Hieber.

(X, Y) : Distributed as $Y_i = f_0(X_i) + \varepsilon_i$, iid $X_i \in [0,1]^d$, $\varepsilon_i \overset{iid}{\sim} N(0,1)$

f_0 takes the form $f_0 = g_q \circ g_{q-1} \circ \cdots \circ g_1 \circ g_0$ and each $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$

$d_0 = d$, $d_{q+1} = 1$, and each g_i is t_i-variate where $t_i \le d_i$ and $g_{ij} \in C^{\beta_i}_{t_i}([a_i, b_i]^{t_i}, K)$ $i \in [q]$.

Network Class : $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$, $\vec{x} \mapsto f(\vec{x}) = W_L \sigma_{v_L} W_{L-1} \sigma_{v_{L-1}} \cdots W_1 \sigma_{v_1} W_0 x$.

s.t. $\max_{j=0,\cdots,L} \|W_j\|_\infty \vee |v_j|_\infty \le 1$, $\sum_{j=0}^{L} \|W_j\|_0 + |v_j|_0 \le s$, $\||f|_\infty\|_\infty \le F$.

Required Network Complexity : (i) $F \ge \max(K, 1)$ (ii) $\sum_{i=0}^{q} \log_2(4t_i \vee 4\beta_i) \log_2 n \le L \lesssim n\phi_n$

(iii) $n\phi_n \lesssim \min_{i=1,\cdots,L} p_i$ (iv) $s \asymp n\phi_n \log n$.

where $\beta_i^* := \beta_i \prod_{l=i+1}^{q} (\beta_l \wedge 1)$, $\phi_n := \max_{i=0,\cdots,q} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}$.

Results : $\exists$ constants C, C' dependent on $q, \vec{d}, \vec{t}, \vec{\beta}, K$ such that for any $\hat{f}_n$ in network class above satisfying required complexity,

(i) if $\Delta_n(\hat{f}_n, f_0) := E_{f_0}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_n(X_i))^2 - \inf_{f \in \mathcal{F}(L,p,s,F)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2\right] \le C\phi_n L \log_2^2 n$, then,

$$R(\hat{f}_n, f_0) := E_{f_0}[(\hat{f}_n(X) - f_0(X))^2] \le C' \phi_n L \log_2^2 n.$$

(ii) if $\Delta_n(\hat{f}_n, f_0) \ge C\phi_n L \log^2 n$. then

$$\frac{1}{C'} \Delta_n(\hat{f}_n, f_0) \le R(\hat{f}_n, f_0) \le C' \Delta_n(\hat{f}_n, f_0).$$

*. Under the assumption that if X has Lebesgue density with positive lower & upper bounds + $t_i \le \min\{d_0, \cdots, d_{i-1}\}$ $\forall i \in [q]$, then $\inf_{\hat{f}_n} \sup_{f_0 \in \text{setting}} R(\hat{f}_n, f_0) \ge c\phi_n$. where inf is over all estimators.

Arora

(x,y) : $\|x\|_2 = 1$, $|y| \le 1$, $(x,y) \sim D$ over $\mathbb{R}^d \times \mathbb{R}$, $\left(H^{\infty}_{ij} = E_{w \sim N(0,I)}\left[x_i^T x_j \mathbb{1}(w^T x_i \ge 0, w^T x_j \ge 0)\right] = \frac{x_i^T x_j (\pi - \arccos(x_i^T x_j))}{2\pi} \right) \Rightarrow \lambda_{min}(H^{\infty}) > 0$

Network Class : $f_{w,a}(x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^T x)$, $\sigma(z) = \max(0,z)$

Training : $\Phi(w) := \frac{1}{2} \sum_{i=1}^{n} (y_i - f_{w,a}(x_i))^2$, $w_r(k+1) - w_r(k) = -\eta \frac{\partial \Phi(w)}{\partial w_r} = -\eta \frac{a_r}{\sqrt{m}} \sum_{i=1}^{n} (f_{w(k),a}(x_i) - y_i) \mathbb{1}(w_r(k)^T x_i \ge 0) x_i$

where $\eta > 0$ is the learning rate. $w_r(0) \sim N(0, \kappa^2 I)$, $a_r \sim unif(-1,1)$

Required Network Complexity

(i) $\kappa = O\left(\frac{\epsilon \delta}{\sqrt{n}}\right)$, $m = \Omega\left(\frac{n^7}{\lambda_0^4 \kappa^2 \delta^4 \epsilon^2}\right)$, $\eta = O\left(\frac{\lambda_0}{n^2}\right)$

(ii) $\kappa = O\left(\frac{\lambda_0 \delta}{n}\right.$, $m \ge \kappa^{-2} poly(n, \lambda_0^{-1}, \delta^{-1})$

Results (i) With probability $1-\delta$, $\forall k = 0,1,2,\cdots$, $\|y - u(k)\|_2 \le \sqrt{\sum_{i=1}^{n} (1-\eta\lambda_i)^{2k} (v_i^T y)^2} \pm \epsilon$.

where $u_i = f_{w,a}(x_i)$, v_i are from $H^{\infty} = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ (orthonormal eigenvectors).

