Note: Noise Conditions and Convergence Analysis of SGD under Polyak-Łojasiewicz Inequality

Atsushi Nitanda

November 3, 2022

We consider the stochastic gradient descent (SGD) for solving the following problem:

$$\min_{w\in\mathbb{R}^d} F(w),$$

where $F:\mathbb{R}^d\to\mathbb{R}$ is a differentiable function. Let us denote by $G_t(w)$ a stochastic gradient at t-th iterate, which satisfies the following:

Assumption 1 (Stochastic gradient)**.** *$G_t(w)$ is an unbiased estimator of the gradient $\nabla F(w)$, that is, $\mathbb{E}[G_t(w)]=\nabla F(w)$. Moreover, $\{G_t(w)\}_{t=1,2,\dots}$ are independent copies each other.*

Then, SGD is defined as follows: for an initial point $w_1\in\mathbb{R}^d$,

$$w_{t+1}\leftarrow w_t-\eta_t G_t(w_t), \tag{1}$$

where $\eta_t>0$ $(t=1,2,\dots)$ are step sizes.

Example 1 (Risk Minimization)**.** *Let $\ell(w,z)$ be a loss function consisting of the hypothesis function parameterized by $w\in\mathbb{R}^d$ and the data $z\in\mathbb{R}^p$. Let μ be an empirical/true data distribution over the data space and Z be a random variable following μ. Then, the objective function is defined by*

$$F(w)=\mathbb{E}_{Z\sim\mu}[\ell(w,Z)].$$

Given i.i.d. random variables $\{Z_t\}_{t=0}^{\infty}$ with the same distribution as Z, the standard stochastic gradient at t-th iterate is defined as $G_t(w)=\nabla_w\ell(w,Z_t)$. That is, SGD is described as follows:

$$w_{t+1}\leftarrow w_t-\eta_t\nabla_w\ell(w_t,Z_t).$$

Note that we can further include the ℓ_2-regularization in the objective f.

In this note, we make the following standard assumptions on F:

Assumption 2 (Lipschitz smoothness)**.** *There is a constant $L>0$ such that for any $w,w'\in\mathbb{R}^d$,*

$$F(w')\le F(w)+\nabla F(w)^\top(w'-w)+\frac{L}{2}\|w'-w\|_2^2.$$

Assumption 3 (Polyak-Łojasiewicz (PL) Inequality)**.** *Let $F_*=\inf_{w\in\mathbb{R}^d}F(w)$. There is a constant $\mu>0$ such that for any $w\in\mathbb{R}^d$,*

$$\|\nabla F(w)\|_2^2\ge 2\mu(F(w)-F_*).$$

To guarantee the convergence of SGD, we need a condition on the second moment or variance of the stochastic gradient:

$$\mathbb{E}[\|G_t(w)\|_2^2] = \|\nabla F(w)\|_2^2 + \mathbb{E}[\|G_t(w) - \nabla F(w)\|_2^2].$$

So far, several conditions were proposed as summarized below:

- Bounded gradient (BG):
$$\mathbb{E}[\|G_t(w)\|_2^2] \leq \sigma^2.$$
- Bounded variance (BV), (Ghadimi and Lan, 2013):
$$\mathbb{E}[\|G_t(w)\|_2^2] \leq \|\nabla F(w)\|_2^2 + \sigma^2.$$
- Strong growth condition (SGC), (Vaswani et al., 2019):
$$\mathbb{E}[\|G_t(w)\|_2^2] \leq \alpha\|\nabla F(w)\|_2^2.$$
- Weak growth condition (WGC), (Vaswani et al., 2019):
$$\mathbb{E}[\|G_t(w)\|_2^2] \leq 2\alpha(F(w) - F_*).$$
- Relaxed growth condition (RGC), (Bottou et al., 2018):
$$\mathbb{E}[\|G_t(w)\|_2^2] \leq \alpha\|\nabla F(w)\|_2^2 + \beta.$$
- Expected smoothness (ES), (Gower et al., 2021b, 2019):
$$\mathbb{E}[\|G_t(w) - G_t(w_*)\|_2^2] \leq 2\alpha(F(w) - F_*),$$
where $w_* = \arg\min_{w\in\mathbb{R}^d} F(w)$.

