The heavy-tail phenomenon in SGD

Fonction à minimiser sur $\mathbb{R}^d$:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

f_i différentiables
L-smooth
μ-convexes / μ-PL.

Mini-batch-SGD :

$$x_{n+1} = x_n - \eta \times \widehat{\nabla f}(x_n)$$

step-size η.

avec $\widehat{\nabla f}(x_n) = \frac{1}{b}\sum_{i\in B} \nabla f_i(x_n)$.

et $B \sim \mathrm{Unif}(\{B \subseteq [m],\ |B| = b\})$.

Théorème Si les f_i sont μ-fortement convexes et L-smooth, alors $\forall \eta < \frac{1}{2L}$

$$\mathbb{E}\left[|x_n - x_*|^2\right] \leq (1-\eta\mu)^n |x_0 - x_*|^2 + \frac{2\eta}{\mu}\sigma^2$$

où σ^2 dépend de f et b : $\sigma^2 = \mathrm{Var}(\widehat{\nabla f}(x_*))$.

En particulier, $(x_n - x_*)_n$ est une suite L^2-bornée donc tendue : quitte à extraire on peut supposer que $x_n \xrightarrow[n\to\infty]{\text{loi}} x_*$.

Que dire de la loi limite et de sa dépendance en $\eta, b, L,$ et ?

Écrivons $\nabla f_i(x) \approx \nabla f_i(x_*) + \nabla^2 f_i(x_*)(x_n - x_*)$ autour de x_*. Si x_* est un minimum de chaque f_i (interpolation) alors $\nabla f_i(x_*) = 0$.

On a donc :

$$x_{n+1} = x_n - \frac{\eta}{b}\left(\sum_{i\in\mathcal{B}} \nabla f_i(x_n)\right)$$

$$\approx x_n - \frac{\eta}{b}\sum_{i\in\mathcal{B}_n} \nabla^2 f_i(x_*)x_n + \frac{\eta}{b}\sum_{i\in\mathcal{B}_n} \nabla^2 f_i(x_*)x_*$$

$$\approx \underbrace{\left(Id - \frac{\eta}{b}\sum_{i\in\mathcal{B}_n} \nabla^2 f_i(x_*)\right)}_{= A_n}x_n + \underbrace{\left(\frac{\eta}{b}\sum_{i\in\mathcal{B}_n} \nabla^2 f_i(x_*)x_*\right)}_{= B_n}$$

A_n : matrice aléatoire de taille $d \times d$

B_n : vecteur aléatoire de taille d.

On a donc $x_{n+1} \approx A_n x_n + B_n$. (1)

(1) est exacte dans le cas où les f_i sont toutes quadratiques. C'est le cas de la régression linéaire classique : training data (z_i, y_i) avec $y_i = \langle x_*, z_i\rangle + \varepsilon$ $z_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$ $x \in \mathbb{R}^d$

loss : $f(x) = \frac{1}{2n}\sum_{i=1}^{n} \underbrace{|\langle x, z_i\rangle - y_i|^2}_{f_i(x)}$.

$\nabla f_i(x) = (\langle x, z_i\rangle - y_i) z_i$

$\nabla^2 f_i(x) = z_i z_i^*$.

La récursion (1) s'écrit alors

$$x_{n+1} = \left(I - \frac{\eta}{b} \sum_{i \in B_n} z_i z_i^T \right) x_n + \frac{\eta}{b} \sum_{i \in B_n} z_i z_i^T X$$

On suppose qu'on est dans un cadre de SGD online : les batchs (y_i, z_i) sont générés à chaque étape et jamais réutilisés.

Hypothèses $z_i \sim N(0, I)$ dimension 1

$x_0 = 1$

$$\boxed{x_{n+1} = \left(I - \frac{\eta}{b} K_n^2 \right) x_n + \frac{\eta}{b} K_n^2}$$

où $K_n^2 \sim \chi^2(b)$, iid des (x_n).

Récursions affines et queue de proba.

