An Optimal Transport View on Generalization

Nemo Fournier

January 13, 2020

Outline

Framework

Main results

An application

Deep Neural Networks

instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
hypothesis space $\mathcal{W}$
loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathrm{R}^+$

learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
underlying distribution D
training sample $S_n \sim D^{\otimes n}$

instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
hypothesis space $\mathcal{W}$
loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathrm{R}^+$

learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
underlying distribution D
training sample $S_n \sim D^{\otimes n}$

risk $R(w) = \mathbb{E}_{z \sim D}[\ell(z, w)]$

empirical risk $R_{S_n}(w) = \mathbb{E}_{z \sim S_n}[\ell(z, h)] = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, w)$

generalization error $G\left(D, P_{W|S_n}\right) = \mathbb{E}\left[R(W) - R_{S_n}(W)\right]$

μ and ν two measures on $\mathcal{W}$

coupling T measure on $\mathcal{W} \times \mathcal{W}$ such that $\begin{cases} T(X, \mathcal{W}) = \mu(X) \\ T(\mathcal{W}, X) = \nu(X) \end{cases}$

wasserstein distance

$$\mathbb{W}_1(\mu, \nu) = \inf_{T \in \Gamma(\mu,\nu)} \mathbb{E}_{(W,W') \sim T}[d_{\mathcal{W}}(W, W')]$$

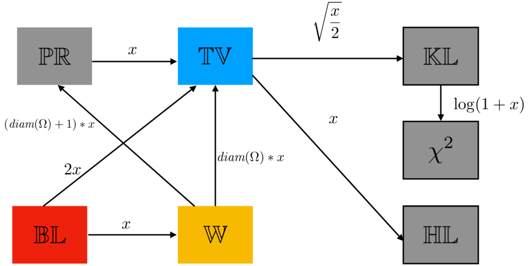
algorithmic transport cost of algorithm $\mathcal{A}$ $(P_{W|S_n})$

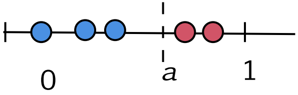
$$Opt\left(D, P_{W|S_n}\right) = \mathbb{E}_{z \sim D}\left[\mathbb{W}_1\left(P_W, P_{W|z}\right)\right]$$

$$A.G.T.\ Opt\left(D, P_{W|S_n}\right) = \mathbb{E}_{z\sim D}\left[\mathbb{W}_1\left(P_W, P_{W|z}\right)\right]$$

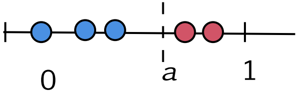
main theorem

$$G\left(D, P_{W|S_n}\right) = \mathbb{E}\left[R(W) - R_{S_n}(W)\right] \leq K \times Opt\left(D, P_{W|S_n}\right)$$





$$\mathcal{A} : \begin{cases} \mathcal{Z}^n & \to & \mathcal{W} \\ S_n = \{(x_1, y_1), \dots, (x_n, y_n)\} & \mapsto & \begin{cases} \max\limits_{\substack{1 \le i \le n \\ \text{s.t. } y_i = 0}} x_i \text{ if } \{i \mid y_i = 0\} \neq \emptyset \\ 0 \text{ otherwise} \end{cases} \end{cases}$$



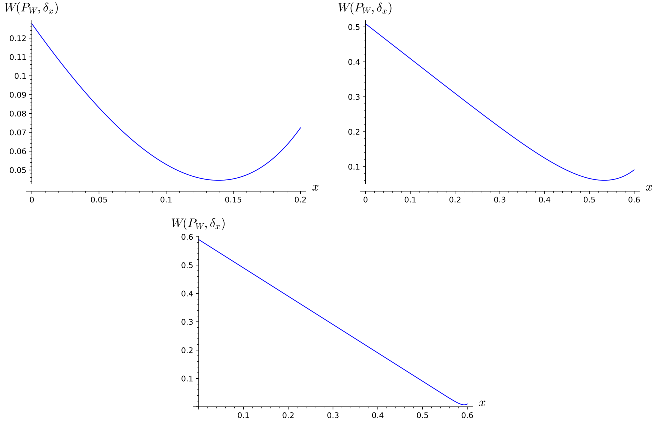
$$\mathcal{A}:\begin{cases} \mathcal{Z}^n & \to & \mathcal{W} \\ S_n = \{(x_1,y_1),\dots,(x_n,y_n)\} & \mapsto & \begin{cases} \max\limits_{\substack{1\le i\le n \\ \text{s.t. } y_i=0}} x_i \text{ if } \{i \mid y_i = 0\} \neq \emptyset \\ 0 \text{ otherwise} \end{cases} \end{cases}$$

$$P_W(w) = (1-a)^{n-k} + n(w+1-a)^n$$

$$P_{W|z} = \delta_x \text{ if } x \le a \qquad P_{W|z} = \delta_0 \text{ otherwise}$$

$$\mathbb{W}_1(\mu,\delta_t) = \mathbb{E}_{X\sim\mu}[d(X,t)] \implies \mathbb{W}_1(P_W,\delta_x) = \int_0^a |x-w| P_W(w)\mathrm{d}w$$

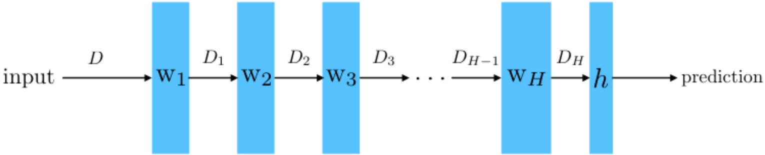
$$\mathbb{W}_1(P_W,\delta_x) = \frac{1}{2\,(an+a)}\Big(a^2((-a+1)^n+2)n + a^2(3\,(-a+1)^n+2) + 2\,((-a+1)^n n + (-a+1)^n)x^2$$
$$-4\Big(a^2 - ax - a\Big)(-a+x+1)^n - 2\,a((-a+1)^n+1) - 2\,(a(2\,(-a+1)^n+1)n$$
$$+a(2\,(-a+1)^n+1))\,x\Big)$$



$$Opt\left(D, P_{W_n}\right) = \int_0^a \mathbb{W}_1\left(P_W, \delta_x\right) \mathrm{d}x + (1-a)\mathbb{W}_1\left(P_W, \delta_0\right)$$

$$Opt\left(D, P_{W_n}\right) = \frac{1}{6(n^2+3n+2)}\Big(2a^2(2(-a+1)^n-3) - \Big(a^2(4(-a+1)^n+3) - 3a((-a+1)^n+2)\Big)n^2$$
$$-3\Big(3a^2+3a((-a+1)^n-2) - 2(-a+1)^n+2\Big)n\Big) - 6a((-a+1)^n-2)$$

$$G\left(D, P_{W|S_n}\right) \leq 1 \times Opt\left(D, P_{W_n}\right)$$



$$\mathbb{E}\left[R(W) - R_{S_n}(W)\right] \le \exp\left(-\frac{H}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{K^2R^2I(S_n;W)}{2n}}$$

Powerful theoretical tool (average case, link with information theory)

Quite quickly too convoluted to provide concrete bounds