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 $f(x) : \mathcal{X} \to \mathbb{R}$ be a fun*
$g_x \in \mathbb{R}^d$ *is called a subgradient of f at x if for all $y \in \mathcal{X}$ we h*

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

Example: $|x|$

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

Lemma (Existence and convexity). *Let $f : \mathcal{X} \to \mathbb{R}$ 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

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

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

$$\left.\begin{aligned} f(ty+(1-t)x)-f(x) &\le g^\top t(y-x) \quad (1),\\ f(ty+(1-t)x)-f(y) &\le g^\top(1-t)(x-y) \quad (2). \end{aligned}\right\} \xRightarrow{(1-t)\cdot(1)+t\cdot(2)}$$

$$\boxed{f(ty+(1-t)x) \le (1-t)f(x)+tf(y).}$$

Converse is also true! Application of Supporting Hyperplane Theorem…

$$\left.\begin{aligned} f(ty+(1-t)x)-f(x) &\le g^\top t(y-x) \quad (1), \\ f(ty+(1-t)x)-f(y) &\le g^\top (1-t)(x-y) \quad (2). \end{aligned}\right\} \overset{(1-t)\cdot(1)+t\cdot(2)}{\Longrightarrow}$$

$$\boxed{f(ty+(1-t)x) \le (1-t)f(x)+tf(y).}$$

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 $\mathbf{0} \in \partial f(x)$.*

Proof. It is a global minimum if and only if $\mathbf{0} \in \partial f(x)$.

Moreover, for $t > 0$ small enough,

$$\boxed{\text{Hence } f(x) \le f(y).}$$

$$f(x) \le f(tx+(1-t)y) \le tf(x)+(1-t)f(y).$$

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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 $T = \frac{R^2L^2}{\epsilon^2}$*

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

with appropriately choosing $\alpha = \frac{\epsilon}{L^2}$.

Analysis of GD for L-Lipschitz

Proof. It holds that

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

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq g_{x_t}^\top (x_t - x^*) \text{ def. subgradient,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq g_{x_t}^\top (x_t - x^*) \text{ def. subgradient,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \text{ law of Cosines,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq g_{x_t}^\top (x_t - x^*) \text{ def. subgradient,} \\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) + \frac{\alpha}{2}\|g_{x_t}\|_2^2 \text{ Def. of GD,}
\end{aligned}
$$

Analysis of GD for L-Lipschitz

Proof. It holds that

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

Exercise 3 (General case). *Suppose $f(x)$ is L-Lipschitz continous and $\partial f(x) \neq \emptyset$. Then $\forall x \in dom(f)$*

$$\|g_x\|_2 \leq L \textit{ where } g_x \in \partial f(x).$$

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

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) &\leq \frac{1}{2\alpha T}(\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\alpha L^2}{2}. \\ &\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \text{ by choosing appropriately } \alpha, T.\end{aligned}$$

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 $f : \mathbb{R}^d \to \mathbb{R}$ 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

- α_k is called the stepsize. Intuitively the smaller, the slower the algorithm.
- α_k must depend on k (vanishing to talk about convergence).
- v_k and moreover x_k are random vectors!

Analysis of SGD for μ-convex

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

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

Remarks

- α_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 ϵ.
- 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 μ-convex

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 μ-convex

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

From strong convexity we get

$$\mathbb{E}\left[(x_t - x^*)^\top \nabla^t\right] \geq \mathbb{E}\left[f(x_t) - f(x^*) + \frac{\mu}{2}\|x_t - x^*\|_2^2\right].$$

Claim.

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

Proof of Claim. Law of Cosines gives

$$\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \geq 2\alpha_t(x_t - x^*)^\top v_t - a_t^2 \|v_t\|_2^2$$

Law of total expectation … Tower property!

Analysis of SGD for μ-convex

Proof of Cont.

Combining the two above we get (lin. expectation)

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

Analysis of SGD for μ-convex

Proof of Cont.

Combining the two above we get (lin. expectation)

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

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

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

Analysis of SGD for μ-convex

Proof of Cont.

Combining the two above we get (lin. expectation)

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

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

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

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

- a 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 ϵ.

Analysis of SGD (general)

Proof. (Recall and add expectation)

$$
\begin{aligned}
\mathbb{E}_{1:T}\left[f(x_t) - f(x^*)\right] &\le \mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t] \\
&= \mathbb{E}_{1:t-1}[\mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t | v_1, \ldots, v_{t-1}]]
\end{aligned}
$$

Analysis of SGD (general)

Proof. (Recall and add expectation)

$$\begin{aligned}
\mathbb{E}_{1:T}\left[f(x_t) - f(x^*)\right] &\leq \mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t] \\
&= \mathbb{E}_{1:t-1}[\mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t | v_1, ..., v_{t-1}]] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top \mathbb{E}_{1:t-1}[\nabla^t | v_1, ..., v_{t-1}] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top v_t
\end{aligned}$$

Analysis of SGD (general)

Proof. (Recall and add expectation)

$$
\begin{aligned}
\mathbb{E}_{1:T}\left[f(x_t) - f(x^*)\right] &\le \mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t] \\
&= \mathbb{E}_{1:t-1}[\mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t | v_1, \ldots, v_{t-1}]] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top \mathbb{E}_{1:t-1}[\nabla^t | v_1, \ldots, v_{t-1}] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top v_t \\
&\le \mathbb{E}_{1:T}\left[\frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)\right] + \frac{\alpha\rho^2}{2}.
\end{aligned}
$$

Recall $||v_t|| \le \rho$!

Analysis of SGD (general)

Proof. (Recall and add expectation)

$$
\begin{aligned}
\mathbb{E}_{1:T}\left[f(x_t) - f(x^*)\right] &\le \mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t] \\
&= \mathbb{E}_{1:t-1}[\mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t | v_1, \ldots, v_{t-1}]] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top \mathbb{E}_{1:t-1}[\nabla^t | v_1, \ldots, v_{t-1}] \\
&= \mathbb{E}_{1:T}[(x_t - x^*)]^\top v_t \\
&\le \mathbb{E}_{1:T}\left[\frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)\right] + \frac{\alpha\rho^2}{2}.
\end{aligned}
$$

Recall $||v_t|| \le \rho$!

Taking the telescopic sum we have

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

Example: Coordinate Descent

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

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

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 μ-convex.
  - Next Lecture we will see examples related to MLE.
- Next week we will talk about online learning/optimization!