Analysis of Stochastic Gradient Descent

Materials from *Gartner, B., He, N., and Jaggi, M. Lectures notes on Optimization for Data Science.*

May 19, 2026

Stochastic Optimization

General form:

$$\min_{\mathbf{x}\in\mathbf{C}} \; F(x) = \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \qquad \text{(Stochastic Optimization)}$$

where $f(\mathbf{x}, \boldsymbol{\xi})$ is a function involving the decision variable $\mathbf{x}$ and a random variable (vector) $\boldsymbol{\xi} \sim \mathbb{P}(\boldsymbol{\xi})$.

Stochastic Optimization

General form:

$$\min_{\mathbf{x}\in\mathsf{C}} \; F(x) = \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \qquad \text{(Stochastic Optimization)}$$

where $f(\mathbf{x}, \boldsymbol{\xi})$ is a function involving the decision variable $\mathbf{x}$ and a random variable (vector) $\boldsymbol{\xi} \sim \mathbb{P}(\boldsymbol{\xi})$.

In particular, given a set of function $\{f_i(\mathbf{x})\}_{i=1}^n$ and let $\boldsymbol{\xi} \sim \texttt{Unif}\{1, \dots, n\}$. Define $f(\mathbf{x}, \boldsymbol{\xi}) = f_{\boldsymbol{\xi}}(\mathbf{x})$, then

$$\min_{\mathbf{x}\in\mathsf{C}} \; F(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \qquad \text{(Finte-Sum Problem)}$$

Stochastic Gradient Descent

Assume that $f(\mathbf{x}, \boldsymbol{\xi})$ is continuously differentiable for any instance of $\boldsymbol{\xi}$. The Stochastic Gradient Descent (SGD) updates the numerical solution by

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t) \right),$$

where $\boldsymbol{\xi}_t \sim \mathbb{P}(\boldsymbol{\xi})$ and ∇ is taken over the argument $\mathbf{x}$, starting from a deterministic initialization $\mathbf{x}_1$.

Stochastic Gradient Descent

Assume that $f(\mathbf{x}, \boldsymbol{\xi})$ is continuously differentiable for any instance of $\boldsymbol{\xi}$. The Stochastic Gradient Descent (SGD) updates the numerical solution by

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t) \right),$$

where $\boldsymbol{\xi}_t \sim \mathbb{P}(\boldsymbol{\xi})$ and ∇ is taken over the argument $\mathbf{x}$, starting from a deterministic initialization $\mathbf{x}_1$.

The corresponding update formula of solving (Finte-Sum Problem) by SGD is written as

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f_{i_t}(\mathbf{x}_t) \right), \quad i_t \sim \texttt{Unif}\{1, \dots, n\}.$$

Stochastic Gradient Descent

Assume that $f(\mathbf{x}, \boldsymbol{\xi})$ is continuously differentiable for any instance of $\boldsymbol{\xi}$. The Stochastic Gradient Descent (SGD) updates the numerical solution by

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t) \right),$$

where $\boldsymbol{\xi}_t \sim \mathbb{P}(\boldsymbol{\xi})$ and ∇ is taken over the argument $\mathbf{x}$, starting from a deterministic initialization $\mathbf{x}_1$.

The corresponding update formula of solving (Finte-Sum Problem) by SGD is written as

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f_{i_t}(\mathbf{x}_t) \right), \quad i_t \sim \texttt{Unif}\{1, \dots, n\}.$$

Note that

$$\mathbb{E}[\nabla f_{i_t}(\mathbf{x}_t) | \mathbf{x}_t] = \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t) \cdot \mathbb{P}(i_t = i) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t) = \nabla F(\mathbf{x}_t)$$

Stochastic Gradient Descent

In our analysis, we always assume that

$$\mathbb{E}[\nabla f(\mathbf{x}, \boldsymbol{\xi})] = \int \nabla f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{P}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \nabla \int f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{P}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \nabla F(\mathbf{x}).$$

Stochastic Gradient Descent

Example

For ordinary least squares regression,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2, \quad f_i(\mathbf{x}) = (\mathbf{a}_i^T\mathbf{x} - b_i)^2$$

where $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ is the set of training samples.

Stochastic Gradient Descent

Example

For ordinary least squares regression,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2, \quad f_i(\mathbf{x}) = (\mathbf{a}_i^T\mathbf{x} - b_i)^2$$

where $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ is the set of training samples.

$$\texttt{Gradient Descent (GD)}: \quad \mathbf{x}_{t+1} = \mathbf{x}_t - 2\gamma_t \sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)\mathbf{a}_i$$

Stochastic Gradient Descent

Example

For ordinary least squares regression,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2, \quad f_i(\mathbf{x}) = (\mathbf{a}_i^T\mathbf{x} - b_i)^2$$

where $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ is the set of training samples.

$$\texttt{Gradient Descent (GD)}: \quad \mathbf{x}_{t+1} = \mathbf{x}_t - 2\gamma_t \sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)\mathbf{a}_i$$

$$\texttt{SGD}: \quad \mathbf{x}_{t+1} = \mathbf{x}_t - 2\gamma_t(\mathbf{a}_{i_t}^T\mathbf{x} - b_{i_t})\mathbf{a}_{i_t}, \quad i_t \sim \texttt{Unif}\{1, \dots, n\}$$

