我们只讨论随机梯度算法（SGD）的收敛性。
为了方便，我们记 $f(x, Y_i) = f_i(x)$. 每次随机选取 $i_k \in \{1, 2, ..., n\}$, SGD 迭代为

$$x^{k+1} = x^k - \alpha_k \nabla f_{i_k}(x^k).$$

假设：我们将在以下假设下分析随机梯度率：

- f 有下界（不一定是凸的）。
- ∇f 是 L-Lipschitz 连续的。
- 对于某个常数 σ^2 和所有 x，有 $E[\|\nabla f_i(x)\|_2] \le \sigma^2$（“方差”有界）。

可以放宽噪声边界到更实际的假设 $E[\|\nabla f_i(x_k) - \nabla f(x_k)\|_2] \le \sigma^2$。这只是在结果中引入了一些额外的项。

- 由' 方差' 有上界可得,

$$\begin{aligned}\mathbb{E}\left[f\left(x^{k+1}\right)\right] &\leq f\left(x^{k}\right)-\alpha_{k}\left\|\nabla f\left(x^{k}\right)\right\|^{2}+\alpha_{k}^{2} \frac{L}{2} \mathbb{E}\left[\left\|\nabla f_{i_{k}}\left(x^{k}\right)\right\|^{2}\right] \\ &\leq f\left(x^{k}\right)-\alpha_{k}\left\|\nabla f\left(x^{k}\right)\right\|^{2}+\alpha_{k}^{2} \frac{L \sigma^{2}}{2}\end{aligned}$$

- 整理可得

$$\alpha_{k}\left\|\nabla f\left(x^{k}\right)\right\|^{2} \leq f\left(x^{k}\right)-\mathbb{E}\left[f\left(x^{k+1}\right)\right]+\alpha_{k}^{2} \frac{L \sigma^{2}}{2}.$$

- 对 $k=1, \ldots t$ 求和得

$$\sum_{k=1}^{t} \alpha_{k-1} \mathbb{E}\left\|\nabla f\left(x^{k-1}\right)\right\|^{2} \leq \sum_{k=1}^{t}\left[\mathbb{E} f\left(x^{k-1}\right)-\mathbb{E} f\left(x^{k}\right)\right]+\sum_{k=1}^{t} \alpha_{k-1}^{2} \frac{L \sigma^{2}}{2}$$

- 继续处理上述不等式:

$$\sum_{k=1}^{t} \alpha_{k-1} \mathbb{E} \underbrace{\left\| \nabla f\left(x^{k-1}\right) \right\|^2}_{\text{bound by min}} \leq \sum_{k=1}^{t} [\underbrace{\mathbb{E} f\left(x^{k-1}\right) - \mathbb{E} f\left(x^{k}\right)}_{\text{telescope}}] + \sum_{k=1}^{t} \alpha_{k-1}^2 \underbrace{\frac{L\sigma^2}{2}}_{\text{no } k}.$$

- 化简得

$$\min_{k=0,1,\ldots,t-1} \left\{ \mathbb{E} \left\| \nabla f\left(x^k\right) \right\|^2 \right\} \sum_{k=0}^{t-1} \alpha_k \leq f\left(x^0\right) - \mathbb{E} f\left(x^t\right) + \frac{L\sigma^2}{2} \sum_{k=0}^{t-1} \alpha_k^2.$$

- 因为 $\mathbb{E} f\left(x^k\right) \geq f^*$ ，两边同时除以 $\sum_k \alpha_{k-1}$ 得

$$\min_{k=0,1,\ldots,t-1} \left\{ \mathbb{E} \left\| \nabla f\left(x^k\right) \right\|^2 \right\} \leq \frac{f\left(x^0\right) - f^*}{\sum_{k=0}^{t-1} \alpha_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} \alpha_k^2}{\sum_{k=0}^{t-1} \alpha_k}$$

- 结果表明:

$$\min_{k=0,1,\ldots,t-1}\left\{\mathbb{E}\left\|\nabla f\left(x^k\right)\right\|^2\right\} \le \frac{f\left(x^0\right)-f^*}{\sum_{k=0}^{t-1}\alpha_k}+\frac{L\sigma^2}{2}\frac{\sum_{k=0}^{t-1}\alpha_k^2}{\sum_{k=0}^{t-1}\alpha_k}.$$

- 若 $\sigma^2=0$, then we could use a constant step-size and would get a $O(1/t)$ rate.
- 由于随机性存在, 收敛速度取决于 $\sum_k \alpha_k^2/\sum_k \alpha_k$.
- 单调下降步长: set $\alpha_k=\alpha/k$ for some α.
  - Gives $\sum_k \alpha_k = O(\log(t))$ and $\sum_k \alpha_k^2 = O(1)$, so error at t is $O(1/\log(t))$.
- 更大的下降步长: set $\alpha_k=\alpha/\sqrt{k}$ for some α.
  - Gives $\sum_k \alpha_k = O(\sqrt{k})$ and $\sum_k \alpha_k^2 = O(\log(k))$, so error at t is $O(\log(t)/\sqrt{t})$.
- \- 常数步长: set $\alpha_k=\alpha$ for some α.
  - Gives $\sum_k \alpha_k = k\alpha$ and $\sum_k \alpha_k^2 = k\alpha^2$, so error at t is $O(1/t)+O(\alpha)$