1 Assumptions

We are interested in composite minimization problems where the objective function is the sum of two functions $F(x) := g(x) + f(x)$.

- f is smooth (differentiable)
- $\nabla f(x)$ is Lipschitz continuous with constant $L > 0$.
- g convex
- dom $g \cap$ dom $f \neq \varnothing$.

Ideally, we need additional assumptions on f and g (as well as other functions stated in this lecture notes), i.e., they should be proper and closed and satisfy one of the assumptions of the Weierstrass' theorem (see lecture notes for Lecture 12). These assumptions guarantee that our problem has a non-empty set of minimizers that can be attained. However, we will not focus on these assumptions since they require more advanced knowledge of convex analysis.

1.1 Optimality conditions

Lemma 1 (Fermat's optimality condition)**.** *Let h be convex. Then x^* is a stationary point of h if and only if $0 \in \partial h(x^*)$.*

Proof. Check previous lecture notes. □

Lemma 2. *Let g be a proper, closed and convex function with bounded domain. Then for any $x, u \in$ dom g the following three claims are equivalent*

- $u = prox_g(x)$
- $x - u \in \partial g(u)$
- $(x-u)^T(y-u) \leq g(y) - g(u)\ \forall y \in$ *dom* g.

Proof. Check previous lecture notes. □

Let's introduce new notation that is standard in the optimization literature. Let us introduce the **gradient mapping**

$$G(x) := \frac{1}{\alpha}(x - x^+) = \frac{1}{\alpha}(x - \text{prox}_{\alpha g}(x - \alpha \nabla f(x))). \tag{1}$$

Furthermore, using the definition of the gradient mapping we can rewrite the definition of proximal gradient as

$$x^+ = x - \alpha G(x).$$

However, note that $G(x)$ is not a sub-gradient for our problem. The gradient mapping is a very important quantity because we can relate it to the optimality conditions of the composite problem.

Theorem 1. *Let's assume that the Assumptions in Subsection 1 hold. For $x^* \in$ dom $g \cap$ dom f it holds that $G(x^*) = 0$ if and only if x^* is a stationary point of F ($g + f$).*

Proof. Check previous lecture note. □

2 Descent lemma and termination of Armijo line-search for proximal gradient

For gradient descent we proved that the objective function decreases monotonically. To prove that we used the fundamental theorem of calculus and the definition of the gradient. This gave us

$$F(x_{k+1}) \leq F(x_k) - \frac{\alpha_k}{2}\|\nabla F(x_k))\|_2^2. \tag{2}$$

For the composite problem, function F might not be differentiable because g might not be differentiable. Therefore, the above inequality (sufficient decrease inequality) does not apply. However, we can still follow similar reasoning for proximal gradient.

Lemma 3 (Sufficient decrease lemma)**.** *Let $\alpha_k < 2/L$, then*

$$F(x_{k+1}) \leq F(x_k) + \left(\frac{L}{2} - \frac{1}{\alpha_k}\right)\|x_{k+1} - x_k\|_2^2.$$

If we set $\alpha_k = 1/L$ we get

$$F(x_{k+1}) \leq F(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

Using the definition of the gradient mapping we get

$$F(x_{k+1}) \leq F(x_k) - \frac{1}{2L}\|G(x_k)\|_2^2.$$

Proof. Using the FToC for f (we can do this because we assume that $\nabla f(x)$ is Lipschitz continuous), we

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2. \tag{3}$$

Using the definition of proximal gradient $x_{k+1} = \text{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$ and Lemma 2 for function $\alpha_k g$, $y := x_k$ and $x := x_k - \alpha_k \nabla f(x_k)$ we get

$$(x_k - \alpha_k \nabla f(x_k) - x_{k+1})^T(x_k - x_{k+1}) \leq \alpha_k g(x_k) - \alpha_k g(x_{k+1}),$$

which is equivalent to

$$-\alpha_k \nabla f(x_k)^T(x_k - x_{k+1}) \leq \alpha_k g(x_k) - \alpha_k g(x_{k+1}) - \|x_{k+1} - x_k\|_2^2,$$

by dividing by α_k (and assuming that $\alpha_k > 0$) we get

$$-\nabla f(x_k)^T(x_k - x_{k+1}) \leq g(x_k) - g(x_{k+1}) - \frac{1}{\alpha_k}\|x_{k+1} - x_k\|_2^2.$$

By re-arranging we get

$$g(x_{k+1}) \leq g(x_k) - \nabla f(x_k)^T(x_{k+1} - x_k) - \frac{1}{\alpha_k}\|x_{k+1} - x_k\|_2^2 \tag{4}$$

Setting $y := x_{k+1}$ and $x := x_k$ in equation (3) and adding it to (4) we get

$$F(x_{k+1}) \leq F(x_k) + \left(\frac{L}{2} - \frac{1}{\alpha_k}\right)\|x_{k+1} - x_k\|_2^2$$