Stochastic Gradient Descent

Example

For ordinary least squares regression,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)^2, \quad f_i(\mathbf{x}) = (\mathbf{a}_i^T\mathbf{x} - b_i)^2$$

where $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ is the set of training samples.

$$\texttt{Gradient Descent (GD)}: \quad \mathbf{x}_{t+1} = \mathbf{x}_t - 2\gamma_t \sum_{i=1}^{n}(\mathbf{a}_i^T\mathbf{x} - b_i)\mathbf{a}_i$$

$$\texttt{SGD}: \quad \mathbf{x}_{t+1} = \mathbf{x}_t - 2\gamma_t(\mathbf{a}_{i_t}^T\mathbf{x} - b_{i_t})\mathbf{a}_{i_t}, \quad i_t \sim \texttt{Unif}\{1, \dots, n\}$$

- **GD: n samples are used in per-iteration.**
- **SGD: one sample is used in per-iteration.**

Stochastic Gradient Descent

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}} \left(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t) \right)$$

- Since $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is a random variable, we cannot guarantee that $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t) \to 0$ as $t \to \infty$. Therefore, $\gamma_t \to 0$ as $t \to \infty$ is needed to ensure convergence.
- $\{\mathbf{x}_t\}_{t \geq 1}$ is a random process of which randomness originates from $\boldsymbol{\xi}_t$, and so is $\{F(\mathbf{x}_t)\}_{t \geq 1}$.

Convergence for Strongly Convex Functions

Theorem 1 (Convergent Step Size)

Assume that $F(\mathbf{x})$ is μ-strongly convex, and $\exists M > 0$, s.t. $\mathbb{E}[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|_2^2] \leq M^2$, $\forall \mathbf{x} \in \mathsf{C}$, then SGD with $\gamma_t = \frac{\gamma}{t}$ at iteration t where $\gamma > 1/(2\mu)$ satisfies

$$\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|^2\right] \leq \frac{C(\gamma)}{t}, \quad \text{where } C(\gamma) = \frac{\gamma^2 M^2}{2\mu\gamma - 1}$$

Proof of Theorem 1.

Let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$. Given $\mathbf{x}_t$, by the non-expansive property of the projection operator, we have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\Pi_{\mathsf{C}}(\mathbf{x}_t - \gamma_t \mathbf{g}_{\boldsymbol{\xi}_t}) - \mathbf{x}^*\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^* - \gamma_t \mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\gamma_t \langle \mathbf{g}_{\boldsymbol{\xi}_t}, \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma_t^2 \|\mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \quad \text{a.s.}
\end{aligned}$$

Proof of Theorem 1.

Let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$. Given $\mathbf{x}_t$, by the non-expansive property of the projection operator, we have

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\Pi_{\mathsf{C}}(\mathbf{x}_t - \gamma_t \mathbf{g}_{\boldsymbol{\xi}_t}) - \mathbf{x}^*\|^2 \\ &\leq \|\mathbf{x}_t - \mathbf{x}^* - \gamma_t \mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\gamma_t \langle \mathbf{g}_{\boldsymbol{\xi}_t}, \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma_t^2 \|\mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \quad \text{a.s.}\end{aligned}$$

Taking expectation both sides conditioned on $\mathbf{x}_t$ yields

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 | \mathbf{x}_t] \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\gamma_t \langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma_t^2 M^2$$

□

Proof of Theorem 1.

By strong convexity of $F(\mathbf{x})$, we have

$$\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle \geq F(\mathbf{x}_t) - F(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2,$$

and

$$\begin{aligned} F(\mathbf{x}_t) - F(\mathbf{x}^*) &\geq \langle \nabla F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \\ &\geq \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \end{aligned}$$

where the last inequality comes from the optimality of $\mathbf{x}^*$.

Proof of Theorem 1.

By strong convexity of $F(\mathbf{x})$, we have

$$\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle \geq F(\mathbf{x}_t) - F(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2,$$

and

$$\begin{aligned} F(\mathbf{x}_t) - F(\mathbf{x}^*) &\geq \langle \nabla F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \\ &\geq \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \end{aligned}$$

where the last inequality comes from the optimality of $\mathbf{x}^*$.

Putting together, we have

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 | \mathbf{x}_t] \leq (1 - 2\mu\gamma_t)\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \gamma_t^2 M^2.$$

□

Proof of Theorem 1.

And by the tower formula, we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|^2] &\leq (1-2\mu\gamma_t)\mathbb{E}[\|\mathbf{x}_t-\mathbf{x}^*\|^2]+\gamma_t^2M^2\\ &\leq (1-\frac{2\mu\gamma}{t})\mathbb{E}[\|\mathbf{x}_t-\mathbf{x}^*\|^2]+\frac{\gamma^2M^2}{t^2}.\end{aligned}$$

Proof of Theorem 1.

