L02
Subgradients and Stochastic Gradient Descent

50.579 Optimization for Machine Learning
Ioannis Panageas
ISTD, SUTD
Definitions

Definition (Subgradients). Let \( f(x) : \mathcal{X} \to \mathbb{R} \) be a function, with \( \mathcal{X} \subset \mathbb{R}^d \). \( g_x \in \mathbb{R}^d \) is called a subgradient of \( f \) at \( x \) if for all \( y \in \mathcal{X} \) we have

\[
  f(y) - f(x) \geq g_x^\top (y - x).
\]

You can define the set of subgradients at \( x \), we denote it by \( \partial f(x) \).
Definitions

Definition (Subgradients). Let $f(x) : \mathcal{X} \rightarrow \mathbb{R}$ be a function. A vector $g_x \in \mathbb{R}^d$ is called a subgradient of $f$ at $x$ if for all $y \in \mathcal{X}$ we have

$$f(y) - f(x) \geq g_x^\top (y - x).$$

You can define the set of subgradients at $x$, we denote it by $\partial f(x)$.

Lemma (Existence and convexity). Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function such that $\partial f(x) \neq \emptyset$ for all $x$. It holds that $f$ is convex.

Proof. It holds that there exists a vector $g$ such that

$$f(ty + (1-t)x) - f(x) \leq g^\top t(y - x),$$

$$f(ty + (1-t)x) - f(y) \leq g^\top (1-t)(x - y).$$
\[
f(ty + (1 - t)x) - f(x) \leq g^\top t(y - x) \quad (1), \\
f(ty + (1 - t)x) - f(y) \leq g^\top (1 - t)(x - y) \quad (2). \\
\]

\[
\begin{align*}
(1 - t) \cdot (1 + t) \cdot (2) \\
\end{align*}
\]

Converse is also true! Application of Supporting Hyperplane Theorem...
\[ f(ty + (1-t)x) - f(x) \leq g^\top t(y - x) \quad (1), \]
\[ f(ty + (1-t)x) - f(y) \leq g^\top (1-t)(x - y) \quad (2). \]
\[ \Rightarrow \]

Converse is also true! Application of Supporting Hyperplane Theorem...

**Lemma (Local minima are global minima).** Let \( f : \mathcal{X} \to \mathbb{R} \) be a convex function. If \( x \) is a local minimum then it is a global minimum. This happens if and only if \( 0 \in \partial f(x) \).

**Proof.** It is a global minimum if and only if \( 0 \in \partial f(x) \).

Moreover, for \( t > 0 \) small enough,  
\[ f(x) \leq f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \]

Hence \( f(x) \leq f(y) \).
**Definitions**

**Definition (Revisited Gradient Descent).** Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex function not necessarily differentiable in some convex set $\mathcal{X}$. GD is defined iteratively:

$$x_{k+1} = x_k - \alpha g_{x_k}.$$ 

**Remarks**

- $g_{x_k} \in \partial f(x_k)$ is the subgradient computed at $x_k$.
- Same guarantees as classic and projected GD.
Analysis of GD for $L$-Lipschitz

**Theorem (Gradient Descent).** Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable, convex (want to minimize) and $L$-Lipschitz. Let $R = \|x_1 - x^*\|_2$, the distance between the initial point $x_0$ and minimizer $x^*$. It holds for $T = \frac{R^2 L^2}{\varepsilon^2}$

$$f \left( \frac{1}{T} \sum_{t=1}^{T} x_t \right) - f(x^*) \leq \varepsilon,$$

with appropriately choosing $\alpha = \frac{\varepsilon}{L^2}$. 
Analysis of GD for \( L \)-Lipschitz

\[ f(x_t) - f(x^*) \leq \nabla x_t(x_t - x^*) \text{ def. subgradient}, \]

\textit{Proof.} It holds that
Analysis of GD for $L$-Lipschitz

*Proof.* It holds that

\[
    f(x_t) - f(x^*) \leq g_x^\top(x_t - x^*) \text{ def. subgradient,}
\]

\[
    = \frac{1}{\alpha} (x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,}
\]
Analysis of GD for $L$-Lipschitz

