Generalization Error of Stochastic Gradient Descent with Momentum under Heavy-Tailed Noise.

Thanh Dang

Florida State University, td22v@fsu.edu

Joint work with Melih Barsbey (Imperial College London), AKM Rokonuzzaman Sonet (Florida State University), Mert Gurbuzbalaban (Rutgers University), Umut Simsekli (Inria) and Lingjiong Zhu (Florida State University)

Stochastic gradient descent

- Many supervised learning problems can be framed as the *empirical risk minimization* (ERM) problem

$$\min_{\theta \in \mathbb{R}^d} \left\{ \widehat{F}(\theta, X_n) := \frac{1}{n} \sum_{i=1}^{n} f(\theta, x_i) \right\}.$$

$X_n = \{x_1, \ldots, x_n\} \subset \mathcal{X}^n$ is a dataset with i.i.d. observations.

- A popular algorithm to solve the above problem is stochastic gradient descent (SGD)

$$\theta_{k+1} = \theta_k - \eta \nabla \tilde{F}(\theta_k, X_n). \tag{1}$$

Here $\nabla \tilde{F}(\theta, X_n) := \frac{1}{b} \sum_{i \in \Omega_k, |\Omega_k| = b} \nabla f(\theta, x_i)$, with Ω_k being a random subset of $\{1, \ldots, n\}$; and $|\Omega_k| = b \ll n$ is mini-batch size.

- The stochastic part of (1) comes from the noise

$$U_{k+1} := \eta \left(\nabla \tilde{F}(\theta_k, X_n) - \nabla \widehat{F}(\theta_k, X_n) \right). \tag{2}$$

- The standard assumption is that U_{k+1} is Gaussian (Welling & Teh 2011).

Heavy-tailed SGD

- Simsekli et al. (2019), Zhang et al. (2020): over large iterates of SGD, the noise term U_{k+1} has heavy tails (sup-exponential or polynomial).
- Panigrahi et al. (2019), Gurbuzbalaban et al. (2021): smaller batch size and larger step size of SGD are associated with heavier tails.
- Martin & Mahoney (2019), Simsekli et al. (2020), Raj et al. (2020): heaviness of the tails is positively correlated with generalization performance of SGD (aka how well SGD works on unseen data).
- Simsekli and co-authors (2017, 2020) propose modeling the noise U_{k+1} in SGD as α-stable distribution to simulate heavy tails $\Rightarrow$ Fractional Langevin Monte Carlo (LMC).

We will consider the continuous proxy of fractional LMC with momentum.

$$d\theta_t = v_t dt, \qquad dv_t = -\gamma v_t dt - \beta \nabla \widehat{F}(\theta_t, X_n) dt + \zeta dL_t.$$

$\gamma > 0$ is the momentum parameter. $L_t, t \geqslant 0$ is an α-stable Lévy process with stability parameter $\alpha \in (1, 2)$. X_n is the dataset.

Generalization Error

Define $\hat{R}(x, X_n) := \frac{1}{n}\sum_{i=1}^{n} \ell(x, x_i)$ and $R(x) := \mathbb{E}_{X\sim\mathcal{D}}[\ell(x, X)]$, where $\mathcal{D}$ is the unknown probability distribution over the data space $\mathcal{X}$. The expected generalization error is

$$\mathbb{E}_{(\theta_t, v_t), X_n}\left[\hat{R}((\theta_t, v_t), X_n) - R((\theta_t, v_t))\right].$$

Theorem

Assume appropriate assumptions on $\hat{F}$ *and* $\sup_{x,y\in\mathcal{X}} \|x - y\| \leqslant \mathbf{D}$. *Then*

$$\left|\mathbb{E}_{(\theta_\infty, v_\infty), X_n}\left[\hat{R}((\theta_\infty, v_\infty), X_n)\right] - R((\theta_\infty, v_\infty))\right| \leqslant C\frac{1}{n}\mathbf{D}^{5/2},$$

Other contributions:

- for quadratic losses: generalization bound of SGDm is larger than that of SGD $\Rightarrow$ momentum+heavy tails can be bad for generalization (confirmed by synthetic experiment and experiment on neural networks).
- generalization bound of the discretization

$$V_{k+1} = V_k - \eta\gamma V_k - \eta\nabla\widehat{F}(\Theta_k, X_n) + \zeta\xi_{k+1}, \qquad \Theta_{k+1} = \Theta_k + \eta V_{k+1}.$$