And by the tower formula, we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|^2] &\leq (1-2\mu\gamma_t)\mathbb{E}[\|\mathbf{x}_t-\mathbf{x}^*\|^2]+\gamma_t^2M^2\\ &\leq (1-\frac{2\mu\gamma}{t})\mathbb{E}[\|\mathbf{x}_t-\mathbf{x}^*\|^2]+\frac{\gamma^2M^2}{t^2}.\end{aligned}$$

Let $t=1$,

$$0\leq \mathbb{E}[\|\mathbf{x}_2-\mathbf{x}^*\|^2]\leq (1-2\mu\gamma)\|\mathbf{x}_1-\mathbf{x}^*\|^2+\gamma^2M^2.$$

Since $\gamma > 1/(2\mu)$,

$$\|\mathbf{x}_1-\mathbf{x}^*\|^2\leq \frac{\gamma^2M^2}{2\mu\gamma-1}=\frac{C(\gamma)}{1}.$$

We complete the proof by induction. □

Convergence for Strongly Convex Functions

Theorem 2 (Constant Step Size)

Assume that $F(\mathbf{x})$ is both μ-strongly convex and L-smooth. Moreover, assume that stochastic gradient satisfies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|_2^2\right] \leq \sigma^2 + c\|\nabla F(\mathbf{x})\|^2.$$

Then, SGD with constant stepsize $\gamma_t \equiv \gamma \leq \frac{1}{cL}$ achieves:

$$\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \frac{\gamma L \sigma^2}{2\mu} + (1 - \gamma\mu)^{t-1} \left(F(\mathbf{x}_1) - F(\mathbf{x}^*)\right),$$

where $\mathbf{x}^*$ is the optimal solution and $\mathsf{C} = \mathbb{R}^d$.

- $\lim_{t\to\infty} (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \leq \frac{\gamma L \sigma^2}{2\mu}$
- **What if $\mathsf{C} \subset \mathbb{R}^d$?**

Proof of Theorem 2.

Since $\mathsf{C} = \mathbb{R}^d$, given $\mathbf{x}_t$ and let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$, we have

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}}(\mathbf{x}_t - \gamma \mathbf{g}_{\boldsymbol{\xi}_t}) = \mathbf{x}_t - \gamma \mathbf{g}_{\boldsymbol{\xi}}, \quad \boldsymbol{\xi}_t \sim \texttt{Unif}\{1, \dots, n\}.$$

And the L-smoothness of $F(\mathbf{x})$ yields

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) &\leq \langle \nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= -\gamma \langle \nabla F(\mathbf{x}_t), \mathbf{g}_{\boldsymbol{\xi}_t} \rangle + \frac{L\gamma^2}{2} \|\mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \quad \text{a.s.} \end{aligned}$$

Proof of Theorem 2.

Since $\mathsf{C} = \mathbb{R}^d$, given $\mathbf{x}_t$ and let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$, we have

$$\mathbf{x}_{t+1} = \Pi_{\mathsf{C}}(\mathbf{x}_t - \gamma \mathbf{g}_{\boldsymbol{\xi}_t}) = \mathbf{x}_t - \gamma \mathbf{g}_{\boldsymbol{\xi}}, \quad \boldsymbol{\xi}_t \sim \texttt{Unif}\{1, \dots, n\}.$$

And the L-smoothness of $F(\mathbf{x})$ yields

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) &\leq \langle \nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= -\gamma \langle \nabla F(\mathbf{x}_t), \mathbf{g}_{\boldsymbol{\xi}_t} \rangle + \frac{L\gamma^2}{2} \|\mathbf{g}_{\boldsymbol{\xi}_t}\|^2 \quad \text{a.s.} \end{aligned}$$

Taking expectation both sides conditioned on $\mathbf{x}_t$ yields

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \langle \nabla F(\mathbf{x}_t), \mathbb{E}[\mathbf{g}_{\boldsymbol{\xi}_t}] \rangle + \frac{L\gamma^2}{2} \mathbb{E}[\|\mathbf{g}_{\boldsymbol{\xi}_t}\|^2].$$

□

Proof of Theorem 2.

By the assumption and $\mathbb{E}[\mathbf{g}_{\boldsymbol{\xi}_t}] = \nabla F(\mathbf{x}_t)$, we get

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{c\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2}. \qquad (\spadesuit)$$

Proof of Theorem 2.

By the assumption and $\mathbb{E}[\mathbf{g}_{\boldsymbol{\xi}_t}] = \nabla F(\mathbf{x}_t)$, we get

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{c\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2}. \qquad (\spadesuit)$$

Since $F(x)$ is μ-strongly convex,

$$\frac{1}{2\mu}\|F(\mathbf{x}_t)\|^2 \geq F(\mathbf{x}_t) - F(\mathbf{x}^*) \quad \text{a.s.} \qquad (\diamondsuit)$$

Proof of Theorem 2.

By the assumption and $\mathbb{E}[\mathbf{g}_{\boldsymbol{\xi}_t}] = \nabla F(\mathbf{x}_t)$, we get

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{c\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2}. \qquad (\spadesuit)$$

Since $F(x)$ is μ-strongly convex,

$$\frac{1}{2\mu}\|F(\mathbf{x}_t)\|^2 \geq F(\mathbf{x}_t) - F(\mathbf{x}^*) \quad \text{a.s.} \qquad (\diamondsuit)$$

Define $\Delta_t = F(\mathbf{x}_t) - F(\mathbf{x}_*)$. Note that $\gamma \leq \frac{1}{cL}$ and hence $1 - \frac{\gamma c L}{2} \geq 0$.