This proves the first part of the lemma. Setting $\alpha_k = 1/L$ and using the definition of $G(x)$ in (1) we get the other two claims. □

Lemma 4 (Armijo line-search terminates)**.** *Any $\alpha \leq \frac{2(1-\theta)}{L}$ satisfies the termination criterion of Armijo line-search for proximal gradient.*

Proof. The main idea of this proof is to show that there exists a positive α such that

$$F(x(\alpha)) - F(x_k) - \theta(\ell(x(\alpha)) - \ell(x_k)) \leq 0.$$

We will do this by attempting to upper bound $F(x(\alpha)) - F(x_k)$ and $\ell(x(\alpha)) - \ell(x_k)$ separately. To upper bound $F(x(\alpha)) - F(x_k)$ we will use the FToC for function f. To upper bound $\ell(x(\alpha)) - \ell(x_k)$ we should use convexity of function ℓ.

Using the FToC for f (we can do this because we assume that $\nabla f(x)$ is Lipschitz continuous), we

$$f(x(\alpha)) \leq f(x_k) + \nabla f(x_k)^T(x(\alpha) - x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2.$$

Thus

$$g(x(\alpha)) + f(x(\alpha)) \leq g(x(\alpha)) + f(x_k) + \nabla f(x_k)^T(x(\alpha) - x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2.$$

Add and subtract $g(x_k)$ in the right hand side to get

$$g(x(\alpha)) + f(x(\alpha)) \leq g(x(\alpha)) - g(x_k) + g(x_k) + f(x_k) + \nabla f(x_k)^T(x(\alpha) - x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2.$$

Using the definition that $F(x) = g(x) + f(x)$ we get

$$F(x(\alpha)) \leq F(x_k) + g(x(\alpha)) - g(x_k) + \nabla f(x_k)^T(x(\alpha) - x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2. \tag{5}$$

Let $\ell(x) := g(x) + f(x_k) + \nabla f(x_k)^T(x - x_k)$. Then (5) is equivalent to

$$F(x(\alpha)) \leq F(x_k) + \ell(x(\alpha)) - \ell(x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2. \tag{6}$$

This complete the upper bound for $F(x(\alpha)) - F(x_k)$.

Regarding the upper bound $\ell(x(\alpha)) - \ell(x_k)$, note that $\ell(x)$ is a convex as a function of x because g is convex. Also, define $\bar{x} := x_k - G(x_k)$, then I have that $\alpha\bar{x} + (1-\alpha)x_k = x(\alpha)$. Therefore, using convexity of ℓ and the previous definition we get

$$\ell(x(\alpha)) = \ell(\alpha\bar{x} + (1-\alpha)x_k) \leq \alpha\ell(\bar{x}) + (1-\alpha)\ell(x_k). \tag{7}$$

Now let's put (6) and (7) together:

$$\begin{aligned} F(x(\alpha)) - F(x_k) - \theta(\ell(x(\alpha)) - \ell(x_k)) &\leq \theta(\ell(x(\alpha)) - \ell(x_k)) + \ell(x(\alpha)) - \ell(x_k) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2 \\ &\leq (1-\theta)(\ell(x(\alpha)) - \ell(x_k)) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2 \\ &\leq \alpha(1-\theta)(\ell(\bar{x}) - \ell(x_k)) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2. \end{aligned} \tag{8}$$

Now we have to upper bound $\ell(\bar{x}) - \ell(x_k)$ in order to show that there exists α such that

$$\alpha(1-\theta)(\ell(\bar{x}) - \ell(x_k)) + \frac{L}{2}\|x(\alpha) - x_k\|_2^2 \leq 0.$$

Note that $\bar{x} = \text{prox}_g(x_k - \nabla f(x_k))$ and Lemma 2 for function g, $y := x_k$ and $x := x_k - \nabla f(x_k)$ we get

$$(x_k - \nabla f(x_k) - \bar{x})^T (x_k - \bar{x}) \leq g(x_k) - g(\bar{x}),$$

which is equivalent to

$$-\nabla f(x_k)^T (x_k - \bar{x}) \leq g(x_k) - g(\bar{x}) - \|\bar{x} - x_k\|_2^2,$$

Re-arrange to get

$$g(\bar{x}) - g(x_k) + \nabla f(x_k)^T (\bar{x} - x_k) \leq -\|\bar{x} - x_k\|_2^2,$$

which is equivalent to

$$\ell(\bar{x}) - \ell(x_k) \leq -\|\bar{x} - x_k\|_2^2. \tag{9}$$

Replacing (9) in (8) we get

$$F(x(\alpha)) - F(x_k) - \theta(\ell(x(\alpha)) - \ell(x_k)) \leq -\alpha(1-\theta)\|\bar{x} - x_k\|_2^2 + \frac{L}{2}\|x(\alpha) - x_k\|_2^2.$$

