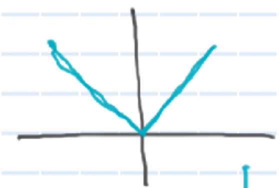
Mathematical foundations of deep neural networks



1/ Intro optimisation $\hat{x} \in \underset{x}{\mathrm{Argmin}}\, f(x)$

2/ Basics $\Gamma_0(\mathcal{H})$: convex, l.s.c, proper. (non-smooth)

3/ Subdifferential

4/ Conjugate.

5/ Proximity operator

6/ Algorithms Gradient descent / Proximal point algo.
Forward - Backward algo.
Douglas - Rachford algo.

7/ FISTA

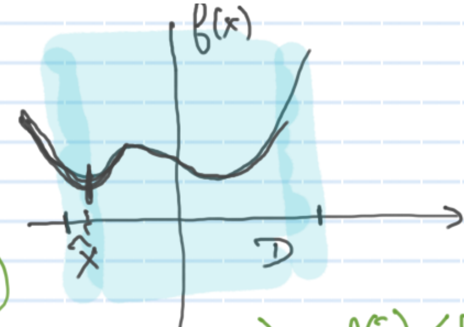
8/ Primal-dual algo.

9/ Stochastic algo

10/ Unfolded algo.

1- Introduction.

- Solve minimization (maximization) problems involving
  - a cost function $f : \mathbb{R}^N \to \mathbb{R}$ (reward)
  - a subset D of $\mathbb{R}^N$



- Goal: We want to

$$\text{Find } \hat{x} \in D \text{ such that } (\forall x \in D)\ f(\hat{x}) \leq f(x) \qquad f(\hat{x}) \geq f(x)$$

$$\iff \text{Find } \hat{x} \in D \text{ such that } f(\hat{x}) = \inf_{x \in D} f(x)$$

$$\iff \text{Find } \boxed{\hat{x} \in \operatorname{Argmin}_{x \in D} f(x)}$$

$$\iff (\forall x \in D)\ -f(\hat{x}) \leq -f(x)$$

$$\iff \dots -f(\hat{x}) \leq -f(x)$$

$$\iff \hat{x} \in \operatorname{Argmin}_{x} -f(x)$$

Without loss of generality, we focus on minimization problems.

- Reformulation with indicator function. $(\forall x \in \mathbb{R}^N)$ $C \subset \mathbb{R}^N$.

$$i_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise} \end{cases}$$

unifying view of constrained and unconstrained optimisation

$$\text{Find } \hat{x} \in \operatorname{Argmin}_{x \in D} f(x) \iff \text{Find } \hat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}^N} f(x) + i_D(x)$$

$$\iff \text{Find } \hat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}^N} \tilde{f}(x)$$

$\hat{x} \in \text{Argmin}_x f(x)$

Main questions

1/ Existence / Uniqueness of $\hat{x}$

2/ Necessary and suffisant conditions for $\hat{x}$ to be a solution

3/ Design an algorithm to approximate a solution when no closed form is available, i.e. building a sequence $(x_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to+\infty} x_n = \hat{x}$

Example : $\mathcal{D} = \{(u_\ell, z_\ell) \in \mathcal{H}\times\mathcal{G} \mid \ell = \{1,\dots L\}\}$

Standard learning.

x image classif $\begin{cases} u_\ell \in \mathbb{R}^N & N \text{ \# pixels.} \\ z_\ell \in \{-1,1\}. \end{cases}$

x

logistic sparse class. $\rightarrow$ $\hat{x} \in \text{Arg}\min_x \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} \log\left(1+e^{-z_\ell x^T u_\ell}\right)}_{f(x)} + \lambda \|x\|_1$

non-sm

non-smooth

SVM class. $\rightarrow$ $\hat{x} \in \text{Argmin}\ \frac{1}{L}\sum_{\ell=1}^{L} \max\left(0, 1 - z_\ell x^T u_\ell\right) + \lambda\|x\|_2^2$

