IFT 6132 – Advanced Structured Prediction **Winter 2019**

Lecture 9 — February 5th, Université de Montréal

Lecturer: Simon Lacoste-Julien *Scribe: Sarah Benamara, Elyes Lamouchi*

9.1 Convex Optimization:

Some definitions for context: [S.Boyd and L.Vandenberghe, 2004], [Zhou, 2018]

Definition. a set C is convex if $\forall x, y \in C, \exists \theta \in [0,1]$ such that $\theta x + (1-\theta)y \in C$.

Definition. a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\begin{cases} i) \text{ the domain of f is a convex set} \\ ii) \ \forall x, y \in dom(f), \theta \in [0,1], f(\theta x + (1-\theta)y)) \\ \leq \theta f(x) + (1-\theta) f(y) \end{cases}$

Definition. f is said to be μ-strongly convex if $\exists \mu > 0$ such that $f(x) + \mu ||x||^2$ is convex.

Let f be a μ-strongly convex, and $g(x) = f(x) + \mu ||x||^2$.

Definition. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and let $x \in dom(f)$. An element $u \in \mathbb{R}^n$ is called a **subgradient** of f at x if:

$$\langle u, y - x \rangle \leq f(y) - f(x), \forall y \in dom(f)$$

We call the collection of all subgradients of $f(x)$ at x the **subdifferential** of f at x, and we denote it by $\partial f(x)$.

Definition. The directional derivative of g is $g\prime(x, v) \triangleq \lim\limits_{n \to \infty} \frac{g(x+\alpha v) - g(x)}{\alpha}$

By convexity of g we have, $\forall \alpha \in [0,1]$:

$$g(\alpha y + (1-\alpha)x) \leq \alpha g(y) + (1-\alpha) g(x) \iff \frac{g(\alpha y + (1-\alpha)x) - g(x)}{\alpha} \leq g(y) - g(x)$$
$$\iff * \lim_{n \to \infty} \frac{g(\alpha y + (1-\alpha)x) - g(x)}{\alpha} \leq g(y) - g(x) \iff g\prime(x, y - x) \leq g(y) - g(x)$$

Note. If g is convex in $y - x$ and g is finite in x then the limit in * exists.

Now the trick is to link the product of the subgradient and $y - x$ with the directional derivative.

Proposition. $g(x + t(y-x)) = \sup\limits_{v \in \partial g(x)} \langle v, y - x \rangle$

Proof. $\forall v \in \partial g(x), z \in dom(g) \underset{\text{by definition}}{\longrightarrow} \langle v, z - x \rangle \leq g(z) - g(x)$.

$\forall t \in [0,1], \forall x, y \in dom(g)$ *such that* $z = ty + (1-t)x \underset{\text{by convexity of } g}{\longrightarrow} z \in dom(g)$.
$\iff \langle v, z - x\rangle = \langle v, (x + t(y-x)) - x\rangle \le g(x + t(y-x)) - g(x)$
$\iff t\langle v, y - x\rangle \le g(x + t(y-x)) - g(x) \iff \langle v, y - x\rangle \le \frac{g(x+t(y-x))-g(x)}{t} = g\prime(x, y-x)$.

Thus we have the inequatily,

$$\langle v, y - x\rangle \le g(y) - g(x)$$

Proposition. If f, h are convex functions on $\mathbb{R}^n$ and $\lambda, \mu > 0$ then the subgradient of the function $g(x) = \lambda f(x) + \mu h(x)$ satisfies $\partial g(x) = \lambda \partial f(x) + \mu \partial h(x)$.

Moreover f and $h(x) = ||x||^2$ are convex functions, thus the subgradient of the function $g(x) = f(x) + \mu h(x)$ satisfies $\partial g(x) = \partial f(x) + \mu \partial h(x)$.
Let u, w be such that $u \in \partial f(x)$ and $w \in \partial h(x)$,

$$\langle v, y - x\rangle = \langle u + \mu w, y - x\rangle$$

Therefore, by the first order convexity condition on g i.e $\langle v, y - x\rangle \le g(y) - g(x)$, we have

$$f(y) \ge f(x) + \langle u, y - x\rangle + \frac{\mu}{2}||y - x||^2, \forall u \in \partial f(x)$$

$\implies$ **We have thus proven that saying that $f(x)$ is μ-strongly convex is equivalent to saying that,** $\forall x, y \in dom(f), f(y) \ge f(x) + \langle u, y - x\rangle + \frac{\mu}{2}||y - x||^2, \forall u \in \partial f(x)$.

