$\int \phi(x,y)d\pi = \int \phi(x, T(x))dm_0.$$

In particular,

$$\int_{\mathbb{R}^{2d}} \varphi(x)d\pi = \int_{\mathbb{R}^d} \varphi(x)dm_0 \qquad \int_{\mathbb{R}^{2d}} \varphi(y)d\pi = \int_{\mathbb{R}^d} \varphi(y)dm_1.$$

In the Kantorowich problem, the mass at a point x is sent not to a point $T(x)$ but to a distributed plan according to $\pi(x,y)$.

Kantorowich problem

Find a probability measure $\pi(x, y)$ that minimizes

$$\int_{\mathbb{R}^{2d}} |x - y|^2 d\pi$$

under the marginal constraints

$$\int_{\mathbb{R}^{2d}} \varphi(x) d\pi = \int_{\mathbb{R}^{d}} \varphi(x) dm_0 \qquad \int_{\mathbb{R}^{2d}} \varphi(y) d\pi = \int_{\mathbb{R}^{d}} \varphi(y) dm_1.$$

Wasserstein distance

- The solution π^* to the previous problem is called an optimal mass transfer plan.
- The 2-Wasserstein distance between m_0 and m_1 is

$$W_2^2(m_0, m_1) = \int |x - y|^2 d\pi^*.$$

- The 2-Wassersstein distance are often better to measure distances between probability measures than the L^p distances

$$\|m_0 - m_1\|_{L^p}^p = \int |m_0(x) - m_1(x)|^p.$$

- p-Wasserstein distances are defined analogously.

Wasserstein and integration

Let f be a Lipschitz function, m_0 and m_1 probability measures. Then

$$\left| \int fm_0 - \int fm_1 \right| = \left| \int f(x) - f(y) d\pi^* \right|$$
$$\leq C \int |x - y| d\pi^* \leq CW_1(m_0, m_1).$$

Data and probability measures

Often in machine learning, data is

- Data is a (large) collection of points in $\mathbb{R}^d$, x_i, sampled independently from a common distribution (often unknown) m
- Data can be identified with the empirical measure

$$\frac{1}{N}\sum_i \delta_{x_i}$$

- Empirical measures approximate (eg in Wasserstein sense) the common distribution.

More distances on probability measures

There a number of useful distances on probability measures that sometimes are simpler to compute that Wasserstein. One of them is the Kullback Leibler divergence

$$\mathcal{D}_{KL}(m_0, m_1) = \int m_0 \log \frac{m_0}{m_1}.$$

Generative problem

Given a reference probability distribution m_0 (let's say a Gaussian) and a target distribution m_1 (let's say images of persons). We would like to find a map T

$$T\sharp m_0 = m_1.$$

Thus "a random point sampled from m_0 becames a random image of a person $T(x)$ sampled according to m_1".

- We can think as the points sampled according to m_0 as a vector of features
- The map T transforms features into images.

The generative models

The previous problem is intractable because m_1 is often not known, rather only samples are available. Furthermore, the set of all maps is too large. Rather, we fix an admissble set of maps $\mathcal{A}$ and seek to find

$$\min_{T\in\mathcal{A}} \mathcal{D}_{KL}(T\sharp m_0, m_1).$$

- Neural networks provide a good way to construct large classes of maps $\mathcal{A}$
- $\mathcal{D}_{KL}$ is computed through sampling methods.

