SGD:

A sequence of i.i.d. data $Y_1, Y_2, \ldots \in \mathcal{Y}_n \subseteq \mathbb{R}^{d_n}$
$\sim P_n \in M_1(\mathbb{R}^{d_n})$

$L_n : \mathcal{X}_n \times \mathcal{Y}_n \to \mathbb{R}$

$\mathcal{X}_n \subseteq \mathbb{R}^{p_n}$

Online SGD with constant learning rate δ_n

$$X_l = X_{l-1} - \delta_n \nabla L_n(X_{l-1}, Y_l)$$

$$X_0 \sim \mu_n \in M_1(\mathcal{X}_n)$$

Consider a summary statistics $u_n \in C^2(\mathbb{R}^{p_n}; \mathbb{R}^k)$

$u_n = (u_n^1(x), \ldots, u_n^k(x))$

Goal: $\underline{\boxed{u_n(X_l)}} \to ? \; \boxed{u(t)}$

δ_n Do interpolation

$$H(x, Y) = L_n(x, Y) - \boxed{\Phi(x)}$$

where $\Phi(x) = \mathbb{E}_Y H(x, Y)$

$$\underline{\underline{V}}(x) = \mathbb{E}[\nabla H(x) \otimes \nabla H(x)] = (V_{ij})$$

Thm 2.2

$$\mathcal{A}_n = \langle \nabla \Phi, \nabla \rangle$$

$$\mathcal{L}_n = \frac{1}{2} \sum_{i,j} V_{ij} \partial^2_{ij}$$

Let

$(X_l^{\delta_n})_l$ be the SGD initialized from $X_0 \sim \mu_n \in M_1(\mathbb{R}^{P_n})$, δ_n, $Y_l \overset{\text{iid}}{\sim} P_n$

$(u_n, L_n, P_n) \rightsquigarrow \delta_n$ - localization

L_n: Loss

H

$\mathbb{R}^k = \bigcup_k E_k$, E_k: cpt

Suppose $\exists\ h: \mathbb{R}^k \to \mathbb{R}^k$, $\Sigma: \mathbb{R}^k \to P_k$ (≥ 0), s.t. $\forall K$-cpt

$$\sup_{x \in u_n^{-1}(E_K)} \| (-\mathcal{A}_n + \delta_n \mathcal{L}_n) u_n(x) - h(u_n(x)) \| \to 0$$

$$\sup_{x \in u_n^{-1}(E_K)} \| \delta_n J_n V J_n^T - \Sigma(u_n(x)) \| \to 0.$$

$$J_n = (\nabla u_n)$$

Then $(u_n(t))_t \xleftarrow{\text{interpolation}} \left(u_n\left(X^{\delta_n}_{\lfloor t/\delta_n \rfloor}\right)_t\right)$

$$X_0 \sim \mu_n$$

$(u_n)_* \mu_n \xrightarrow{\text{as } n\to\infty} \nu$ weakly, then

$(u_n(t))_t \xrightarrow{\text{as } n\to\infty} (u_t)_t$, where u_t solves

$$\begin{cases} du_t = h(u_t)\,dt + \sqrt{\Sigma(u_t)}\,dB_t . \\ u_0 \sim \nu \end{cases}$$

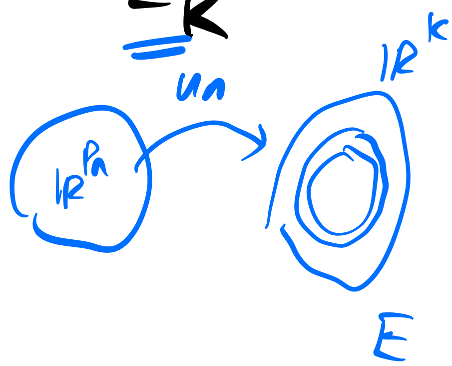
Pf

τ_K^n :

interplocted process

$=$ exit time of $u_n(t)$ from E_K

$$E_K^* = u_n^{-1}(E_K)$$



$$L_K^\infty = L^\infty(E_K^*)$$

$$f_l = f(X_l) \qquad H^l_{l-1} = H(X_{l-1}, Y_l)$$

Def: 2.1 (u_n, L_n, P_n) is δ_n-localizable

if: $\exists (E_K)_K$ s.t. $\mathbb{R}^k = \bigcup_K E_K$, $K \uparrow$ cpt.