9.2 Fundamental Descent Lemma:

Definition. A function f has Lipschitz continuous gradient if: $\forall x, y \in dom(f), \exists L > 0$ such that $||\nabla f(x) - \nabla f(y)||^2 \le L||x - y||^2$.
We say that a function is L-smooth if its gradient is Lipschitz continuous for some $L > 0$.

Theorem. f is L-smooth $\implies f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}||y - x||^2$.
Proof. [P.Bertsekas,]: Let g be such that:

$$g(\alpha) \triangleq f(x + \alpha(y-x)) \longrightarrow f(y) - f(x) = g(1) - g(0) = \int_0^1 \frac{dg}{d\alpha}(\alpha)d\alpha = \int_0^1 \frac{df}{d\alpha}(x + \alpha(y-x))d\alpha =$$

$$\int_0^1 (y-x)\nabla f(x + \alpha(y-x))d\alpha = \int_0^1 (y-x)\Big[\nabla f(x + \alpha(y-x)) - \nabla f(x)\Big]d\alpha + \int_0^1 (y-x)\nabla f(x)d\alpha$$

$$\le \Big|\Big|\int_0^1 (y-x)\nabla f(x + \alpha(y-x))d\alpha\Big|\Big| + \int_0^1 (y-x)\nabla f(x)d\alpha$$

$$\underset{\text{by the triangle inequality}}{\leq} \int_0^1 \Big|\Big|(y-x)\nabla f(x+\alpha(y-x))\Big|\Big| d\alpha + \int_0^1 (y-x)\nabla f(x) d\alpha$$
$$\underset{\text{by Cauchy-Schwartz}}{\leq} \int_0^1 ||y-x||.||\nabla f(x+\alpha(y-x))||d\alpha + (y-x)\Delta f(x)$$
$$\underset{\text{by Lipschitz continuity of } f}{\leq} ||y-x|| \int_0^1 L\alpha||y-x||d\alpha + (y-x)\Delta f(x)$$
$$= \langle \nabla f(x), y-x\rangle + L||y-x||^2$$

Usefulness of this lemma:

Proposition. Let f be an $L-$Lipschitz function, $\forall x \in dom(f), \exists \gamma$ such that $x - \gamma\nabla f(x) \in dom(f) \Longrightarrow f(x-\gamma\nabla f(x)) \leq f(x) - \big[\gamma(1-\frac{\gamma L}{2})\big].||\nabla f(x)||^2$.

Proof. $f(x-\gamma\nabla f(x)) \leq f(x) - \gamma\langle\nabla f(x), \nabla f(x)\rangle + \frac{L}{2}||x-\gamma\nabla f(x) - x||^2 = f(x) - \big[\gamma(1-\frac{\gamma L}{2})\big].||\nabla f(x)||^2$

Gradient Descent is an optimization method, in wich we minimize an objective function f by following the direction of steepest descent (i.e the gradient) in each iteration : $x_{t+1} \leftarrow x_t - \gamma\nabla f(x_t)$. This is where the above proposition comes in handy:

$$f(x-\gamma\nabla f(x)) - f(x) \leq \big[\gamma(\frac{\gamma L}{2}-1)\big].||\nabla f(x)||^2$$
$$\Longleftrightarrow \min_\gamma\big[f(x-\gamma\nabla f(x)) - f(x)\big] \leq \min_\gamma\big[\gamma(\frac{\gamma L}{2}-1)\big].||\nabla f(x)||^2$$
$$\Longleftrightarrow f(y_{\gamma*}) \leq f(x) - \frac{1}{2L}||\nabla f(x)||^2, \text{where} y_{\gamma*} = y_{\frac{1}{L}} = x - \frac{1}{L}\nabla f(x).$$

9.3 Convergence rates of gradient descent:

The L-smooth, μ-convex case: [Liu, 2015]

As shown above, we take the step to be $\gamma = \frac{1}{L}$ so that, at the k^{th} iteration we have: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \Longleftrightarrow -\frac{1}{L}\nabla f(x_k) = x_{k+1} - x_k$.