There are obvious relations among the above conditions. For instance,

- (BG) implies (BV).
- (BV)/(SGC) implies (RGC).
- (SGC) with Lipschitz smoothness implies (WGC).
- (WGC) with PL-inequality implies (SGC).
- (WGC) implies (ES) since $G_t(w_*) = 0$ under (WGC).

Moreover, (WGC) and (SGC) imply the interpolation condition, that is, $G_t(w_*) = 0$ (a.e.). In particular, all stochastic gradients $G_t(w_*) = 0$ for finite-sum problems. This means the vanishing of gradient and stochastic noise at the solution w_*.

Recently, Khaled and Richtárik (2020) proposed the following general condition.

Assumption 4 (ABC condition[1])**.** *There exist constants A, B, and $C \geq 0$ such that for any $w \in \mathbb{R}^d$,*

$$\mathbb{E}[\|G_t(w)\|_2^2] \leq 2A(F(w) - F_*) + B\|\nabla F(w)\|_2^2 + C.$$

[1]This condition is named the expected smoothness in Khaled and Richtárik (2020), but we refer it to as ABC condition according to Gower et al. (2021a) to avoid confusion.

We can easily see that all conditions (BG), (BV), (SGC), (WGC), (RGC), and (ES) imply ABC condition. The following proposition provides an example of this condition.

Proposition 1 (Minibatch stochastic gradient)**.** *Let us consider the risk minimization in Example 1. We suppose each loss $\ell(w,z)$ is β-Lipschitz smooth in w. Let $b \in \mathbb{Z}$ be the minibatch size. For independent copies $Z_t^{(1)}, \dots, Z_t^{(b)}$ of Z_t, we define the minibatch variant of the stochastic gradients as follows:*

$$G_t(w) = \frac{1}{b}\sum_{j=1}^{b} \nabla_w \ell(w, Z_t^{(j)}).$$

Then, Assumption 4 is satisfied with $A = \frac{\beta}{b}, B = \frac{b-1}{b}$, and $C = \frac{2\beta}{b}(F_* - l_*)$, where $\ell_* = \inf_{w,z} \ell(w,z)$.

Proof. Since $\ell(w, Z_t^{(j)})$ is β-Lipschitz smooth in w, it follows that for any $w, w' \in \mathbb{R}^d$,

$$\ell(w', Z_t^{(j)}) \le \ell(w, Z_t^{(j)}) + \nabla_w \ell(w, Z_t^{(j)})^\top (w' - w) + \frac{\beta}{2}\|w' - w\|_2^2.$$

By minimizing both sides with respect to w', we get

$$\ell_* \le \ell(w, Z_t^{(j)}) - \frac{1}{2\beta}\|\nabla_w \ell(w, Z_t^{(j)})\|_2^2.$$

By taking the expectation, we get

$$\mathbb{E}[\|\nabla_w \ell(w, Z_t^{(j)})\|_2^2] \le 2\beta(F(w) - l_*). \tag{2}$$

Next, we evaluate $\mathbb{E}[\|G_t(w)\|_2^2]$ as follows:

$$\begin{aligned}
\mathbb{E}[\|G_t(w)\|_2^2] &= \mathbb{E}\left[\frac{1}{b^2}\left\|\sum_{j=1}^{b} \nabla_w \ell(w, Z_t^{(j)})\right\|_2^2\right] \\
&= \mathbb{E}\left[\frac{1}{b^2}\sum_{j=1}^{b}\left\|\nabla_w \ell(w, Z_t^{(j)})\right\|_2^2 + \frac{1}{b^2}\sum_{i\neq j} \nabla_w \ell(w, Z_t^{(i)})^\top \nabla_w \ell(w, Z_t^{(j)})\right] \\
&= \frac{1}{b}\mathbb{E}\left[\|\nabla_w \ell(w, Z_t)\|_2^2\right] + \frac{b-1}{b}\left\|\mathbb{E}\left[\nabla_w \ell(w, Z_t)\right]\right\|_2^2. \qquad (3)
\end{aligned}$$