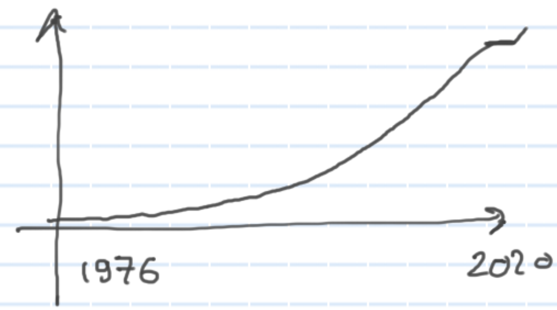
Deep learning $\quad \min_\theta \frac{1}{L}\sum_{\ell=1}^{L} f_1\left(z_\ell, d_\theta(u_\ell)\right) + f_2(\theta)$

$$d_\theta(u) = \eta^{[K]}\left(W^{[K]}\eta^{[K-1]}\left(W^{[K-1]} \cdots \eta^{[1]}\left(W^{[1]}u\right)\right)\right)$$

weights

activation functions.

× Google scholar. "Proximal algorithms"



× $\min\limits_{x} \sum_{s=1}^{S} f_s(x)$

$$\hat{x} \in \operatorname{Arg\,min}_{x} \underbrace{F(x) + g(Lx)}_{\varphi(x)} \iff 0 \in \partial \varphi(\hat{x})$$

$$\iff 0 \in \partial (F + g \circ L)\hat{x}$$

$$p = \mathrm{prox}_{\tau f}(x) \iff x - p \in \tau \partial f(p)$$

$$\mathrm{prox}_f(x) = \arg\min_{y} f(y) + \frac{1}{2}\|x-y\|^2$$

$$x_{n+1} = \mathrm{prox}_{\tau f}(x_n) \iff x_n - x_{n+1} \in \tau \partial f(x_{n+1})$$

$$\iff x_{n+1} = x_n - \tau u_n \quad \text{with} \quad u_n \in \partial f(x_{n+1})$$

$$\text{prox}_{\frac{1}{2}\|A\cdot - z\|_2^2}(x) = \arg\min_{y} \frac{1}{2}\|x-y\|^2 + \frac{1}{2}\|Ay - z\|_2^2$$

$$y - x + A^*(Ay - z) = 0$$

$$y = (I + A^*A)^{-1}(x + A^*z)$$

$$\text{prox}_{f\circ L}(x) = \arg\min_{y} \frac{1}{2}\|x-y\|^2 + f(Ly)$$

$$= L^* \,\text{prox}_f(Lx)$$

$$f = \|\cdot\|_1$$

$$\text{prox}_{\|W\cdot\|_1}(x)$$

① , $\min_{x \in C} f(x) \iff \min_x \underbrace{f(x) + i_C(x)}_{\tilde{f}(x) \,:\, \mathcal{H} \to]-\infty, +\infty]}$.

, $f \in \Gamma_0(\mathcal{H})$

, $\partial f = \{\nabla f\}$ if f diff. $\qquad \partial i_C \leftarrow$ normal cone N_C

, f^* conjugate $\qquad (i_C)^* = \sigma_C \leftarrow$ support funct.

, $\mathrm{prox}_f = \arg\min_y \frac{1}{2}\|y - \cdot\|^2 + f(y) \qquad \mathrm{prox}_{i_C} = P_C$.

② $\boxed{\min_x f(x) + g(Lx)}$

$\partial(f + g \circ L) = \partial f + L^* \partial g(L\,\cdot)$ if $f, g \in \Gamma_0(\mathcal{H})$ + add assu on dom

$\mathrm{prox}_{g \circ L}$ closed form expression if $LL^* = \nu\, \mathrm{Id}$.

$\mathrm{prox}_{(f+g)} \leftarrow$ difficult.

$\Rightarrow$ Splitting proximal algo. $\qquad$ (ISTA) Forward - backward