$$\left(x^{(t+1)}-x^*\right)=\left(x^{(t)}-x^*\right)-\eta\left(\nabla f(x^{(t)})-\nabla f(x^*)\right)$$
$$=\left(x^{(t)}-x^*\right)-\eta\nabla^2 f(\zeta^{(t)})\left(x^{(t)}-x^*\right)$$
$$=\left(I-\eta\nabla^2 f(\zeta^{(t)})\right)\left(x^{(t)}-x^*\right).$$

orm inequality $\|Az\|_2 \le \|A\|\|z\|_2$:

$$\|x^{(t+1)}-x^*\|_2 \le \|I-\eta\nabla^2 f(\zeta^{(t)})\|\|x^{(t)}-x^*\|_2$$
$$\le \max\left(|1-\eta\mu|,|1-\eta L|\right)\|x^{(t)}-x^*\|_2.$$

the definition of PSD matrices and the properties of the $\frac{2}{+\mu}$, we have

$$\|x^{(t+1)}-x^*\|_2 \le \left(\frac{L-\mu}{L+\mu}\right)\cdot\|x^{(t)}-x^*\|_2,$$

$$\|x^{(N)}-x^*\|_2 \le \left(\frac{L-\mu}{L+\mu}\right)^N\cdot\|x^{(0)}-x^*\|_2,$$

$$\mathbb{E}x^{(t+1)}=\mathbb{E}x^{(t)}-\alpha_t\mathbb{E}\nabla f(x^{(t)};y_{i_t})$$
$$=\mathbb{E}x^{(t)}-\alpha_t\frac{1}{N}\sum_{i=1}^N\nabla f(x^{(t)};y_i).$$

Strongly convex functions are easier to optimize and allow for faster convergence guarantees.

- $f(x)$ is strongly convex, if there is a $\mu>0$, such that the function $h(x)=f(x)-\frac{\mu}{2}\|x\|^2$ is convex.
- A differentiable function f is strongly convex if there exists a constant $\mu>0$ such that:

$$f(y)\ge f(x)+\nabla f(x)^T(y-x)+\frac{\mu}{2}\|y-x\|^2,\quad \forall x,y\in\mathbb{R}^n.$$

- A function for which the Hessian exists is strongly convex if:

$$\nabla^2 f(x)\succeq \mu I,$$

where $\mu>0$ is a constant. This means that the smallest eigenvalue of the Hessian is at least μ.

Definition 2.1. A *probability space* is the triple $(\Omega,\mathcal{F},\mathbb{P})$. Here,

- Ω is called the *sample space*, which is the set of all possible outcomes of a random experiment.
- $\mathcal{F}$ is called the *event space*, which is a set whose elements $A\in\mathcal{F}$ are subsets of Ω, i.e. $\mathcal{F}\subseteq 2^\Omega$. Further, $\mathcal{F}$ should satisfy
  1. $\emptyset\in\mathcal{F}$, i.e., the empty set is an element of $\mathcal{F}$.
  2. $A\in\mathcal{F}\Rightarrow\Omega\setminus A\in\mathcal{F}$, i.e., $\mathcal{F}$ is closed under complements.
  3. If $A_1,A_2,...\in\mathcal{F}\Rightarrow\cup_i A_i\in\mathcal{F}$, i.e., $\mathcal{F}$ is closed under countable unions.
- $\mathbb{P}:\mathcal{F}\to\mathbb{R}$ is called the probability measure, which should satisfy the axioms of probability:

Definition 2.4 (Random variables)**.** A random variable X is a measurable function from Ω, to (for example) the reals $\mathbb{R}$,

$$X:\Omega\to\mathbb{R}.$$

Proposition 2 (Properties of CDFs)

- $0\le F_X(x)\le 1.$
- $\lim_{x\to\infty}F_X(x)=1.$
- $\lim_{x\to-\infty}F_X(x)=0.$
- $x\le y\Rightarrow F_X(x)\le F_X(y).$
- *$F_X(x)$ is right continuous.*