Outline

Calculus of variations

Optimal control

Hamilton-Jacobi equations

Reinforcement learning

Variational problems in the space of maps

Infinite dimensional control

Control formulation of (deep) supervised learning

Control formulation of (deep) unsupervised learning

Further mathematical issues

Control in the space of maps

- A velocity field (control), $v(x, t)$ induces a trajectory in the space of maps (flow)

$$\dot{T}(x; t) = v(T(x; t), t).$$

- We seek to find v that minimizes

$$\int_0^T L(T, v)dt + \psi(T(\cdot, T)).$$

- This is an infinite dimensional control problem! **Here L takes a map and a vector field and returns a real number, so L is not a function in $\mathbb{R}^d$!**

Benamou-Brenier formulation of optimal transport

Given m_0 and m_1, find velocity field (control), $v(x,t)$ that:

- $T(\cdot,1)\sharp m_0 = m_1$.
- minimizes

$$\int_0^1 \int \frac{|v(T(x;t),t)|^2}{2} m_0(x)dxdt$$

It turns out this problem is equivalent to the optimal transport problem and $T(\cdot,1)$ is an optimal transport map.

Euler-Arnold variational problem

Given a Lebesgue measure-preserving map $\bar{T}$, find a **divergence-free** velocity field (control), $v(x,t)$ that:

- $T(\cdot,1)=\bar{T},\ T(\cdot,0)=I$
- minimizes
$$\int_0^1\int \frac{|v(T(x;t),t)|^2}{2}\,dxdt,$$
among all such divergence-free velocities.

It turns out a solution to this variational problem solves the Euler equation in fluid mechanics

$$v_t + v\nabla v = \nabla p \qquad \operatorname{div} v = 0.$$

Deep learning as a (discrete) control problem

- Deep learning is a discrete control problem in spaces of maps.
- A layer is a parametrized map N_{θ_i}
- The goal in deep learning is to choose m parameters such that the map

$$T = N_{\theta_m} \circ \dots N_{\theta_1}$$

minimizes some functional.

Outline

Supervised learning problem - data-centered formulation

Given pairs $(x_i, y_i) \in \mathbb{R}^{2d}$ find a map

$$T = N_{\theta_m} \circ \ldots N_{\theta_1}$$

that minimizes

$$\sum |y_i - T(x_i)|^2.$$

Supervised learning problem - abstract version

Given a probability measure μ in $\mathbb{R}^{2d}$ find a transformation

$$T = N_{\theta_m} \circ \dots N_{\theta_1}$$

that minimizes

$$\int |y - T(x)|^2 d\mu(x, y).$$

Deep linear learning

We consider the **linear** neural networks

$$N_{b_n}(x) = Ax + b_n$$

Then, the composition

$$T = N_{b_m} \circ \ldots \circ N_{b_1}(x)$$

gives the flow map corresponding to

$$x_{n+1} = Ax_n + b_n.$$

LQ control problem - recap

Consider the discrete dynamics with control b_n

$$x_{n+1} = Ax_n + b_n.$$

and consider the problem of minimizing

$$|x_m|^2$$

(here, $L = 0$).

LQ control problem as a deep learning (supervised) problem

Let $N_{b_n}(x) = Ax + b_n$, and consider the linear NN

$$T = N_{b_m} \circ \ldots \circ N_{b_1}(x),$$

to minimize

$$\int |T[x] - y|^2 \delta_{x_0}(x) \delta_0(y) dxdy.$$

This is equivalent to minimize $|T(x_0)|^2 = |x_m|^2$.

ResNets

A Residual Network (ResNet) is a NN architecure of the form

$$N_\theta(x) = x + hf(x, \theta).$$

For example

$$N_{A,b}(x) = x + h\sigma(Ax + b)$$

Where $\sigma(z) = z^+$ (taken coordinatewise).

Resnet continuous limit

The ResNet dynamics is

$$x_{n+1} = x_n + h\sigma(A_n x_n + b_n)$$

which is the Euler discretization of the ODE

$$\dot{\mathbf{x}} = \sigma(A(t)\mathbf{x} + b(t)).$$

Resnet limit

Given the controls $A(t)$ and $b(t)$ the limit map is determined by

$$\frac{d}{dt}T(x;t) = \sigma(A(t)T(x;t) + b(t)).$$

Supervised learning problem

Find $A(t)$ and $b(t)$ that minimize

$$\int |y - T(x,1)|^2 d\mu(x,y).$$

with

$$\frac{d}{dt}T(x;t) = \sigma(A(t)T(x;t) + b(t)),$$

and $T(x,0) = x$.