Proof of Theorem 2.

By the assumption and $\mathbb{E}[\mathbf{g}_{\boldsymbol{\xi}_t}] = \nabla F(\mathbf{x}_t)$, we get

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{c\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2}. \qquad (\spadesuit)$$

Since $F(x)$ is μ-strongly convex,

$$\frac{1}{2\mu}\|F(\mathbf{x}_t)\|^2 \geq F(\mathbf{x}_t) - F(\mathbf{x}^*) \quad \text{a.s.} \qquad (\diamondsuit)$$

Define $\Delta_t = F(\mathbf{x}_t) - F(\mathbf{x}_*)$. Note that $\gamma \leq \frac{1}{cL}$ and hence $1 - \frac{\gamma c L}{2} \geq 0$. Combing $(\spadesuit)$ and $(\diamondsuit)$ yields

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}] &\leq (1 - 2\gamma\mu + \gamma^2 c\mu L)\, \mathbb{E}[\Delta_t] + \frac{\gamma^2 L \sigma^2}{2} \\ &\leq (1 - \gamma\mu)\, \mathbb{E}[\Delta_t] + \frac{\gamma^2 L \sigma^2}{2}.\end{aligned}$$

□

Proof of Theorem 2.

The above inequality can be rearranged as

$$\mathbb{E}[\Delta_{t+1}] - \frac{\gamma L \sigma^2}{2\mu} \leq (1 - \gamma\mu) \left[\mathbb{E}[\Delta_t] - \frac{\gamma L \sigma^2}{2\mu} \right]$$

Proof of Theorem 2.

The above inequality can be rearranged as

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}] - \frac{\gamma L \sigma^2}{2\mu} &\leq (1-\gamma\mu)\left[\mathbb{E}[\Delta_t] - \frac{\gamma L \sigma^2}{2\mu}\right] \\ &\leq (1-\gamma\mu)^t \left[\mathbb{E}[\Delta_1] - \frac{\gamma L \sigma^2}{2\mu}\right] \\ &\leq (1-\gamma\mu)^t \mathbb{E}[\Delta_1] \quad \forall t \geq 1\end{aligned}$$

□

Convergence for Nonconvex Functions

Theorem 3

Assume that $F(\mathbf{x})$ is L-smooth and the stochastic gradient has bounded variance, namely, $\mathbb{E}\|\nabla f(\mathbf{x}, \boldsymbol{\xi}) - \nabla F(\mathbf{x})\|^2 \leq \sigma^2$. Then SGD with stepsize $\gamma_t \equiv \gamma := \min\{\frac{1}{L}, \frac{\gamma_0}{\sigma\sqrt{T}}\}$ achieves

$$\mathbb{E}\left[\|\nabla F(\hat{\mathbf{x}}_T)\|^2\right] \leq \frac{2L(F(\mathbf{x}_1) - F(\mathbf{x}^*))}{T} + \frac{\sigma}{\sqrt{T}}\left(\frac{2(F(\mathbf{x}_1) - F(\mathbf{x}^*))}{\gamma_0} + L\gamma_0\right)$$

where $\hat{\mathbf{x}}_T$ is selected uniformly at random from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ and $\mathsf{C} = \mathbb{R}^d$.

Proof of Theorem 3.

From the bounded variance of the stochastic gradient, we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|^2] \leq \sigma^2 + \|\nabla F(\mathbf{x})\|^2$$

Proof of Theorem 3.

From the bounded variance of the stochastic gradient, we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|^2] \leq \sigma^2 + \|\nabla F(\mathbf{x})\|^2$$

Given $\mathbf{x}_t$ and let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$, following the proof of Theorem 2 yields

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2},$$

which can be rearranged as

$$\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})] + \frac{\gamma^2 L \sigma^2}{2}$$

Proof of Theorem 3.

From the bounded variance of the stochastic gradient, we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|^2] \leq \sigma^2 + \|\nabla F(\mathbf{x})\|^2$$

Given $\mathbf{x}_t$ and let $\mathbf{g}_{\boldsymbol{\xi}_t} = \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_t)$, following the proof of Theorem 2 yields

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2},$$

which can be rearranged as

$$\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla F(\mathbf{x}_t)\|^2 \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})] + \frac{\gamma^2 L \sigma^2}{2}$$

Note that $\gamma \leq \frac{1}{L}$, and hence $1 - \frac{\gamma L}{2} \geq \frac{1}{2}$.

$$\frac{\gamma}{2} \|\nabla F(\mathbf{x}_t)\|^2 \leq F(\mathbf{x}_t) - F(\mathbf{x}^*) - (\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)]) + \frac{\gamma^2 L \sigma^2}{2}.$$

Proof of Theorem 3.

Taking expectation and telescoping, we get

$$\frac{\gamma}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq F(\mathbf{x}_1) - F(\mathbf{x}^*) - (\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)]) + \frac{\gamma^2 L\sigma^2}{2}\cdot T$$

$$\leq F(\mathbf{x}_1) - F(\mathbf{x}^*) + \frac{\gamma^2 L\sigma^2}{2}\cdot T$$

Proof of Theorem 3.

