Theorem  Suppose \( f : \mathbb{R}^d \to \mathbb{R} \) is \( L \)-smooth and \( g(x, z) \) is an unbiased estimator such that
\[
E[\|g(x, z) - \nabla f(x)\|^2] \leq \sigma^2 \quad \forall x.
\]
Then the iterates of stochastic gradient descent with \( 0 < \alpha_k < 2/L \) satisfy
\[
E \left[ \min_{k=1}^{T} \| \nabla f(x_k) \|^2 \right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2}{2} \sum_{k=0}^{T} \alpha_k^2}{\sum_{k=0}^{T} \alpha_k (1 - \frac{L \alpha_k}{2})} - 1.
\]

Relevant properties of the expectation
\( \triangleright \) Linearity
Given \( X_1, \ldots, X_n \) r.v. and constants \( \lambda_1, \ldots, \lambda_n \), we have

\[
E \left[ \sum_{i=1}^{n} \lambda_i X_i \right] = \sum_{i=1}^{n} \lambda_i E X_i.
\]

**Tower law**

Given two random variables \( X, Y \)

\[
E_X \left[ E[Y \mid X] \right] = E[Y]
\]

**Proof:** By the Taylor Approximation Theorem

\[
f(x_{k+1}) = f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} \| x_{k+1} - x_k \|^2
\]

Conditioning on \( x_k \)

\[
E[f(x_{k+1}) \mid x_k] \leq f(x_k) - \alpha_k E \left[ \nabla f(x_k)^T g_k \mid x_k \right] + \frac{\alpha_k^2}{2} E \left[ \| g_k \|^2 \mid x_k \right]
\]

**Linearity**

\[
\leq f(x_k) - \alpha_k \nabla f(x_k)^T E[g_k \mid x_k] + \frac{\alpha_k^2}{2} E \left[ \| g_k \|^2 \mid x_k \right]
\]
\[(\heartsuit)\quad \leq f(x_k) - \alpha_k \| \nabla f(x_k) \|^2 \]
\[\quad + \frac{L\alpha_k^2}{2} \left[ \sigma^2 + \| \nabla f(x_k) \|^2 \right] \]
\[\quad = f(x_k) - \left( \alpha_k + \frac{L\alpha_k^2}{2} \right) \| \nabla f(x_k) \|^2 \]
\[\quad + \frac{L\alpha_k^2}{2} \sigma^2. \]

By Tower Law
\[E \left[ f(x_{k+1}) \right] \leq E f(x_k) - \left( \alpha_k + \frac{L\alpha_k^2}{2} \right) E \| \nabla f(x_k) \|^2 \]
\[\quad + \frac{L\alpha_k^2}{2} \sigma^2. \]

By recursively applying this formula
\[E \left[ f(x_{T+1}) \right] \leq E f(x_0) - \sum_{k=0}^{T} \left( \alpha_k - \frac{L\alpha_k^2}{2} \right) E \| \nabla f(x_k) \|^2 \]
\[\quad + \sum_{k=0}^{T} \frac{L\alpha_k^2}{2} \sigma^2 \]

The result follows from reordering and using the fact that
\[E \left[ \min_{k \leq T} \| \nabla f(x_k) \|^2 \right] \leq \sum_{k=0}^{T} \left( \alpha_k - \frac{L\alpha_k^2}{2} \right) \]
\[\leq \sum_{k=0}^{T} \left( \alpha_k - \frac{L\alpha_k^2}{2} \right) E \left[ \| \nabla f(x_k) \|^2 \right]. \]
Consequences

If \( \alpha_k = \frac{1}{L\sqrt{T+1}} \Rightarrow 1 - \frac{L\alpha_k}{2} = \frac{1}{2} \).

Thus we derive

\[
E\left[ \min_{k \leq T} \|A f(x_k)\|^2 \right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2}{2L}}{\frac{1}{2} - \sqrt{T+1}} \\
= O\left( \frac{1}{1^T} \right).
\]

By Jensen’s inequality

\[
E \min_{k \leq T} \|A f(x_k)\| = O\left( T^{-\frac{1}{2}} \right).
\]

This is rather slow, however it improves when we have convexity.

Convex guarantees

Theorem Consider the same setting as the previous Theorem, further assume that \( \alpha_k = \alpha \leq \frac{1}{L} \), \( f \) is convex and \( x^* \in \text{argmin} f \). Then

\[
E \left[ \min_{k \leq T} \{ f(x_k) - f(x^*) \} \right] \leq \frac{\|x_0 - x^*\|^2}{2\alpha (k+1)} + \alpha \sigma^2.
\]
In particular if $\alpha = \frac{1}{\sqrt{T+1}}$ and $T \geq L^2$

$$\mathbb{E} \left[ \min_{k \leq T} \{ f(x_k) - f(x^*) \} \right] \leq \frac{1}{2} \| x_0 - x^* \|^2 + \frac{2\sigma^2}{2\sqrt{T+1}}.$$

**Proof** When $\alpha \leq \frac{1}{L}$, (P) gives

$$\mathbb{E} \left[ f(x_{k+1}) \right] \leq f(x_k) - \frac{\alpha}{2} \| \nabla f(x_k) \|^2 + \frac{\alpha \sigma^2}{2}$$

By convexity

$$\leq f(x^*) - \frac{\alpha}{2} \mathbb{E} \left[ \| g(x_k, z) \| \right]^2 + \frac{\alpha \sigma^2}{2}$$