Key concepts and ideas

- Supervised learning can be set up as control problem in spaces of maps
- Resnets are particularly suitable to obtain continuous limits

Key questions in supervised learning

- Controllability (what maps can be generated by a particular architecture)
- Convergence
- Approximation - can we use continuous limits to design better neural networks.

Outline

Unsupervised learning - clustering

The clustering problem can set up as follows. Find a map

$$T : \mathbb{R}^d \to \{1, \dots, n\}$$

that minimizes some measure of dissimilarity among data.

Example - center of mass clustering

Given a probability measure in $\mathbb{R}^d$ find

$$\min_T \sum_i \int_{\Omega_i} |x - \bar{x}_i|^2 d\mu(x),$$

where $\Omega_i = T^{-1}(i)$ and

$$\bar{x}_i = \frac{\sum_i \int_{\Omega_i} x d\mu}{\sum_i \int_{\Omega_i} 1 d\mu}.$$

Auto-Encoders

Given a probability measure μ in $\mathbb{R}^d$ find two maps $T_E : \mathbb{R}^d \to \mathbb{R}^r$ and $T_D : \mathbb{R}^r \to \mathbb{R}^d$ that minimize

$$\int |x - T_D \circ T_E(x)|^2 d\mu.$$

When $r < d$ the encoder map T_E provides a low-dimensional representation of the data.

Key concepts and ideas

- Many unsupervised learning can be set up as variational problems or control problems in the space of maps.
- Measures of dissimilarity are functionals on spaces of maps.

Key questions in unsupervised learning

- Good measures of dissimilarity for clustering
- Choice of admissible classes of maps

Outline

Choice of architectures

- Approximation properties (finite-element spaces and piecewise linear NN)
- Group invariance (eg translation invariance in convolutional NN)

Training

- Training refers to the process of finding the optimal set of parameters
- Stochastic gradient descent and its variants seem to be the tool of choice. But better global minimizers may be possible.

Temporal structure

- Here, our data was taken as points in $\mathbb{R}^d$. But there are other interesting classes of data that are important in applications, for example, infinite sequences.
- Probability concepts such as independence, Markov property, ergodicity, are of great relevance for the formulation of the problems.
- Structure of NN must take into account the data structure (eg recurrent NN preserve non-anticipatory character)

Applications in PDEs

- Dataless training is also an area of great interest to solving high-dimensional PDEs where few numerical methods are available.
- This is often formulated as a minimization problem such as

$$\min_{T} \|F(x, T(x), DT(x), \ldots)\|$$

- Here, the choice of the norm is crucial and new questions arise - for example, what are good choices of the norm $\|\cdot\|$ that ensure that whever the previous minimization problem gives a small number we have that T is close to the solution of the problem

$$F(x, u, Du, \ldots) = 0.$$

Implementation issues

- We did not discuss any implementation issues, but this is in fact a crucial matter in applications.
- Neural networks and machine learning exist for quite a while but only became popular once powerful enough computers were available and flexible implementations (Keras, Tensor flow, ...) were built.

Take home message

- Many problems in machine learning are variational or control problems similar to well known and well studied mathematical problems.
- Calculus of variations and control theory give important insights in understanding machine learning problems

Further references (just names and highlights)

- Jinchao Xu (KAUST, AMCS) - approximation properties of Neural Networks
- Peter Markowich (KAUST, AMCS) - ResNets and control theory
- Peter Richtarik (KAUST, CS/AMCS) - Stochastic gradient descent, optimization...
- J. Schmidtuber (KAUST) - foundations of machine learning
- E. Zuazua - connection with control theory, see recent works
- Weinan E - high-dimensional PDE, dynamical system approach
- Carola Schönlieb and co-authors.... dynamical systems based NN, structure preservation