Taking expectation and telescoping, we get

$$\frac{\gamma}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq F(\mathbf{x}_1) - F(\mathbf{x}^*) - (\mathbb{E}[F(\mathbf{x}_{T+1}) - F(\mathbf{x}^*)]) + \frac{\gamma^2 L\sigma^2}{2}\cdot T$$
$$\leq F(\mathbf{x}_1) - F(\mathbf{x}^*) + \frac{\gamma^2 L\sigma^2}{2}\cdot T$$

And

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \frac{2F(\mathbf{x}_1) - F(\mathbf{x}^*)}{T\gamma} + \gamma L\sigma^2$$

Since $\gamma := \min\{\frac{1}{L}, \frac{\gamma_0}{\sigma\sqrt{T}}\}$, $\gamma \leq \frac{\gamma_0}{\sigma\sqrt{T}}$ and $\frac{1}{r} = \max\{L, \frac{\sigma\sqrt{T}}{\gamma_0}\}$ □

Proof of Theorem 3.

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{T}\|\nabla F(\mathbf{x}_t)\|^2\right] \leq \frac{2F(\mathbf{x}_1)-F(\mathbf{x}^*)}{T}\cdot\max\{L,\frac{\sigma\sqrt{T}}{\gamma_0}\}+L\sigma^2\cdot\frac{\gamma_0}{\sigma\sqrt{T}}$$
$$\leq \frac{2L(F(\mathbf{x}_1)-F(\mathbf{x}^*))}{T}+\frac{\sigma}{\sqrt{T}}\left(\frac{2(F(\mathbf{x}_1)-F(\mathbf{x}^*))}{\gamma_0}+L\gamma_0\right)$$

□

Complexity

Table: Complexity comparison for GD and SGD, $\kappa = \frac{L}{\mu}$.

Problem class	Method	Iter. complexity	Iter. cost	Total
Strongly convex and smooth	GD	$\mathcal{O}(\kappa \log(1/\varepsilon))$	$\mathcal{O}(n)$	$\mathcal{O}(n\kappa \log(1/\varepsilon))$
	SGD	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(1)$	$\mathcal{O}(1/\varepsilon)$
Non-convex	GD	$\mathcal{O}(1/\varepsilon^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n/\varepsilon^2)$
	SGD	$\mathcal{O}(1/\varepsilon^4)$	$\mathcal{O}(1)$	$\mathcal{O}(1/\varepsilon^4)$

Remark on Theorem 2 and Theorem 3

Both theorems depend on the bounded variance assumption

$$\mathbb{E}[\|\nabla f(\mathbf{x}, \boldsymbol{\xi})\|^2] \leq \sigma^2 + \|\nabla F(\mathbf{x})\|^2,$$

where σ^2 manifests the noisy level. The noise gradient estimate prevents convergence to the optimal in Theorem 2. A larger σ^2 may also lead to a worse bound in Theorem 3. Noise reduction methods are developed as a remedy; see (Bottou, Curtis, and Nocedal, 2018, Section 5).

References

[BCN2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223–311.

The following notes are not used in the lecture.

Noise Reduction Methods

Consider SGD for stochastic optimization.

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; F(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \qquad \text{(Stochastic Optimization)}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t) \qquad \text{(SGD)}$$

where $\boldsymbol{\xi}_t$ is generated from some distribution $\mathbb{P}(\boldsymbol{\xi}_t)$ independently at every iteration.

Given $\mathbf{x}_t$, the variance of random vector $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is defined as

$$\mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] = \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] - \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)]\|^2$$

Given $\mathbf{x}_t$, the variance of random vector $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is defined as

$$\mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] = \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] - \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)]\|^2$$

Assumption 1 (First and second moment limits)

(a) There exists scalars $\beta_G \geq \beta > 0$ such that

$$\nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \geq \beta \|\nabla F(\mathbf{x}_t)\|^2 \quad \text{and} \quad \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_k)]\| \leq \beta_G \|\nabla F(\mathbf{x}_t)\|.$$

(b) There exits scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $t \in \mathbb{N}$

$$\mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M + M_V \|\nabla F(\mathbf{x}_t)\|^2.$$

Given $\mathbf{x}_t$, the variance of random vector $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is defined as

$$\mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] = \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] - \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)]\|^2$$

Assumption 1 (First and second moment limits)

(a) There exists scalars $\beta_G \geq \beta > 0$ such that

$$\nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \geq \beta \|\nabla F(\mathbf{x}_t)\|^2 \quad \text{and} \quad \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_k)]\| \leq \beta_G \|\nabla F(\mathbf{x}_t)\|.$$

(b) There exits scalars $M \geq 0$ and $M_V \geq 0$ such that, for all $t \in \mathbb{N}$

$$\mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M + M_V \|\nabla F(\mathbf{x}_t)\|^2.$$

Combing Assumption 1 and the definition of variance yields

$$\mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] \leq M + M_G \|\nabla F(\mathbf{x}_t)\|^2,$$

where $M_G = M_V + \beta_G^2 \geq \beta^2 > 0$

Noise Reduction Methods

- `Dynamic sampling methods:`to achieve noise reduction by gradually increasing the mini-batch size used in the gradient computation.
- `Gradient aggregation methods:`store gradient estimates corresponding to samples employed in previous iterations and define the search direction as weighted as a weighted average of these estimates.
- `Iterate averaging methods:`maintain an average of iterates computed during the optimization process.