*Proof.* It holds that

$$f(x_t) - f(x^*) \leq g_{x_t}^\top (x_t - x^*) \text{ def. subgradient},$$

$$= \frac{1}{\alpha} (x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD},$$

$$= \frac{1}{2\alpha} \left( \|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) \text{ law of Cosines},$$
Analysis of GD for $L$-Lipschitz Optimization for Machine Learning

**Proof.** It holds that

$$f(x_t) - f(x^*) \leq g_{x_t}^\top (x_t - x^*) \text{ def. subgradient},$$

$$= \frac{1}{\alpha} (x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD},$$

$$= \frac{1}{2\alpha} \left( \|x_t - x^*\|^2_2 + \|x_t - x_{t+1}\|^2_2 - \|x_{t+1} - x^*\|^2_2 \right) \text{ law of Cosines},$$

$$= \frac{1}{2\alpha} \left( \|x_t - x^*\|^2_2 - \|x_{t+1} - x^*\|^2_2 \right) + \frac{\alpha}{2} \|g_{x_t}\|^2_2 \text{ Def. of GD},$$
Proof. It holds that

\[ f(x_t) - f(x^*) \leq g^\top_{x_t} (x_t - x^*) \text{ def. subgradient,} \]
\[ = \frac{1}{\alpha} (x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \]
\[ = \frac{1}{2\alpha} \left( \|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) \text{ law of Cosines,} \]
\[ = \frac{1}{2\alpha} \left( \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\alpha}{2} \|g_{x_t}\|_2^2 \text{ Def. of GD,} \]
\[ \leq \frac{1}{2\alpha} \left( \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\alpha L^2}{2} \text{ Exercise 3.} \]

Exercise 3 (General case). Suppose \( f(x) \) is \( L \)-Lipschitz continuous and \( \partial f(x) \neq \emptyset \). Then \( \forall x \in \text{dom}(f) \)

\[ \|g_x\|_2 \leq L \text{ where } g_x \in \partial f(x). \]
Analysis of GD for $L$-Lipschitz

Proof cont. Since

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha} \left( \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\alpha L^2}{2},$$

taking the telescopic sum we have

$$\frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2\alpha T} (\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\alpha L^2}{2}.$$ 

$$\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \text{ by choosing appropriately } \alpha, T.$$

The claim follows by convexity since $\frac{1}{T} \sum_{t=1}^{T} f(x_t) \geq f \left( \frac{1}{T} \sum_{t=1}^{T} f(x_t) \right)$ (Jensen’s inequality).
Stochastic Gradient Descent (SGD)

**Definition (Stochastic Gradient Descent).** Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex (want to minimize). The algorithm below is called stochastic gradient descent

$$x_{k+1} = x_k - \alpha_k v_k,$$

where $\mathbb{E}[v_k | x_k] \in \partial f(x_k)$.

**Remarks**

- $\alpha_k$ is called the **stepsize**. Intuitively the smaller, the slower the algorithm.
- $\alpha_k$ must depend on $k$ (vanishing to talk about convergence).
- $v_k$ and moreover $x_k$ are random vectors!
Analysis of SGD for $\mu$-convex

**Theorem (Stochastic Gradient Descent).** Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mu$-strongly convex (want to minimize). Moreover assume that $\mathbb{E}[\|v_k\|^2] \leq \rho^2$. Let $x^*$ be a minimizer. It holds for $\alpha_k = \frac{1}{\mu k}$,

$$
\mathbb{E} \left[ f \left( \frac{1}{T} \sum_t x_t \right) \right] - f(x^*) \leq \frac{\rho^2}{2\mu T} (1 + \log T).
$$

**Remarks**

- $\alpha_k$ scales as $\frac{1}{k}$ and is vanishing to talk about convergence.
- For $T = \Theta \left( \frac{1}{\epsilon} \log \frac{1}{\epsilon} \right)$ we get error $\epsilon$.
- Rakhlin, Shamir & Sridharan (2012) derived a convergence rate in which the $\log T$ is eliminated for a variant.
- Shamir & Zhang (2013) shown theorem above for last iterate $x_T$!
Analysis of SGD for \( \mu \)-convex Optimization for Machine Learning

**Proof of Theorem.** Set \( \nabla^t = \mathbb{E}[v_t | x_t] \).

From strong convexity we get

\[
(x_t - x^*)^\top \nabla^t \geq f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|_2^2.
\]
Analysis of SGD for $\mu$-convex

Proof of Theorem. Set $\nabla^t = \mathbb{E}[v_t|x_t]$.

