Neural networks and PDE

Benny Avelin
(J. Work with K. Nyström)

The 6th Uppsala University - Tokyo Tech Joint Symposium

Outline

1. Motivation
2. Connection to PDE
3. Some results

Motivation (What is machine learning)

Three types of problems

- Supervised Learning
  - Learning with a teacher
  - **Ex:** Regression / Classification
- Unsupervised Learning
  - Learning representations
  - **Ex:** Density estimation, dimensionality reduction, etc.
- Reinforcement Learning
  - Learning with a critic
  - **Ex:** Optimal control

Supervised Learning

- Classification
  - Image classification: x-image, y-class. Could be object identification like saying ‘this is the image of a cat’.
  - Text classification: Given a snippet of text, what is its subject?
- Regression
  - What is the weight of a person given the height? x-height, y-length.
  - Object location: Given that you have an image with a ball in it, where in the image is the ball.

Risk and Hypothesis

- Let us consider data $(x, y) \sim \mu$, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.
- A hypothesis is a function $h : \mathbb{R}^n \to \mathbb{R}^m$,
- A loss-function $L : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$,

$$R(h) = \mathbb{E}_\mu \left[L(h(x), y) \right], \quad \textbf{Risk}$$

- Given a data-set $D = \{(x_1, y_1), \dots (x_N, y_N)\}$ which are sampled i.i.d. from μ we also define,

$$R_{emp,D}(h) = \frac{1}{N} \sum_{i=1}^{N} \left[L(h(x_i), y_i) \right], \quad \textbf{Empirical Risk}$$

- Call a set of hypothesis $\mathcal{H}$, the hypothesis space.
- test

Risk and Hypothesis

- Find $h^* \in \mathcal{H}$ such that,

$$R(h^*) = \min_{\mathcal{H}} R(h), \quad \textbf{Risk minimization}$$

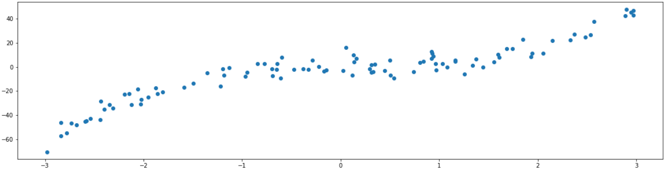
- We dont have access to μ but we have access to a given data-set D, we could try to find $h^*_D \in \mathcal{H}$ such that,

$$R_{emp,D}(h^*_D) = \min_{\mathcal{H}} R_{emp,D}(h)$$

- We cannot find h^*_D in general. Instead we try to find $h \in \mathcal{H}$ such that $R_{emp,D}(h)$ is as small as possible,

$$R_{emp,D}(h^*_D) \leq R_{emp,D}(h), \quad \textbf{Empirical Risk Min}$$

Simple example



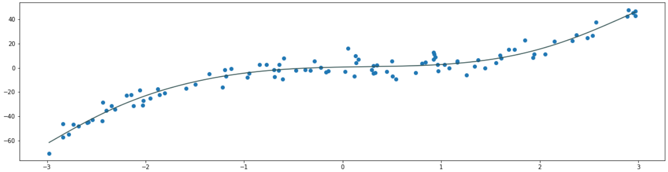
- Let $\mathcal{H}$ be functions of the form

$$h(x) = \sum_{i=1}^{M} v_i \sigma(w_i \cdot x + b_i)$$

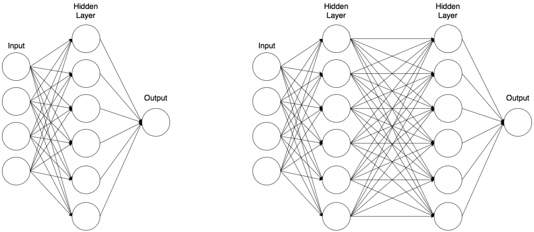
  for paramters v_i, w_i, b_i. $\sigma(x) = \frac{1}{1+e^{-x}}$
- Let the loss function be quadratic $L(x, y) = (x - y)^2$
- Goal: find h_D^* that minimizes

$$R_{emp,D}(h) = \frac{1}{N} \sum_{i=1}^{N} (h(x_i) - y_i)^2$$

Simple Example (Neural networks)



- $h(x)$ is actually a 'single hidden layer neural network'



$$h(x) = \sigma\big(W^2(\sigma(W^1 x + B^1)) + B^2\big)$$

How do we minimize risk?