Combining (2) and (3), we get

$$\begin{aligned}
\mathbb{E}[\|G_t(w)\|_2^2] &\le \frac{2\beta}{b}(F(w) - l_*) + \frac{b-1}{b}\|\nabla F(w)\|_2^2 \\
&= \frac{2\beta}{b}(F(w) - F_*) + \frac{b-1}{b}\|\nabla F(w)\|_2^2 + \frac{2\beta}{b}(F_* - l_*).
\end{aligned}$$

□

The counterpart of Proposition 1 for the finite-sum setting was obtained by Sebbouh et al. (2021). These are basically extensions of the result of Bassily et al. (2018) to non-interpolation settings. Indeed, the linear convergence rate obtained by Bassily et al. (2018) can be recovered in the interpolation setting (i.e., $F_* = l_* \Leftrightarrow C = 0$). See Theorem 1.

We next give a convergence analysis of SGD with simple step-size schedules. The proof technique is based on that of Bottou et al. (2018).

Theorem 1. *Suppose Assumptions 2, 3, and 4 hold. We take step size η_t so that $\eta_t \leq \min\{\frac{\mu}{2AL}, \frac{1}{2BL}\}$. Then, the following results hold for constant step-size and decreasing step-size schedules.*

Case of constant step-size*: for the constant step-size $\eta_t = \eta$, it follows that*

$$\mathbb{E}[F(w_T)] - F_* \leq \frac{\eta LC}{2\mu} + (1 - \eta\mu)^{T-1} \left(F(w_1) - F_*\right).$$

Case of decreasing step-size*: for the decreasing step-size $\eta_t = \frac{2}{\mu(\gamma+t)}$ where $\gamma > 0$ is a hyper-parameter which should be chosen so that η_t satisfies the above conditions, it follows that*

$$\mathbb{E}[F(w_T)] - F_* \leq \frac{\nu}{\gamma + T},$$

where ν is a constant defined below:

$$\nu = \max\left\{(\gamma+1)(F(w_1) - F_*), \frac{2LC}{\mu^2}\right\}.$$

Proof. By applying Lipschitz smoothness (Assumption 2), ABC condition (Assumption 4), and PL-inequality (Assumption 3), we get $F(w_{t+1}) \leq F(w_t) - \eta \nabla F(w_t)^\top G(w_t) + \frac{L}{2}\|G(w_t)\|_2^2$. Hence,

$$\begin{aligned}
\mathbb{E}[F(w_{t+1})] &\leq \mathbb{E}[F(w_t)] - \eta_t \mathbb{E}[\|\nabla F(w_t)\|_2^2] + \frac{\eta_t^2 L}{2}\mathbb{E}\left[\|G_t(w_t)\|_2^2\right] \\
&\leq \mathbb{E}[F(w_t)] - \eta_t\left(1 - \frac{\eta_t BL}{2}\right)\mathbb{E}[\|\nabla F(w_t)\|_2^2] + \eta_t^2 LA(\mathbb{E}[F(w_t)] - F_*) + \frac{\eta_t^2 LC}{2} \\
&\leq \mathbb{E}[F(w_t)] - \frac{3\eta_t}{4}\mathbb{E}[\|\nabla F(w_t)\|_2^2] + \eta_t^2 LA(\mathbb{E}[F(w_t)] - F_*) + \frac{\eta_t^2 LC}{2} \\
&\leq \mathbb{E}[F(w_t)] - \eta_t\left(\frac{3\mu}{2} - \eta_t LA\right)(\mathbb{E}[F(w_t)] - F_*) + \frac{\eta_t^2 LC}{2} \\
&\leq \mathbb{E}[F(w_t)] - \eta_t\mu(\mathbb{E}[F(w_t)] - F_*) + \frac{\eta_t^2 LC}{2}.
\end{aligned}$$

By subtracting F_* from both sides, we get

$$\mathbb{E}[F(w_{t+1})] - F_* \leq (1 - \eta_t\mu)(\mathbb{E}[F(w_t)] - F_*) + \frac{\eta_t^2 LC}{2}. \tag{4}$$

Next, we consider two step-size schedules separately.