$E_K \to C(K)$

i) $\max_{1 \le i \le k} \sup_{\substack{x \in u_n^{-1}(E_K) \\ u_n(x) \in \mathbb{R}^k}} \| \nabla^j u_i \|_{op} \le C(K) \cdot \delta_n^{-(3-j)/2}$

ii) $\sup_{x \in u_n^{-1}(E_K)} \| \nabla \Phi \| \le C(K)$

$\Phi(x) = \mathbb{E}_Y (x, Y)$

$\sup_{x \in u_n^{-1}(E_K)} \mathbb{E}[\| \nabla H \|^8] \le C(K) \, \delta_n^{-4}$

iii) $\max_{1 \le i \le K} \sup_{x \in u_n^{-1}(E_K)} \mathbb{E}[\langle \nabla H, \nabla u_i \rangle^4] \le C(K) \, \delta_n^{-2}$

— .. — $\mathbb{E}[\langle \nabla^2 u_i, \nabla H \otimes \nabla H - V \rangle^2]$

H

$$= O(\delta_n^{-3})$$

$$L_n = \overbrace{L_n - \Phi(x)}^{\parallel} + \Phi$$

$$X_l = X_{l-1} - \delta_n \nabla L_n (X_{l-1}, Y_l)$$

$$= X_{l-1} - \delta_n \left(\nabla \Phi (X_{l-1}) + \nabla H (X_{l-1}, Y_l) \right)$$

$$f_l = f\left(X_{l-1} - \delta \nabla \Phi_{l-1} - \delta \nabla H^l_{l-1}\right)$$

$$= f_{l-1} - \delta \langle \nabla f_{l-1}, \nabla \Phi_{l-1} + \nabla H^l_{l-1} \rangle \; \star$$

$$+ \frac{1}{2} \delta^2 \langle \nabla^2 f_{l-1}, (\nabla \Phi_{l-1} + \nabla H^l_{l-1}) \otimes (\nabla \Phi_{l-1} + \nabla H^l_{l-1}) \rangle$$

$$+ O(\delta^3 \|\nabla^3 f\|_{L^\infty_K} \|\nabla L\|^3_{L^\infty_{K,n}})$$

Organize the above to have:

$$f_l = f_{l-1} + \delta [A^f_l - A^f_{l-1}]$$

$$+ \delta [M^f_l - M^f_{l-1}]$$

$$+ \text{h.o.T.}$$

where $\Rightarrow$ Pre-visible

$\|\nabla\phi\| \le C(K)$

$$A_l^f - A_{l-1}^f = -\langle \nabla\Phi_{l-1}, \nabla\rangle f_{l-1}$$

$$+ \delta \mathcal{L}_n f_{l-1} + \tfrac{1}{2}\delta \langle \nabla\Phi\otimes\nabla\Phi, \nabla\delta\rangle_{l-1}$$

where $\mathcal{L}_n f_{l-1} = \frac{1}{2}\langle \nabla^2 f, V\rangle = \frac{1}{2}\sum_{i,j} v_{ij}\partial_{v_i}\partial_{v_j}$

$$\boxed{V = \mathbb{E}\langle \nabla H_{l-1} \otimes \nabla H_{l-1}\rangle}$$

$\mathbb{E}\|\nabla H\|^2$

$\le (\mathbb{E}\|\nabla H\|^8)^{1/4}$

$\lesssim \delta_n^{-1}$

$$M_l^f - M_{l-1}^f = -\langle \nabla H^l, \nabla f\rangle_{l-1}$$

$(M_l^f)_{l \le K}$ martingale.

$$+ \delta(\Sigma_l^f - \Sigma_{l-1}^f)$$

where

$$\Sigma_l^f - \Sigma_{l-1}^f = (\nabla^2 f : \nabla\Phi \otimes \nabla H^l)_{l-1}$$

$$+ \tfrac{1}{2}\langle \nabla^2 f, \nabla H^l \otimes \nabla H^l - V\rangle_{l-1}$$

$\boxed{f = u_j}$ $\Leftarrow$ j-th coordinate of $\vec{u}_n$

$$\text{H.O.T} = O(\delta^3 \|\nabla^3 u_j\|_{L_K^\infty} \|\nabla L\|^3_{L_K^\infty})$$