Definition 2.6 (Probability density function)**.** Suppose we have a real valued random variable X defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$. If the derivative of the CDF $F_X(x)$ exists, then we define the probability density functuon (pdf) $f_X(x)$ of X to be the derivative of $F_X(x)$. Namely,

$$f_X(x)=\frac{dF_X(x)}{dx}.$$

Example 4. Suppose we toss a fair coin. We have the probability space $(\Omega,\mathcal{F},\mathbb{P})$, where $\Omega=\{H,T\}$, $\mathcal{F}=\{\emptyset,\{H\},\{T\},\{H,T\}\}$ and

$$\mathbb{P}(\emptyset)=0,\ \mathbb{P}(\{H\})=\mathbb{P}(\{T\})=1/2,\ \mathbb{P}(\{H,T\})=1.$$

Example 5. Suppose we toss a fair coin twice. We have the probability space $(\Omega,\mathcal{F},\mathbb{P})$, where $\Omega=\{HH,HT,TH,TT\}$, $\mathcal{F}=2^\Omega$. Suppose we want to assign the probability that the first coin lands heads. Then

$$\mathbb{P}(\{HH,HT\})=1/2.$$

Theorem 1.1. *If F is such that*

- $\|\nabla F\|\le L.$
- $\mathbb{E}_{i_t}\|\nabla f_{i_t}(x^{(t)})\|^2\le M+M_G\|\nabla F(x^{(t)})\|^2.$
- *F is strongly convex, with constant c.*
- *$\alpha_t=\frac{\beta}{\gamma+t}$, where $\beta>1/c$, $\gamma>0$ and $\alpha_1<\frac{1}{LM_G}$, then*

$$\mathbb{E}\left[F(x^{(t)})\right]-F(x^*)\le\frac{v}{\gamma+t}\to 0\ \text{as}\ t\to\infty,$$

where

$$v=\max\left\{\frac{\beta^2LM}{2(\beta c-1)},(\gamma+1)(F(x^{(1)})-F(x^*))\right\}.$$

$$F(x)=\frac{1}{N}\sum_{i=1}^N f(x;y_i)$$
$$=:\frac{1}{N}\sum_{i=1}^N f_i(x).$$

$$\mathbb{E}[(X-\mathbb{E}[X])^2]$$
$$=\mathbb{E}[X^2-2X\mathbb{E}[X]+\mathbb{E}[X]^2]$$
$$=\mathbb{E}[X^2]-2\mathbb{E}[X]\mathbb{E}[X]+\mathbb{E}[X]^2$$
$$=\mathbb{E}[X^2]-\mathbb{E}[X]^2$$

$$\mathrm{Var}(aX)=a^2\mathrm{Var}(X).\qquad \mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

Lemma 2.3. *We have*

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right]-F(x^{(t)})$$
$$\le-\alpha_t\nabla F(x^{(t)})^\top\mathbb{E}_{i_t}\left[\nabla f_{i_t}(x^{(t)})\right]+\frac{1}{2}\alpha_t^2L\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right].$$

Proof. Since $x^{(t+1)}=x^{(t)}-\alpha_t\nabla f_{i_t}(x^{(t)})$. Using Taylor expansion, there exists some z such that

$$F(x^{(t+1)})=F(x^{(t)}-\alpha_t\nabla f_{i_t}(x^{(t)}))$$
$$=F(x^{(t)})-\alpha_t\nabla F(x^{(t)})^\top\nabla f_{i_t}(x^{(t)})+\frac{1}{2}\alpha_t^2\nabla f_{i_t}(x^{(t)})^\top\nabla^2F(z)\nabla f_{i_t}(x^{(t)}).$$