Reducing Noise at a Geometric Rate

Remark on SGD of L-smooth function
Given $\mathbf{x}_t$, if $F(\mathbf{x})$ is L-smooth, SGD satisfies

$$\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq -\gamma_t \nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \frac{1}{2}\gamma_t^2 L \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2].$$

- If $-\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is a descent direction in expectation and if we are able to decrease $\mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2]$ fast enough, then the effect of having noisy directions will not impede a fast rate of convergence.

Reducing Noise at a Geometric Rate

Remark on SGD of L-smooth function
Given $\mathbf{x}_t$, if $F(\mathbf{x})$ is L-smooth, SGD satisfies

$$\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq -\gamma_t \nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \frac{1}{2}\gamma_t^2 L \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2].$$

- If $-\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ is a descent direction in expectation and if we are able to decrease $\mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2]$ fast enough, then the effect of having noisy directions will not impede a fast rate of convergence.
- Such behavior is expected if, in Assumption 1, we suppose instead that the variance of $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ vanishes sufficiently quickly.

Reducing Noise at a Geometric Rate

Theorem 4

Let $F(\mathbf{x})$ be μ-strongly convex and L-smooth. Assume that the Assumption 1 holds with the variance assumption refined to the existence of constant $M \geq 0$ and $\zeta \in (0,1)$ such that, for all $t \in \mathbb{N}$,

$$\texttt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M\zeta^{t-1}. \tag{♯}$$

In addition, suppose that SGD is run with a fixed stepsize $\gamma_t = \gamma$ satisfying

$$0 < \gamma \leq \min\left\{\frac{\beta}{L\beta_G^2}, \frac{1}{\mu\beta}\right\} \tag{🍄}$$

Then, for all $t \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \omega\rho^{t-1}, \tag{🔥}$$

where

$$\omega = \max\left\{\frac{\gamma LM}{\mu\beta}, F(\mathbf{x}_1) - F(\mathbf{x}^*)\right\} \quad \text{and} \quad \rho = \max\left\{1 - \frac{\gamma\mu\beta}{2}, \zeta\right\} < 1. \tag{🍆}$$

Proof of Theorem 4.

It can be shown that given $\mathbf{x}_t$, if $F(\mathbf{x})$ is L-smooth, SGD satisfies

$$\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq -\gamma \nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \frac{1}{2}\gamma^2 L \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2].$$

Proof of Theorem 4.

It can be shown that given $\mathbf{x}_t$, if $F(\mathbf{x})$ is L-smooth, SGD satisfies

$$\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq -\gamma \nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \frac{1}{2}\gamma^2 L \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2].$$

From Assumption 1(a) and (♯), we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] &= \mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)]\|^2 \\ &\leq M\zeta^{t-1} + \beta_G^2 \|\nabla F(\mathbf{x}_t)\|^2. \end{aligned}$$

Proof of Theorem 4.

It can be shown that given $\mathbf{x}_t$, if $F(\mathbf{x})$ is L-smooth, SGD satisfies

$$\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) \leq -\gamma \nabla F(\mathbf{x}_t)^T \mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \frac{1}{2}\gamma^2 L \mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2].$$

From Assumption 1(a) and (♯), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)\|^2] &= \mathtt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] + \|\mathbb{E}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)]\|^2 \\ &\leq M\zeta^{t-1} + \beta_G^2 \|\nabla F(\mathbf{x}_t)\|^2.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_t) &\leq -\gamma\beta\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\gamma^2 L \beta_G^2 \|\nabla F(\mathbf{x}_t)\|^2 + \frac{L}{2}\gamma^2 M \zeta^{t-1} \\ &\leq -\gamma\left(\beta - \frac{1}{2}L\beta_G^2\gamma\right)\|\nabla F(\mathbf{x}_t)\|^2 + \frac{L}{2}\gamma^2 M\zeta^{t-1}\end{aligned}$$

□

Proof of Theorem 4.

Since $F(\mathbf{x})$ is μ-strongly convex

$$\|\nabla F(\mathbf{x}_t)\|^2 \geq 2\mu \left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right)$$

Denote $\Delta_t = \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)]$.

Proof of Theorem 4.

Since $F(\mathbf{x})$ is μ-strongly convex

$$\|\nabla F(\mathbf{x}_t)\|^2 \geq 2\mu \left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right)$$

Denote $\Delta_t = \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)]$.

Combing the inequalities and rearranging yield

$$\Delta_{t+1} \leq (1 - \gamma\beta\mu)\Delta_t + \frac{L}{2}\gamma^2 M \zeta^{t-1},$$

where we apply the constraint $\gamma \leq \frac{\beta}{L\beta_G^2}$.

Proof of Theorem 4.

Since $F(\mathbf{x})$ is μ-strongly convex

$$\|\nabla F(\mathbf{x}_t)\|^2 \geq 2\mu \left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right)$$

Denote $\Delta_t = \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)]$.

Combing the inequalities and rearranging yield

$$\Delta_{t+1} \leq (1 - \gamma\beta\mu)\Delta_t + \frac{L}{2}\gamma^2 M \zeta^{t-1},$$

where we apply the constraint $\gamma \leq \frac{\beta}{L\beta_G^2}$.