From strong convexity we get

$$\mathbb{E} \left[ (x_t - x^*)^\top \nabla^t \right] \geq \mathbb{E} \left[ f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|^2 \right].$$

Claim.

$$\mathbb{E}[(x_t - x^*)^\top \nabla^t] \leq \frac{\mathbb{E}[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2]}{2\alpha_t} + \frac{\alpha_t}{2} \rho^2.$$

Proof of Claim. Law of Cosines gives

$$\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \geq 2\alpha_t (x_t - x^*)^\top v_t - a_t^2 \|v_t\|^2.$$

Law of total expectation ... Tower property!
Proof of Cont.

Combining the two above we get (lin. expectation)

$$\mathbb{E} [f(x_t) - f(x^*)] \leq \frac{\mathbb{E} [\|x_t - x^*\|^2_2 (1 - \alpha_t \mu) - \|x_{t+1} - x^*\|^2_2]}{2\alpha_t} + \frac{\alpha_t}{\rho^2}.$$
Analysis of SGD for $\mu$-convex

Proof of Cont.

Combining the two above we get (lin. expectation)

$$
\mathbb{E} [ f(x_t) - f(x^*) ] \leq \frac{\mathbb{E} [ \| x_t - x^* \|^2_2 (1 - \alpha_t \mu) - \| x_{t+1} - x^* \|^2_2 ]}{2\alpha_t} + \frac{\alpha_t}{2} \rho^2.
$$

Therefore (lin. expectation), recall $a_t = \frac{1}{t\mu}$,

$$
\mathbb{E} \left[ \frac{1}{T} \sum_t f(x_t) \right] - f(x^*) \leq \mathbb{E} \left[ -\mu T \| x_T - x^* \|^2_2 \right] + \frac{\rho^2}{2\mu} \frac{1}{T} \sum_t \frac{1}{t}
$$

Optimization for Machine Learning
Analysis of SGD for $\mu$-convex optimization for Machine Learning

Proof of Cont.

Combining the two above we get (lin. expectation)

$$\mathbb{E} \left[ f(x_t) - f(x^*) \right] \leq \frac{\mathbb{E} \left[ \|x_t - x^*\|_2^2 (1 - \alpha_t \mu) - \|x_{t+1} - x^*\|_2^2 \right]}{2\alpha_t} + \frac{\alpha_t}{2} \rho^2.$$ 

Therefore (lin. expectation), recall $a_t = \frac{1}{t\mu}$,

$$\mathbb{E} \left[ \frac{1}{T} \sum_t f(x_t) \right] - f(x^*) \leq \mathbb{E} \left[ -\mu T \|x_T - x^*\|_2^2 \right] + \frac{\rho^2}{2\mu} \frac{1}{T} \sum_t \frac{1}{t} \leq \frac{\rho^2}{2\mu} \left( \frac{1 + \log T}{T} \right).$$
Analysis of SGD (general)

**Theorem (Stochastic Gradient Descent).** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function (want to minimize). Moreover assume that $\|v_k\|_2 \leq \rho$ with probability one. Let $x^*$ be a minimizer. It holds for $\alpha = \frac{R}{\rho \sqrt{k}}$,

$$\mathbb{E} \left[ f \left( \frac{1}{T} \sum_t x_t \right) \right] - f(x^*) \leq \frac{R \rho}{\sqrt{T}}.$$ 

**Remarks**

- $\alpha$ scales as $\sqrt{\frac{1}{k}}$ and is vanishing to talk about convergence but fixed!
- For $T = \Theta \left( \frac{1}{\epsilon^2} \right)$ we get error $\epsilon$. 

