Expectation on a Single Iteration I

- Recall earlier we derived

$$\begin{aligned} &\sum_{t=0}^{T-1} \alpha_t E[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\ \leq\ & f(\boldsymbol{\theta}_0) - f^* + \sum_{t=0}^{T-1} \frac{\alpha_t^2 G^2 L}{2}, \end{aligned} \quad (1)$$

- Let's get a better representation by taking the expectation on a single iteration
- Assume that we run up to a random number of iteration τ

Expectation on a Single Iteration II

- And τ follows the following probability distribution

$$P(\tau = t) = \frac{\alpha_t}{\sum_{k=0}^{T-1} \alpha_k}$$

- The expected squared-norm of the gradient is

$$\begin{aligned}
& E_{\tau, \tilde{i}_0, \ldots, \tilde{i}_{\tau-1}}[\|\nabla f(\boldsymbol{\theta}_\tau)\|^2] \\
= & \sum_{t=0}^{T-1} E_{\tilde{i}_0, \ldots, \tilde{i}_{t-1}}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] P(\tau = t) \\
= & \left(\sum_{k=0}^{T-1} \alpha_k \right)^{-1} \sum_{t=0}^{T-1} \alpha_t E[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \qquad (2)
\end{aligned}$$

Expectation on a Single Iteration III

- From (2) and (1),

$$
\begin{aligned}
& E[\|\nabla f(\boldsymbol{\theta}_\tau)\|^2] \\
\leq \; & \left(\sum_{k=0}^{T-1} \alpha_k\right)^{-1} \left(f(\boldsymbol{\theta}_0) - f^* + \frac{G^2 L}{2} \sum_{t=0}^{T-1} \alpha_t^2\right)
\end{aligned} \tag{3}
$$

Expectation on a Single Iteration IV

- If the learning rate is a constant $\alpha_t = \alpha$, then

$$
\begin{aligned}
& E[\|\nabla f(\boldsymbol{\theta}_\tau)\|^2] \\
\leq \; & (T\alpha)^{-1}\left(f(\boldsymbol{\theta}_0) - f^* + \frac{G^2 L}{2} T\alpha^2\right) \\
= \; & \frac{f(\boldsymbol{\theta}_0) - f^*}{\alpha T} + \frac{\alpha L G^2}{2}
\end{aligned}
$$

- The right-hand side does not go to zero as $T \to \infty$ due to the term

$$\frac{\alpha L G^2}{2}$$

Reducing the Learning Rate I

- We can make it zero by decreasing the learning rate over time
- To make the right-hand side of (3) go to zero, we need

$$\sum_{t=0}^{T-1} \alpha_t \text{ grows much faster than } \sum_{t=0}^{T-1} \alpha_t^2$$

- An example is

$$\alpha_t = \frac{1}{\sqrt{t+1}}$$

Reducing the Learning Rate II

- We have

$$
\begin{aligned}
\sum_{t=0}^{T-1} \alpha_t = \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \\
\approx \int_0^T \frac{1}{\sqrt{x}} dx \\
= 2x^{1/2}\Big|_0^T = 2\sqrt{T}
\end{aligned}
$$

Reducing the Learning Rate III

and

$$\begin{aligned}\sum_{t=0}^{T-1} \alpha_t^2 &= \sum_{t=0}^{T-1} \frac{1}{t+1} \\ &\approx \int_1^{T+1} \frac{1}{x} dx \\ &= \log x \Big|_1^{T+1} = \log(T+1)\end{aligned}$$

Reducing the Learning Rate IV

- Now (3) becomes

$$
\begin{aligned}
& E[\|\nabla f(\boldsymbol{\theta}_\tau)\|^2] \\
\lessapprox\ & (2\sqrt{T})^{-1}\left(f(\boldsymbol{\theta}_0) - f^* + \frac{G^2 L}{2}(\log(T+1))\right) \\
=\ & \frac{2(f(\boldsymbol{\theta}_0) - f^*) + G^2 L \log(T+1)}{4\sqrt{T}} = O(\frac{\log T}{\sqrt{T}})
\end{aligned}
$$

- The value goes to zero as T increases