We are going to prove (🔥) by induction, which obviously holds when $t = 1$. □

Proof of Theorem 4.

Suppose (🔥) holds at t-th iteration, we have

$$\begin{aligned}\Delta_{t+1} &\leq (1-\gamma\beta\mu)\Delta_t + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\ &\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\ &\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{\gamma\mu\beta}{2}\omega\zeta^{t-1} \quad (\omega \geq \gamma LM/(\mu\beta))\end{aligned}$$

Proof of Theorem 4.

Suppose (🔥) holds at t-th iteration, we have

$$
\begin{aligned}
\Delta_{t+1} &\leq (1-\gamma\beta\mu)\Delta_t + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{\gamma\mu\beta}{2}\omega\zeta^{t-1} \quad (\omega \geq \gamma LM/(\mu\beta)) \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{\gamma\mu\beta}{2}\omega\rho^{t-1} \quad (\rho \geq \zeta)
\end{aligned}
$$

Proof of Theorem 4.

Suppose (🔥) holds at t-th iteration, we have

$$\begin{aligned}
\Delta_{t+1} &\leq (1-\gamma\beta\mu)\Delta_t + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{L}{2}\gamma^2 M\zeta^{t-1} \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{\gamma\mu\beta}{2}\omega\zeta^{t-1} \quad (\omega \geq \gamma LM/(\mu\beta)) \\
&\leq (1-\gamma\beta\mu)\omega\rho^{t-1} + \frac{\gamma\mu\beta}{2}\omega\rho^{t-1} \quad (\rho \geq \zeta) \\
&= \left(1-\frac{\gamma\beta\mu}{2}\right)\omega\rho^{t-1} \\
&\leq \omega\rho^t. \quad \left(\rho \geq \left(1-\frac{\gamma\beta\mu}{2}\right)\right)
\end{aligned}$$

We complete the proof. □

Dynamic Sample Size Methods

One may ask how can we design a method that achieve the vanishing bounded variance (♯).

$$\texttt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M\zeta^{t-1}. \tag{♯}$$

For some $\tau > 1$, define the stochastic direction as

$$\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t) := \frac{1}{n_t} \sum_{i \in \mathsf{S}_t} \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i}) \tag{🍉}$$

with $n_t = |\mathsf{S}_t| = \lceil \tau^{t-1} \rceil$.

Dynamic Sample Size Methods

One may ask how can we design a method that achieve the vanishing bounded variance (♯).

$$\texttt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M\zeta^{t-1}. \tag{♯}$$

For some $\tau > 1$, define the stochastic direction as

$$\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t) := \frac{1}{n_t} \sum_{i \in \mathsf{S}_t} \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i}) \tag{🍉}$$

with $n_t = |\mathsf{S}_t| = \lceil \tau^{t-1} \rceil$.

Assume that $\{\boldsymbol{\xi}_{t,i}\}_{i \in \mathsf{S}_t}$ are independently drawn from $\mathbb{P}$ for all $t \in \mathbb{N}$ and the variance of $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i})$ is consistently bounded by $M > 0$.

Dynamic Sample Size Methods

One may ask how can we design a method that achieve the vanishing bounded variance ($\sharp$).

$$\texttt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq M\zeta^{t-1}. \tag{$\sharp$}$$

For some $\tau > 1$, define the stochastic direction as

$$\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t) := \frac{1}{n_t} \sum_{i \in \mathsf{S}_t} \nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i}) \tag{🍉}$$

with $n_t = |\mathsf{S}_t| = \lceil \tau^{t-1} \rceil$.

Assume that $\{\boldsymbol{\xi}_{t,i}\}_{i \in \mathsf{S}_t}$ are independently drawn from $\mathbb{P}$ for all $t \in \mathbb{N}$ and the variance of $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i})$ is consistently bounded by $M > 0$. It can shown that

$$\texttt{Var}[\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)] \leq \frac{M}{n_t},$$

and hence $\mathbf{g}(\mathbf{x}_t, \boldsymbol{\xi}_t)$ satisfies ($\sharp$).

Dynamic Sample Size Methods

Corollary 5

Assume that SGD is run with the stochastic direction (🍉) and $\mathbb{E}[\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i})]$ for all $i \in \mathsf{S}_t$ and $t \in \mathbb{N}$. Then the variance condition (♯) is satisfied, and if all other assumptions of Theorem 4 hold, then the expected optimality gap vanishes linearly in the sense of (🔥).

Dynamic Sample Size Methods

Total work complexity: The number of evaluations of the individual gradients $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i})$ required to compute an ϵ-optimal solution, i.e., to achieve

$$\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \epsilon. \tag{●}$$

Dynamic Sample Size Methods

Theorem 6

Suppose that SGD is run with stochastic direction estimate (🍉) for some $\tau \in (1, (1-\gamma\beta\mu/2)^{-1}]$ and a stepsize γ satisfying (🍄). In addition, suppose assumptions in Theorem 4 hold. Then, the total number of evaluations of a stochastic gradient of the form $\nabla f(\mathbf{x}_t, \boldsymbol{\xi}_{t,i})$ (**total work complexity**) required to obtain an ϵ-optimal solution (🍊) is $\mathcal{O}(\epsilon^{-1})$.

Proof of Theorem 6.