Then

$$F(x^{(t+1)})=F(x^{(t)})-\alpha_t\nabla F(x^{(t)})^\top\nabla f_{i_t}(x^{(t)})+\frac{1}{2}\alpha_t^2\nabla f_{i_t}(x^{(t)})^\top\nabla^2F(z)\nabla f_{i_t}(x^{(t)})$$
$$\le F(x^{(t)})-\alpha_t\nabla F(x^{(t)})^\top\nabla f_{i_t}(x^{(t)})+\frac{1}{2}\alpha_t^2L\|\nabla f_{i_t}(x^{(t)})\|^2.$$

This further implies

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right]-F(x^{(t)})\le-\alpha_t\nabla F(x^{(t)})^\top\mathbb{E}_{i_t}\left[\nabla f_{i_t}(x^{(t)})\right]+\frac{1}{2}\alpha_t^2L\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right].$$

Assumption 2.4. We assume

$$\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\right]\|^2\le M+M_G\|\nabla F(x^{(t)})\|^2,$$

where $M_G\ge 0$.

Lemma 2.5. *Under assumptions 2.1, 2.2 and 2.4, we have*

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right]-F(x^{(t)})\le-\left[\left(1-\frac{1}{2}\alpha_tLM_G\right)\alpha_t\right]\|\nabla F(x^{(t)})\|^2+\frac{1}{2}\alpha_t^2LM.$$

Proof. We have

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right]-F(x^{(t)})\le-\alpha_t\|\nabla F(x^{(t)})\|^2+\frac{1}{2}L\alpha_t^2(M+M_G\|\nabla F(x^{(t)})\|^2)$$
$$=-\left[\left(1-\frac{1}{2}\alpha_tLM_G\right)\alpha_t\right]\|\nabla F(x^{(t)})\|^2+\frac{1}{2}\alpha_t^2LM.$$

$$F(x)-F(x^*)\le\frac{\|\nabla F(x)\|^2}{2c}.$$

$$0<\alpha\le\frac{1}{LM_G}\ \textit{and}\ \alpha\le\frac{1}{c},$$

h step-size α and uniform random choice of i_t from the set $\{1,2,...,N\}$ guar

$$\mathbb{E}\left[F(x^{(t)})-F(x^*)\right]\le\frac{\alpha LM}{2c}+(1-\alpha c)^{t-1}\left[F(x^{(1)})-F(x^*)-\frac{\alpha LM}{2c}\right].$$

$$\mathbb{E}\left[F(x^{(t)})-F(x^*)\right]\le\frac{\alpha LM}{2c}+(1-\alpha c)^{t-1}\left[F(x^{(1)})-F(x^*)-\frac{\alpha LM}{2c}\right].$$

3.1 Mini-Batching

To optimize $F(x)=\frac{1}{N}\sum_{i=1}^N f_i(x)$, the idea is to update

$$x^{(t+1)}=x^{(t)}-\alpha_t g(x^{(t)}),$$

where $g(x)=\frac{1}{m}\sum_{k=1}^m\nabla f_{i_k}(x)$. Here i_k is randomly drawn independently, with equal probability from $\{1,...,N\}$.

Here again, as in "vanilla" SGD:

$$\mathbb{E}g(x)=\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^m\nabla f_{i_k}(x)\right]=\frac{1}{m}\sum_{k=1}^m\mathbb{E}\nabla f_{i_k}=\frac{1}{m}\cdot m\cdot\mathbb{E}\nabla f_{i_k}=\frac{1}{N}\sum_{i=1}^N f_i(x)=\nabla F(x).$$

$$\mathbb{E}\left[F(x^{(t+1)})-F(x^*)\right]-\frac{\alpha LM}{2c}\le(1-\alpha c)^t\left[E\left(F(x^{(1)})-F(x^*)\right)-\frac{\alpha LM}{2c}\right].$$

So far, we showed that under assumptions

- $\|\nabla F(x)\|\le L$ for all x.
- $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right]\le M+M_G\|\nabla F(x^{(t)})\|^2.$
- F is strongly convex with constant c.
- $0<\alpha\le\frac{1}{LM_G}$ and $\alpha\le\frac{1}{c}.$