Case of constant step-size ($\eta_t = \eta$):
Using 4, we get

$$\begin{aligned}
\mathbb{E}[F(w_{t+1})] - F_* - \frac{\eta LC}{2\mu} &\leq (1 - \eta\mu)\left(\mathbb{E}[F(w_t)] - F_* - \frac{\eta LC}{2\mu}\right) \\
&\leq (1 - \eta\mu)^t\left(F(w_1) - F_* - \frac{\eta LC}{2\mu}\right).
\end{aligned}$$

This means

$$\mathbb{E}[F(w_T)] - F_* \leq \frac{\eta LC}{2\mu} + (1 - \eta\mu)^{T-1}\left(F(w_1) - F_*\right).$$

Case of decreasing step-size ($\eta_t = 2/\mu(\gamma+t)$):
We show the following bound by induction:

$$\mathbb{E}[F(w_t)] - F_* \leq \frac{\nu}{\gamma + t}. \tag{5}$$

This bound clearly holds for $t = 1$. Next, we suppose it holds for t. We denote $\hat{t} = \gamma + t$ for simplicity. Then, by (4) we see

$$\begin{aligned}
\mathbb{E}[F(w_t)] - F_* &\leq \left(1 - \frac{2}{\hat{t}}\right)(\mathbb{E}[F(w_t)] - F_*) + \frac{2LC}{\mu^2 \hat{t}^2} \\
&\leq \left(1 - \frac{2}{\hat{t}}\right)\frac{\nu}{\hat{t}} + \frac{2LC}{\mu^2 \hat{t}^2} \\
&= \frac{\hat{t}-1}{\hat{t}^2}\nu - \frac{\nu}{\hat{t}^2} + \frac{2LC}{\mu^2 \hat{t}^2} \\
&\leq \frac{\nu}{\hat{t}},
\end{aligned}$$

where we used $\nu \geq \frac{2LC}{\mu^2}$ for the last inequality. This proves (5) with $t+1$ and concludes the proof. □

Theorem 1 can derive the complexity (number of iterations), under an appropriate choice of step-size, to achieve an ϵ-accurate solution: $\mathbb{E}[F(w_T)] - F_* \leq \epsilon$. The required complexity for the constant step-size schedule is

$$T = O\left(\max\left\{\frac{AL}{\mu^2}, \frac{BL}{\mu}, \frac{CL}{\mu^2 \epsilon}\right\} \log\left(\frac{1}{\epsilon}\right)\right), \tag{6}$$

and for the decreasing step-size schedule

$$T = O\left(\max\left\{\frac{AL}{\mu^2}, \frac{BL}{\mu}, \frac{CL}{\mu^2}\right\} \frac{1}{\epsilon}\right). \tag{7}$$

We note that the logarithmic factor of (6) can be improved by using a refined step-size schedule. For the detail, see Khaled and Richtárik (2020). Moreover, we note that the complexity of (6) implies the linear convergence in the interpolation setting (i.e., $C = 0$).

References

Bassily, R., Belkin, M., and Ma, S. (2018). On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Gower, R., Sebbouh, O., and Loizou, N. (2021a). Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209.

Gower, R. M., Richtárik, P., and Bach, F. (2021b). Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188(1):135–192.

Khaled, A. and Richtárik, P. (2020). Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*.

Sebbouh, O., Gower, R. M., and Defazio, A. (2021). Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204.