Problem Setup

Observations: SGD often generalizes better than adaptive gradient algorithms, e.g., ADAM.

Empirical explanation: adaptive gradient algorithms often converge to sharp minima, while SGD prefers to find flat minima at the flat or asymmetric basins/valleys.

Problem: why SGD often converges to flat minima, while adaptive gradient algorithms, e.g., ADAM, do not?

Stochastic Differential Equation (SDE) Based Analysis

Observation: stochastic gradient noise $\sigma_t$ in SGD and ADAM are heavy-tailed and approximately obey the stable distribution with parameter $\alpha = 2.2$. The SDE driven by $\sigma_t$ and the noise term is a minibatch.

\[ u_t = -\nabla F(\theta_t) - \nabla^2 F(\theta_t) \sigma_t \]

(1)

Assumption: Assume $u_t$ obeys $\sigma_t$ distribution and with a time-dependent covariance matrix $\Sigma_t$.

\[ \Sigma_t = \frac{1}{m} \sum_{i=1}^{m} \nabla^2 F(\theta_t_i) \]

(2)

Levy-driven SDE of SGD:

\[ \theta_{t+1} = \theta_t - \nabla^2 F(\theta_t) \sigma_t \]

(3)

Main Results

Define the escaping time $\Gamma$ from $\Omega$ as (constant $\gamma > 0$ satisfies $\lim_{\alpha \to \infty} \gamma = 0$). Define the escaping set $W^*$ at the basin $\Omega$ as:

\[ W^* = \{ \theta \in \mathbb{R}^d \mid \Sigma_{\alpha} \sigma_t \}

(4)

where $\Sigma_{\alpha} = \lim_{\alpha \to \infty} \Sigma_{\alpha}$ for both SGD and ADAM, and $\Sigma_{\alpha} I$ for SGD and $\Sigma_{\alpha} = \lim_{\alpha \to \infty} \Omega_{\alpha}$ for ADAM.

Assumption 1: For ADAM and SGD, the objective $F(\theta)$ is upper-bounded non-negative, and is locally $\mu$-strongly convex and $\gamma$-smooth in the basin $\Omega$.

Assumption 2: For $\alpha$, $\gamma, \epsilon, \beta, \mu$, and $\rho$ satisfying $\epsilon > 0$ and $\lim_{\alpha \to \infty} \rho = 0$, SGD in (2) and ADAM in (3) obey:

\[ \Gamma = C \left( \frac{1}{m} \| \mathbb{W} \| \right) \]

(5)

Escaping time of SGD and ADAM: Suppose Assumptions 1 and 2 hold. Let $\rho_0 = \sqrt{-\frac{\mu_0}{\mu_1}} \leq \rho_2$, with $\Delta = F(\theta_0) - F(\theta^*)$ and a constant $\Omega_0$. Then for any $\theta_0 \in \Omega, \epsilon \in (0, \gamma]$ and $\rho_0 \in (0, \gamma)$ satisfying $\epsilon > 0$ and $\lim_{\alpha \to \infty} \rho = 0$, SGD in (2) and ADAM in (3) obey:

\[ \Gamma = C \left( \frac{1}{m} \| \mathbb{W} \| \right) \]

(6)

Result 1: Preference to Flat Minima

Definition of “flat minima”: A minimum $\theta^* \in \Omega$ is said to be flat if its basin $\Omega$ has large nonzero Radon measure.

Due to $\| \mathbb{W} \|$, both ADAM and SGD have large escaping time $\Gamma$ at the flat minima.

Define complementary set $\mathbb{W}^*$ of $\mathbb{W}$ as

\[ \mathbb{W}^* = \{ \theta \in \mathbb{R}^d \mid \| \mathbb{W} \| \} \]

(7)

large measure of $\Omega \rightarrow$ large $\| \mathbb{W}^* \|$ with $\| \mathbb{W} \| \rightarrow$ small $\| \mathbb{W} \|$ on $\| \mathbb{W} \| ^\star = constant$

large escaping time $\Gamma$.

Minima with large Radon measure often refers to the conventional flat minima or the minima at the asymmetric basin, since 1) Radon measure positively rely on the volume of the basin and 2) the flat or asymmetric basin often has large volume.

Result 2: Better Sharp Minima Escaping Ability of SGD over ADAM

Factor 1 $\Omega = \mathbb{R}^d$:

- With same learning rate $\epsilon$ in ADAM and SGD, the smaller tail index $\alpha$, the smaller the escaping time $\Gamma$.

- For some iterations, SGD has smaller $\alpha$, as exponential gradient average in ADAM smooths noises.

Factor 2 $\| \mathbb{W} \|$ that positively depends on the volume of escaping set $\mathbb{W}$. Approximating $\Omega$ as a quadratic basin with center $\theta^*$, i.e.

\[ \Omega = \{ \theta \mid \| \mathbb{W} \| = \| \mathbb{W} \| ^\star \}

(8)

with a basin height $h(\theta^*)$ and Hessian matrix $H(\theta^*)$ at $\theta^*$.

Comparison of Escaping Sets of SGD and ADAM: Under the quadratic basin approximation, the escaping set $\mathbb{W}$ is

\[ \mathbb{W}_{\text{ADAM}} \approx \{ \theta \mid \theta \in \mathbb{R}^d \}

(9)

and it satiates $m(\mathbb{W}_{\text{ADAM}}) < m(\mathbb{W}_{\text{SGD}})$.

Conclusion: SGD could better escape from sharp minima and converge to flatter minima, since

- From Factors 1 and 2, SGD has smaller escaping time and is much more unstable. For the same basin, ADAM could stuck in one basin, but SGD could not.

- Both SGD and ADAM prefers to converge to flat minima.