Using the definition of $G(x_k)$ we have that $G(x_k) = \bar{x} - x_k$ and $\alpha G(x_k) = x(\alpha) - x_k$. Thus, we get

$$\begin{aligned} F(x(\alpha)) - F(x_k) - \theta(\ell(x(\alpha)) - \ell(x_k)) &\leq -\alpha(1-\theta)\|\bar{x} - x_k\|_2^2 + \frac{L}{2}\|x(\alpha) - x_k\|_2^2 \\ &\leq -\alpha(1-\theta)\|G(x_k)\|_2^2 + \frac{L\alpha^2}{2}\|G(x_k)\|_2^2 \\ &\leq 0 \quad (\text{for any } \alpha \leq \frac{2(1-\theta)}{L}). \end{aligned}$$

The last inequality implies that for any $\alpha \leq \frac{2(1-\theta)}{L}$ we have

$$F(x(\alpha)) - F(x_k) \leq \theta(\ell(x(\alpha)) - \ell(x_k)),$$

which is the termination condition of Armijo line-search for proximal gradient descent.

□

3 Properties of the proximal operator

Lemma 5 (The proximal operator is a non-expansive, i.e., Lipschitz continuous with constant 1)**.**

$$\|\textit{prox}_{\alpha g}(x) - \textit{prox}_{\alpha g}(y)\|_2 \leq \|x - y\|_2.$$

Proof. Let $u = \text{prox}_g(x)$ and $z = \text{prox}_g(y)$. From convexity of g we have that

$$g(z) \geq g(u) + d^T(z - u)\ \forall z, u \in \text{dom}\ g\ \forall d \in \partial g(u).$$

Using the second case of Lemma 2 we have that $x - u \in \partial g(u)$. Thus, by setting $d := x - u$ we get

$$g(z) \geq g(u) + (x - u)^T (z - u)\ \forall z, u \in \text{dom}\ g\ \forall. \tag{10}$$

Applying the same techniques, but by reversing the roles of z and u we get

$$g(u) \geq g(z) + (y - z)^T (u - z)\ \forall z, u \in \text{dom}\ g\ \forall. \tag{11}$$

Adding (10) and (11) we get

$$\begin{aligned}
0 &\geq (x-u)^T(z-u) + (y-z)^T(u-z) \\
&= (x-u)^T(z-u) - (y-z)^T(z-u) \\
&= (x-u-y+z)^T(z-u) \\
&= \|z-u\|_2^2 + (x-y)^T(z-u) \\
&= \|z-u\|_2^2 - (y-x)^T(z-u).
\end{aligned}$$

By re-arranging we get

$$\|z-u\|_2^2 \leq (y-x)^T(z-u).$$

Using Cauchy-Schwartz we get

$$\|z-u\|_2^2 \leq (y-x)^T(z-u) \leq \|y-x\|_2\|z-u\|_2.$$

Finally, dviding by $\|z-u\|_2$ and replacing the definitions of u and z we get

$$\|z-u\|_2 = \|\text{prox}_g(y) - \text{prox}_g(x)\|_2 \leq \|y-x\|_2.$$

□

Lemma 6 (Lipschitz continuity of the gradient mapping $G(x)$ (1))**.**

$$\|G(x) - G(y)\|_2 \leq \left(\frac{2}{\alpha} + L\right)\|x-y\|_2 \; \forall x, y \in \textit{dom } g.$$

Proof. using the definition of $G(x)$ (1) we get

$$\begin{aligned}
\|G(x) - G(y)\|_2 &= \frac{1}{\alpha}\|x - \text{prox}_{\alpha g}(x - \alpha\nabla f(x)) - y + \text{prox}_{\alpha g}(y - \alpha\nabla f(y))\|_2 \\
&\leq \frac{1}{\alpha}\|x-y\|_2 + \frac{1}{\alpha}\|\text{prox}_{\alpha g}(x - \alpha\nabla f(x)) - \text{prox}_{\alpha g}(y - \alpha\nabla f(y))\|_2 \\
&\leq \frac{1}{\alpha}\|x-y\|_2 + \frac{1}{\alpha}\|x - \alpha\nabla f(x) - y + \alpha\nabla f(y)\|_2 \quad \text{(using Lemma 5)} \\
&\leq \frac{1}{\alpha}\|x-y\|_2 + \frac{1}{\alpha}\|x-y\|_2 + \|\nabla f(x) - \nabla f(y)\|_2 \\
&= \frac{2}{\alpha}\|x-y\|_2 + \|\nabla f(x) - \nabla f(y)\|_2 \\
&\leq \frac{2}{\alpha}\|x-y\|_2 + L\|x-y\|_2 \quad \text{(using } \nabla f(x) \text{ is Lipschitz continuous)} \\
&= \left(\frac{2}{\alpha} + L\right)\|x-y\|_2.
\end{aligned}$$

□