0

$$\delta^3 \sup_{x \in \mathbb{E}_K^*} \mathbb{E}\Big[\|\nabla^3 u_j\| \, \|\nabla L\|^3\Big]$$

$$\nabla L = \nabla \Phi + \nabla H$$

$$\lesssim \delta^3 \underbrace{\|\nabla^3 u_j\|_{L^\infty_K}}_{\leq C(K)} \Big(\underbrace{\|\nabla \Phi\|^3_{L^\infty_K}}_{\leq C(K)} + \underbrace{\sup_{E_K^*} \mathbb{E}\|\nabla H\|^3}\Big)$$

$$\lesssim \delta^{\frac{3}{2}}.$$

$$\mathbb{E}\|\nabla H\|^3 \leq \Big(\mathbb{E}\big(\|\nabla H\|^3\big)^{\frac{8}{3}}\Big)^{\frac{3}{8}}$$

$$\leq C(K)\, \delta_n^{-4 \cdot \frac{3}{8}}$$

$$= C(K) \cdot \delta_n^{-\frac{3}{2}}$$

Since $\delta_n \to 0$ as $n \to \infty$, $\boxed{l \delta_n = t}$

$$u_j(X_j) = u_j(0) + \delta \sum_{l' \leq l} \big(A^{u_j}_{l'} - A^{u_j}_{l'-1}\big) + \delta \sum_{l' \leq l} \big(M^{u_j}_{l'} - M^{u_j}_{l'-1}\big) + o(1).$$

Define.

$$a_j'(s) = A^{u_j}_{[s/\delta]} - A^{u_j}_{[s/\delta]-1}$$

$$b_j'(s) = M^{u_j}_{[s/\delta]} - M^{u_j}_{[s/\delta]-1}$$

$$a_j(s) = \int_0^s a_j'(s')\,ds'$$

$$= a_j(\delta[s/\delta]) + \underbrace{(s-\delta[s/\delta])}_{|\cdot|<\delta}\left(A^{u_j}_{[s/\delta]} - A^{u_j}_{[s/\delta]-1}\right)$$

$b_j(s)$

$$u_n(s) = u_n(0) + a_n(s) + b_n(s) + \underline{o(1)}$$

<u>To show:</u> $(u_n(s \wedge \tau_K^n))$ is tight in $C([0,T])$ with limit points is 2-Hölder for each K.

Only need to show this for

$$v_n(s) = u_n(0) + a_n(s) + b_n(s)$$

<u>Tightness</u> : ① "Uniform bounded"

δ_n - <u>localisable</u>

② $\boxed{\text{equi-continuity}}$

$\Uparrow$

Kolmogorov's continuity Thm.

$$\|v_n(s) - v_n(t)\| \leq \|a_n(s) - a_n(t)\| \checkmark$$
$$+ \|b_n(s) - b_n(t)\| \checkmark$$

$$\mathbb{E}\left[\left|a(s \wedge \tau_K) - a(t \wedge \tau_K)\right|^4\right]$$

$$\lesssim \mathbb{E}\Big|\delta \sum_l \underline{\underline{\left(\left(-A_n + \delta L_n\right) u\right)_l}}\Big|^4 \lesssim_K (t-s)^4$$

$$+ \mathbb{E}\Big|\delta^2 \sum_l \langle \nabla\Phi \otimes \nabla\Phi, \nabla^2 u\rangle_l\Big|^4$$

$$l \sim \left[[s/\delta] \wedge \tau_K/\delta \,,\, [t/\delta] \wedge \tau_K/\delta\right]$$

$$\lesssim_K \mathbb{E}\Big|\delta \sum_l \left(h(u_n)\right)_l\Big|^4 + o(|t-s|^4)$$

$\lesssim_K |t-s|^4$

① $\mathbb{E}|a(s\wedge\tau_K) - a(t\wedge\tau_K)|^4 \lesssim_K (t-s)^4$

② $\mathbb{E}|b(s\wedge\tau_K) - b(t\wedge\tau_K)|^4 \lesssim |t-s|^2$

$\Downarrow$ Kolmogorov's continuity Thm

$(v_n(s\wedge\tau_K))_s$ is (uniformly) $1/4$ - Hölder

$\Longrightarrow$ The sequence $(v_n(s\wedge\tau_K))_s$ is tight with $\frac{1}{4}$ - Hölder limit points

Next:

$(v_n(t\wedge\tau_K) - a_n(t\wedge\tau_K))_t$

is tight and the limits points

are continuous martingales.

Define $v_n^K(t) = v_n(t \wedge \tau_K) \longrightarrow \begin{matrix} v^K \\ a^K \\ b^K \end{matrix}$

a_n^K, b_n^K

$\frac{1}{4}$-Hölder

$$\boxed{\begin{matrix} \delta_n \to 0 \\ d_n \to \infty \\ p_n \to \infty \end{matrix}}$$

ASSUMPTION:

$$\sup_{x \in u_n^{-1}(E_K)} \| (A_n + \delta_n \ell_n) u_n(x) - h(u_n(x)) \| \to 0$$

$$\sup_{x \in u_n^{-1}(E_K)} \| \delta_n J_n V J_n^T - \Sigma(u_n(x)) \| \to 0$$