By Assumption

$$\leq f(x^*) - \frac{\alpha}{2} \mathbb{E} \left[ \| g(x_k, z) \| \right]^2 + \frac{\alpha \sigma^2}{2}$$

Using that

$$\| x_{k+1} - x^* \|^2 = \| x_k - x^* \|^2 - \alpha g_k^T (x^* - x_k) + \alpha^2 \| g_k \|^2$$

$$\leq f(x^*) - \mathbb{E} \left[ \frac{\alpha}{2\alpha} \left[ \| x_{k+1} - x^* \|^2 - \| x_k - x^* \|^2 \right] \right] + \frac{\alpha \sigma^2}{2}$$

By Tower law
\[ \mathbb{E} \left[ f(x_{k+1}) - f(x^*) \right] \leq \frac{1}{2\alpha} \mathbb{E} \left[ \| x_{k+1} - x^* \|^2 - \| x_k - x^* \|^2 \right] + \kappa \sigma^2. \]

Once more the result follows by summing up and dividing by \( t \).

\( \square \)

**Remark**

- The rate above is of the order \( O\left( \frac{1}{\sqrt{t}} \right) \), exactly like the rate for nonsmooth convex optimization.

- In HW 4 you'll show the same rate for stochastic nonsmooth convex opt. There, we will have \( g(x, z) \) s.t.

\[ \mathbb{E} [g(x, z)] \leq 2 f(x). \]

**Extensions**

**Acceleration?**

The noise dominates and leads to slow convergence. Best known rate

\[ O \left( \frac{\| x_0 - x^* \|^2}{t^2} + \frac{\sigma^2}{\sqrt{t}} \right). \]
Randomized coordinate descent

Assume our oracle is

\[ i \sim \text{Unif}\{1, \ldots, d\} \]

\[ g(x, i) = d \cdot \frac{\partial f(x)}{\partial x_i} \cdot e_i. \]

The analysis above yields a guarantee but we can do better.

**Theorem.** Assume \( f: \mathbb{R}^d \to \mathbb{R} \) L-smooth. Then SGD with (\( \cdot \)) and \( \alpha_k = \frac{1}{Ld} \) yields

\[
\mathbb{E} \left[ \min_{k \leq T} \| \nabla f(x_k) \|^2 \right] \leq \frac{2L d \left( f(x_0) - \min_{x \in \mathbb{R}^d} f(x) \right)}{T}.
\]

**Proof.** Indeed this oracle gives descent ATM iter \( K \),

\[
f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \| x_{k+1} - x_k \|^2
\]

\[
= f(x_k) - \frac{1}{Ld} d \cdot \frac{\partial f(x_k)}{\partial x_i} \cdot \nabla f(x_k)^T e_i
\]

\[
+ \frac{1}{2Ld^2} \left( d \frac{\partial f(x_k)}{\partial x_i} \right)^2
\]

\[
= f(x) - \frac{1}{2L} \left( \frac{\partial f(x_k)}{\partial x_i} \right)^2.
\]
Taking expectations,

\[ \mathbb{E} [f(x_{k+1})] \leq \mathbb{E} [f(x_k)] - \frac{1}{2L} \mathbb{E} \left[ \left( \frac{\partial f}{\partial x_i} (x_k) \right)^2 \right] \]

\[ = \mathbb{E} [f(x_k)] - \frac{1}{2L} \mathbb{E} \left[ \| \nabla f(x_k) \|^2 \right] \]

\[ = \mathbb{E} \left[ \left( \frac{\partial f(x)}{\partial x_i} (x) \right)^2 \right] \]

\[ = \frac{1}{d} \| \nabla f(x_k) \|^2 \]

By recursively applying the formula above, we obtain

\[ \mathbb{E} [f(x_{T+1})] \leq \mathbb{E} [f(x_0)] - \frac{1}{2Ld} \sum_{k=0}^{T} \mathbb{E} \left[ \| \nabla f(x_k) \|^2 \right] \]

Reordering and multiplying by \( \frac{1}{T} \), yields

\[ \mathbb{E} \left[ \min_{k \leq T} \| \nabla f(x_k) \|^2 \right] \leq \frac{2Ld \left( f(x_0) - \min_f \right)}{T} \]

This is the same rate as in the deterministic case.

Extensions to greedy and cyclic rules can be found in [Nutini, ICML'15] and [Beck, Tetruashvili, SIAM J. Optim. 15'].
Stochastic Variance Reduced Gradient (SVRG)

Recall the finite sum problem

$$\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

The SVRG reads as follows

**Algorithm**

Set $\hat{x}_0 \leftarrow x_0$

for $i = 0, \ldots$

\[
y_i \leftarrow \hat{x}_i
\]

for $j = 0, \ldots, 2d$

\[
draw l \sim \text{Unif} \{1, \ldots, n\}
\]

\[
g_j \leftarrow \nabla f_l(x_i) + \nabla f_l(y_j) - \nabla f_l(\hat{x}_i)
\]

\[
y_{j+1} \leftarrow y_j - \alpha g_j
\]

end for

\[
\hat{x}_{i+1} \leftarrow \frac{1}{2d+1} \sum_{j=0}^{2d} y_j
\]

end for
Theorem: Assume $f: \mathbb{R}^d \to \mathbb{R}$ 1-smooth, $\mu$-strongly convex. Then, if $\alpha$ sufficiently small, $\gamma \in (0, 1)$.

$$E[f(\tilde{x}_k) - \min_f] \leq \gamma^k [f(\tilde{x}_0) - \min_f].$$

Proof: [Johnson, Zhang 2013]