We run a discrete form of gradient flow on the space of weights

$$dW_t = -\nabla_W R_{emp,D}(h_{W_t})dt$$

- W is usually very high dimensional 1M and up for many problems
- the size of D is also quite big.
- run the following discrete process instead

$$\Delta W_i = -\nabla_W R_{emp,D_i}(h_{W_i})\Delta n$$

- D_i is a subsampled set of D at each time step i, Δn is step-length.
- Called Stochastic Gradient Descent (SGD), or Robins-Monro stochastic approximation.

What are the dynamics of W_i?

- $\nabla_W R_{emp,D_i}(h_W)$ is an unbiased estimate of true gradient $\nabla_W R_{emp,D}(h_W)$.

$$\Delta W_i = -\nabla_W R_{emp,D}(h_{W_i})\Delta n + (\nabla_W R_{emp,D}(h_{W_i}) - \nabla_W R_{emp,D_i}(h_{W_i}))\Delta_n$$

- Identify this as a Euler-Maruyama scheme for the SDE

$$dW_t = -\nabla_W R_{emp,D}(h_{W_t})dt + \sqrt{\Delta n \Sigma(W_t)}dB_t$$

Observations

- If $\Delta n \to 0$ then we regain standard gradient flow.
- It is unclear what Σ actually is
- The density of W_t solves a Fokker-Planck equation.

Fokker-Planck

The density of

$$dW_t = -\nabla_W R_{emp,D}(h_{W_t})dt + \sqrt{\Delta n \Sigma(W_t)}dB_t$$

solves the Fokker planck equation (where $V(W) = R_{emp,D}(h_W)$)

$$\dot{\rho} = \nabla \cdot \left(\rho \nabla V + \frac{\Delta n}{2} \nabla \cdot (\Sigma \rho)\right)$$

Remember: W is high dimensional.

Gradient flow

- If $\Sigma = \sigma I$, $\Delta_n = \alpha$ and V is confining then the SDE becomes the stochastic gradient flow equation on the potential V. The corresponding Fokker planck equation.

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla V + \frac{\sigma\alpha}{2} \nabla \rho \right),$$

  Has the following stationary solution

$$\rho = e^{-\frac{2}{\sigma\alpha} V}$$

- Consider the transformation,

$$\rho_1 = e^{\frac{2}{\sigma\alpha} V} \rho,$$

- Multiply by a compactly supported test function $\varphi \in C_0^\infty$, no time dependence, then for $d\mu = e^{-\frac{2}{\sigma\alpha} V} dx$,

$$\int \frac{\partial \rho_1}{\partial t} \varphi d\mu = \int \nabla \cdot \left(\frac{\sigma\alpha}{2} e^{-\frac{2}{\sigma\alpha} V} \nabla \rho_1 \right) \varphi dx,$$

Gradient flow

- We can perform the integration by parts on the right hand side and get,

$$\int \frac{\partial \rho_1}{\partial t} \varphi d\mu = -\int \frac{\sigma\alpha}{2} \nabla \rho_1 \cdot \nabla \varphi d\mu,$$

- Rescaling the time variable leads to a heat equation w.r.t. the measure $d\mu$

$$\int \frac{\partial \rho_1}{\partial t} \varphi d\mu = -\int \nabla \rho_1 \cdot \nabla \varphi d\mu,$$

Conclusion: Our stochastic gradient flow on f gives rise to a gradient flow of the Dirichlet energy

$$E(\rho) = \frac{1}{2} \int |\nabla \rho|^2 d\mu,$$

in L^2_μ.

Consequences

- The non-convex optimization problem becomes convex in the space of distributions.
- We obtain very good tail bounds on the density.
- It becomes easier to study stability problems, for instance what happens in the infinite layer limit.
- We obtain a lot of tools to study different types of regularizers and other first order optimization methods.
- The better the estimate of the Poincaré inequality related to the measure μ the better control over convergence rate to the limit distribution.

Further reading

B. Avelin, K. Nyström,
Neural ODE as the Deep Limit of ResNets. https://arxiv.org/abs/1906.12183