(ii) For $\delta > 0$, loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$ that's 1-Lipschitz, in first argument with $\ell(y,y) = 0$, with prob $1-\delta$,

for $k \ge \Omega\left(\frac{1}{\eta \lambda_0} \log \frac{n}{\delta}\right)$, $E_{(x,y) \sim D}\left[\ell(f_{w(k),a}(x), y)\right] \le \sqrt{\frac{2 y^T (H^{\infty})^{-1} y}{n}} + O\left(\sqrt{\frac{\log\left(\frac{n}{\lambda_0 \delta}\right)}{n}}\right)$

over train sample & initialization.

under the assumption that $P_D(\lambda_{min}(H^{\infty}) \ge \lambda_0) \ge 1 - \frac{\delta}{3}$ for fixed λ_0.

(iii) If $y_i = g(x_i) = \alpha (\beta^T x_i)^p$ $\forall i \in [n]$, for $p=1$ or $p=2l$ for $l \in \mathbb{N}_+$, $\beta \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$, Then we have

$\sqrt{y^T (H^{\infty})^{-1} y} \le 3p|\alpha| \cdot \|\beta\|_2^p$

Chen (2020)

(i) $H_{ij}(p) = E_p[u^2 \langle \nabla_\theta h(\theta, x_i), \nabla_\theta h(\theta, x_j) \rangle] + E_p[h(\theta, x_i) h(\theta, x_j)]$ is positive definite.

(X, Y) : $\|x_i\|_2 \le 1$, $|y_i| \le 1$, $x \in \mathbb{R}^d$, (ii) $\exists\, p_{true}$ with $D_{\chi^2}(p_{true} \| p_0) < \infty$ s.t. $y = \int u h(\theta, x) p_{true}(\theta, u)\, d\theta\, du$

$S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$.

Network Class : $f(p, x) = \alpha \int_{\mathbb{R}^{d+1}} u h(\theta, x) p(\theta, u)\, d\theta\, du$, p_t is generated from the PDE

Training : $$\frac{dp_t}{dt} = -\nabla_u [p_t(\theta, u) g_1(t, \theta, u)] - \nabla_\theta \cdot [p_t(\theta, u) g_2(t, \theta, u)] + \lambda \Delta p_t(\theta, u)$$

which is the infinite-width, continuous-time version of noisy gradient descent

Required Network Complexity : $\lambda_0 := \sqrt{\frac{\Lambda}{n}}$ where $\Lambda = \lambda_{min}(H(p_0))$.

Activation $h(\theta, x)$ is s.t. $h(\theta, x) = \tilde{h}(\theta^T x)$ where $\tilde{h}$ satisfies $|\tilde{h}(z)| \le G$, $|\tilde{h}'(z)| \le G$, $|\tilde{h}''(z)| \le G$

$|z \tilde{h}'(z)| \le G$, $|\tilde{h}'''(z)| \le G$.

(i) $\alpha \ge 8\sqrt{A_2^2 + \lambda A_1^2}\, \lambda_0^{-2} R^{-1}$, $R = \min\{\sqrt{d+1}, \mathrm{poly}(G, \log(\frac{1}{\lambda_0}))^{-1} \lambda_0^2\}$, $A_1 = 2G(d+1) + 4G\sqrt{d+1}$, $A_2 = 16G\sqrt{d+1} + 4G$

(ii) $\alpha \ge \sqrt{n\lambda} > 0$.

Results : (i) $L(p_t) = E_S[\phi(f(p, x), y)] \le 2\exp(-2\alpha^2 \lambda_0^2 t) + 2A_1^2 \lambda^2 \alpha^{-2} \lambda_0^{-4}$

$D_{KL}(p_t \| p_0) \le 4A_2^2 \alpha^{-2} \lambda_0^{-4} + 4A_1^2 \lambda \alpha^{-2} \lambda_0^{-4}$.

(ii) For any $\delta > 0$, with probability $1-\delta$, $E_D[\ell^{0-1}(f(p^*, x), y)] \le (8G+1)\sqrt{\frac{D_{\chi^2}(p_{true} \| p_0)}{n}} + 6\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$

where p^* is the minimizer of $L(p) + \lambda D_{KL}(p \| p_0)$, $\ell^{0-1}(y', y) = \mathbb{1}(y'y < 0)$

Here, (X_i, Y_i) are iid sampled from unknown distribution D.

Training : Noisy GD , $\{(\theta_j, u_j)\}_{j=1,\cdots,m}$ initialized by $p_0(\theta, u) \sim N(0, I_{d+1 \times d+1})$