On considère l'équation en loi $\boxed{X \overset{\text{loi}}{=} AX + B}$ (2)

Si la récursion $X_{n+1} = A_n X_n + B_n$ converge en loi c'est forcément vers une solution de (2)

On supposera toujours que (A, B) a une densité contre lebesgue. Par simplicité on suppose aussi que $A, B > 0$ mais les résultats s'étendent à tous les cas.

Lemme 1 $X \overset{\text{loi}}{=} \lim_{n\to\infty} \sum_{k=0}^{n} A_n A_{n-1} \cdots A_{n-k} B_k$

la série converge si $\mathbb{E} \ln|A| = \gamma < 0$.

Théorème (Kesten 1973-75)

(1) On suppose qu'il existe $s > 0$ tel que

(i) $\mathbb{E}A^s = 1$ (ii) $\mathbb{E}A^s \ln A < \infty$ (iii) $\mathbb{E}B^s < \infty$.

Alors $\exists c > 0$ tq $\mathbb{P}(X > x) \sim \frac{c}{x^s}$.

(2) Si A peut prendre des valeurs négatives, le même résultat est vrai.

(3) En plusieurs dims on pose $h(s) = \lim_{n\to\infty} \mathbb{E}\left(\|A_n \cdots A_1\|^s\right)^{\frac{1}{n}}$.

Si $\exists s$ tq $h(s) = 1$, $\mathbb{E}\|A\|^s \ln\|A\| < \infty$, $\mathbb{E}\|B\|^s < \infty$

alors $\forall u \in \mathbb{R}^n$, $u \neq 0$ $\exists c(u) > 0$ tq

$$\mathbb{P}(\langle u, X\rangle > x) \sim \frac{c(u)}{x^s}.$$

Application. K_n a tous ses moments exponentiels donc $\mathbb{E}|K_n|^\alpha < \infty$ $\forall \alpha$.

$s \mapsto \mathbb{E}|M_n|^s$ est sta. $\uparrow$ donc $\exists!\, s$ tq

$\mathbb{P}(x_\infty > x) \sim c x^{-s}$.

En plus $s = f\left(\underset{-}{h}, \underset{+}{b}, \underset{-}{d}\right)$

s grand $\Rightarrow$ queue fine

s petit $\Rightarrow$ queue lourde

$s \overset{>}{\sim} 1$

What happens for FC networks / Deep Nets ?

$$f(\theta, x) = \sum_{i=1}^{K} \theta_i^1 \, \varphi(\langle \theta_i^2, x \rangle + \theta_i^3)$$

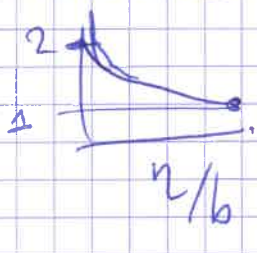
↳ activation.

$\theta_i^1 \in \mathbb{R}$
$\theta_i^2 \in \mathbb{R}^d$
$\theta_i^3 \in \mathbb{R}$ bias.

Idea when $K \to \infty$ the θ_i^2 become independent.

Empirically their distribution seems to become heavy-tailed.

SURTOUT pour les couches externes. les couches du milieu sont HT. $\alpha = f(\eta/b)$



Conséquences sur la Modélisat°

(SGD) $x_{n+1} - x_n = \eta \widehat{\nabla f}(x_n) \longleftrightarrow dX_t = -\nabla f(X_t)\,dt + \Sigma_t \, dL_t$ noise.

Langevin : $X_t \xrightarrow[n \to \infty]{\text{loi}} X \sim \frac{e^{-f}}{Z}$

Si f a une croissance linéaire, e^{-f}/Z a des queues de distribution fines..

Conclusion (SGD) $\longleftrightarrow$ $dX_t = -\nabla f(X_t)\,dt + d L_t^\alpha$

où L^α est un processus de Lévy.

Ultime remarque

Dans le cas où les χ_i sont isotropiques :

$\exists\ \eta_1 < \eta_2$ tq :

$\eta < \eta_1 \Rightarrow S > 2$

$\eta \in (\eta_1, \eta_2) \Rightarrow S \in 2$

$\eta > \eta_2 \Rightarrow S \in (0, 2)$.