Optimization for Machine Learning
Analysis of SGD (general)

Proof. (Recall and add expectation)

\[ E_{1:T} [f(x_t) - f(x^*)] \leq E_{1:T} [(x_t - x^*)^\top \nabla^t] \]

\[ = E_{1:t-1} [E_{1:T} [(x_t - x^*)^\top \nabla^t | \nu_1, ..., \nu_{t-1}]] \]
Proof. (Recall and add expectation)

$$\mathbb{E}_{1:T} [f(x_t) - f(x^*)] \leq \mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t]$$

$$= \mathbb{E}_{1:t-1} [\mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t | \nu_1, ..., \nu_{t-1}]]$$

$$= \mathbb{E}_{1:T} [(x_t - x^*)^\top \mathbb{E}_{1:t-1} [\nabla^t | \nu_1, ..., \nu_{t-1}]]$$

$$= \mathbb{E}_{1:T} [(x_t - x^*)^\top \nu_t]$$
Analysis of SGD (general)

**Proof.** (Recall and add expectation)

\[
\mathbb{E}_{1:T} [f(x_t) - f(x^*)] \leq \mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t]
\]

\[
= \mathbb{E}_{1:t-1} \mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t | \nu_1, ..., \nu_{t-1}]
\]

\[
= \mathbb{E}_{1:T} [(x_t - x^*)]^\top \mathbb{E}_{1:t-1} [\nabla^t | \nu_1, ..., \nu_{t-1}]
\]

\[
= \mathbb{E}_{1:T} [(x_t - x^*)]^\top \nu_t \quad \text{Recall } ||\nu_t|| \leq \rho!
\]

\[
\leq \mathbb{E}_{1:T} \left[ \frac{1}{2\alpha} \left( ||x_t - x^*||_2^2 - ||x_{t+1} - x^*||_2^2 \right) \right] + \frac{\alpha \rho^2}{2}.
\]
Analysis of SGD (general)

Proof. (Recall and add expectation)

\[ \mathbb{E}_{1:T} [f(x_t) - f(x^*)] \leq \mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t] \]

\[ = \mathbb{E}_{1:t-1} [\mathbb{E}_{1:T} [(x_t - x^*)^\top \nabla^t | \nu_1, \ldots, \nu_{t-1}]] \]

\[ = \mathbb{E}_{1:T} [(x_t - x^*)]^\top \mathbb{E}_{1:t-1} [\nabla^t | \nu_1, \ldots, \nu_{t-1}] \]

\[ = \mathbb{E}_{1:T} [(x_t - x^*)]^\top \nu_t \]

\[ \leq \mathbb{E}_{1:T} \left[ \frac{1}{2\alpha} \left( \|x_t - x^*\|^2_2 - \|x_{t+1} - x^*\|^2_2 \right) \right] + \frac{\alpha \rho^2}{2}. \]

Taking the telescopic sum we have

\[ \mathbb{E}_{1:T} \left[ \frac{1}{T} \sum_{t=1}^{T} f(x_t) - f(x^*) \right] \leq \frac{R^2}{2\alpha T} + \frac{\alpha \rho^2}{2}. \]
Example: Coordinate Descent

**Definition (Coordinate Descent).** Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex differentiable function in some convex set $\mathcal{X}$. CD is defined iteratively:

$$\text{Choose coordinate } i \in [d] \text{ and update } x_{k+1} = x_k - \alpha_k \frac{\partial f(x_k)}{\partial x_i} \cdot e_i.$$

**Remarks**

- Similar guarantees with GD as long as each coordinate is taken often.
- If coordinate $i$ is chosen uniformly at random, then instantiation of ?.
Conclusion

• Introduction to Subgradients and SGD.
  – Same guarantees as for differentiable functions.
  – SGD has rate of convergence $O \left( \frac{1}{\epsilon} \ln \frac{1}{\epsilon} \right)$ for $\mu$-convex.
  – Next Lecture we will see examples related to MLE.

• Next week we will talk about online learning/optimization!