By L-smoothness of f, we have:

$$f(x_{k+1}) \leq f(x_k) + \langle\nabla f(x_k), x_{k+1}-x_k\rangle + \frac{L}{2}||x_{k+1}-x_k||^2 = f(x_k) - \frac{1}{L}\nabla f(x_k) + \frac{L}{2}||-\frac{1}{L}\nabla f(x_k)||^2$$
$$= f(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2$$

And let $\Delta_k = f(x_k) - f^*$, where $f^* = \underset{x}{min} f(x)$. By μ-strong convexity we have:

$$f(x) - f(x_k) \geq \langle \nabla f(x_k), x - x_k \rangle + \frac{\mu}{2}||x - x_k||^2 \iff \underset{x}{min} f(x) - f(x_k) \geq \langle \nabla f(x_k), x^* - x_k \rangle + \frac{\mu}{2}||x^* - x_k||^2$$
$$\iff \Delta_{k+1} \underset{f(x_{k+1}(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2}{\leq} \Delta_k - \frac{1}{2L}||\nabla f(x_k)||^2 \underset{||\nabla f(x_k)||^2 \leq 2\mu\Delta_k}{\leq} \left[1 - \frac{\mu}{L}\right]\Delta_k$$

Thus we have,

$$\underset{k \to \infty}{lim} \frac{\Delta_{k+1}}{\Delta_k} = 1 - \frac{\mu}{L}$$

Therefore, taking $\mu \leq L$ we have $1 - \frac{\mu}{L} \in (0, 1)$. A convergence rate of this sort is called ***a linear rate***.

The weakly-convex case:

From convexity it follows,

$$f(x^*) - f(x_k) \geq \langle \nabla f(x_k), x^* - x_k \rangle \iff \Delta_k \leq \langle \nabla f(x_k), x_k - x^* \rangle \underset{\text{By Cauchy-Schwartz}}{\leq} ||x_k - x^*||.||\nabla f(x_k)||$$

Moreover,

$$||x_k - x^*||^2 = ||x_k - \frac{1}{L}\nabla f(x_k) - x^*||^2 = ||x_k - x^*||^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^* \rangle \leq ||x_k - \frac{1}{L}\nabla f(x_k) - x^*||^2$$
$$= ||x_k - x^*||^2 + \frac{1}{L^2}||\nabla f(x_k)||^2 - \frac{2}{L^2}||\nabla f(x_k)||^2$$

By recursion, we have:

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2L\underbrace{||x_0 - x^*||^2}_{r_0^2}}\Delta_k^2 \iff \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{1}{2L.r_0^2} \geq ... \geq \frac{1}{\Delta_0} + \frac{1}{2L.r_0^2}(k+1)$$

Thus we have $\underset{k \to \infty}{lim} \frac{\Delta_{k+1}}{\Delta_k} = 1$. And a rate of convergence of $O(\frac{k}{Lr_0^2})$. Such a convergence is said to have a ***linear rate***.

9.4 Second order methods:

Newton's method: [S.Boyd and L.Vandenberghe, 2004]

By gradient descent, we can also minimize an objective function $f(x)$ by using its second derivative i.e its Hessian matrix $\nabla^2 f(x)$. **The Newton step:** For $x \in dom(f)$, the vector $\Delta x = -\nabla^2 f(x)^{-1}\nabla f(x)$ is called **the Newton step**.

Proposition. The Hessian matrix is semi definite.
The implication of the above proposition is that,

$$\nabla f(x)^T \Delta x = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \underset{\text{positive definiteness}}{<} 0$$

Thus going from one iteration to the next, we are always in descent direction.

Bibliography

[Liu, 2015] Liu, J. (2015). A simple proof of chernoff's bound, 1998. In *Class notes*, pages 23:1,23:3.

[P.Bertsekas,] P.Bertsekas, D. Convergence analysis of gradient methods. In *Class notes*, page 8.

[S.Boyd and L.Vandenberghe, 2004] S.Boyd and L.Vandenberghe (2004). Newton's method. In *Convex Optimization*, pages 67,484. Cambridge University Press.

[Zhou, 2018] Zhou, X. (2018). On the fenchel duality between strong convexity and lipschitz continuous gradient. pages 2,3.