By definition of the stochastic direction estimate, the bounded variance ($\sharp$) hold with $\zeta = 1/\tau$.

Proof of Theorem 6.

By definition of the stochastic direction estimate, the bounded variance ($\sharp$) hold with $\zeta = 1/\tau$.

Suppose $\omega\rho^{\bar{t}-1} \leq \epsilon$ holds for some $\bar{t} \in \mathbb{N}$, which implies

$$\bar{t} - 1 \geq \frac{\log(\epsilon/\omega)}{\log \rho}.$$

Since $\bar{t} - 1$ is integer, we have $\bar{t} - 1 \geq \left\lceil \frac{\log(\omega/\epsilon)}{-\log \rho} \right\rceil$ and, if the equality holds we still ensure an ϵ-optimal solution in the $\bar{t}$-th iteration.

Proof of Theorem 6.

By definition of the stochastic direction estimate, the bounded variance ($\sharp$) hold with $\zeta = 1/\tau$.

Suppose $\omega\rho^{\bar{t}-1} \leq \epsilon$ holds for some $\bar{t} \in \mathbb{N}$, which implies

$$\bar{t} - 1 \geq \frac{\log(\epsilon/\omega)}{\log \rho}.$$

Since $\bar{t} - 1$ is integer, we have $\bar{t} - 1 \geq \left\lceil \frac{\log(\omega/\epsilon)}{-\log \rho} \right\rceil$ and, if the equality holds we still ensure an ϵ-optimal solution in the $\bar{t}$-th iteration.

Then, by (🍉), the number of sample gradients required in $\bar{t}$-th iteration is $\lceil \tau^{\bar{t}-1} \rceil$ where

$$\tau^{\bar{t}-1} = \tau^{\left\lceil \frac{\log(\omega/\epsilon)}{-\log \rho} \right\rceil} \leq C \exp\left(\log\left(\tau^{\frac{\log(\omega/\epsilon)}{-\log \rho}}\right)\right) = C\left(\frac{\omega}{\epsilon}\right)^{\frac{\log \tau}{-\log \rho}} = C\left(\frac{\omega}{\epsilon}\right)^{\kappa},$$

where $C > 0$ is some constant and $\kappa = \frac{\log \tau}{-\log \rho}$. □

Proof of Theorem 6.

Therefore, the total number of sample gradient evaluations for the first $\bar{t}$ iterations is

$$\sum_{j=1}^{\bar{t}} \lceil \tau^{j-1} \rceil \leq 2 \sum_{j=1}^{\bar{t}} \tau^{j-1} = 2 \left(\frac{\tau^{\bar{t}} - 1}{\tau - 1} \right) \leq 2 \left(\frac{\tau C (\omega/\epsilon)^{\kappa} - 1}{\tau - 1} \right)$$
$$\leq 2C \left(\frac{\omega}{\epsilon} \right)^{\kappa} \left(\frac{1}{1 - 1/\tau} \right).$$

Proof of Theorem 6.

Therefore, the total number of sample gradient evaluations for the first $\bar{t}$ iterations is

$$\sum_{j=1}^{\bar{t}} \lceil \tau^{j-1} \rceil \leq 2 \sum_{j=1}^{\bar{t}} \tau^{j-1} = 2 \left(\frac{\tau^{\bar{t}} - 1}{\tau - 1} \right) \leq 2 \left(\frac{\tau C (\omega/\epsilon)^{\kappa} - 1}{\tau - 1} \right)$$
$$\leq 2C \left(\frac{\omega}{\epsilon} \right)^{\kappa} \left(\frac{1}{1 - 1/\tau} \right).$$

From (🍉) and note that $\tau \leq (1 - \gamma\beta\mu/2)^{-1}$, we have $\rho = \zeta^{-1} = 1/\tau$, so that $\kappa = \log \tau / \log \rho = 1$.

Proof of Theorem 6.

Therefore, the total number of sample gradient evaluations for the first $\bar{t}$ iterations is

$$\sum_{j=1}^{\bar{t}} \lceil \tau^{j-1} \rceil \leq 2 \sum_{j=1}^{\bar{t}} \tau^{j-1} = 2 \left(\frac{\tau^{\bar{t}} - 1}{\tau - 1} \right) \leq 2 \left(\frac{\tau C (\omega/\epsilon)^{\kappa} - 1}{\tau - 1} \right)$$
$$\leq 2C \left(\frac{\omega}{\epsilon} \right)^{\kappa} \left(\frac{1}{1 - 1/\tau} \right).$$

From (🍉) and note that $\tau \leq (1 - \gamma\beta\mu/2)^{-1}$, we have $\rho = \zeta^{-1} = 1/\tau$, so that $\kappa = \log \tau / \log \rho = 1$.

For some $\sigma \in (0, 1]$, we have $\tau = (1 - \sigma\gamma\beta\mu/2)^{-1}$, and it leads to

$$\sum_{j=1}^{\bar{t}} \lceil \tau^{j-1} \rceil \leq 2C \left(\frac{\omega}{\epsilon} \right)^{\kappa} \left(\frac{1}{1 - 1/\tau} \right) = \frac{4C\omega}{\sigma\gamma\beta\mu} \cdot \frac{1}{\epsilon}$$

as desired. □