**Continuous Optimization** **Srping 2023**

# Homework 5: Proximal Gradient Descent

*Lecturer: Aurelien Lucchi, Student: Julian Bopp*

### Homework 1 (Proximal operator for quadratics):

Consider the quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}$ where $\mathbf{A} \succcurlyeq 0$.

Prove that

$$\text{prox}_{\eta f}(\mathbf{x}) = (\mathbf{I} + \eta\mathbf{A})^{-1}(\mathbf{x} - \eta\mathbf{b}). \tag{1}$$

Proof: We start by writing down the definition of the prox operator

$$\begin{aligned}\text{prox}_{\eta f} &= \operatorname*{argmin}_{\mathbf{u}} \eta f(\mathbf{u}) + \frac{1}{2}\left\|\mathbf{u} - \mathbf{x}\right\|^2 \\ &= \operatorname*{argmin}_{\mathbf{u}} \underbrace{\frac{1}{2}\mathbf{u}^\top\mathbf{u} + \frac{\eta}{2}\mathbf{u}^\top\mathbf{A}\mathbf{u} + \mathbf{u}^\top(\eta\mathbf{b} - \mathbf{x}) + \eta\mathbf{c} + \frac{1}{2}\mathbf{x}^\top\mathbf{x}}_{g(\mathbf{u}):=}\end{aligned}$$

The minimum is attained where $\nabla g(\mathbf{u}) = 0$, because it is a quadratic convex function.
The gradient is given by

$$\nabla g(\mathbf{u}) = \mathbf{u} + \eta\mathbf{A}\mathbf{u} + \eta\mathbf{b} - \mathbf{x}.$$

Solving this for $\nabla g(\mathbf{u}) = 0$ yields

$$\mathbf{u} = (\mathbf{I} + \eta\mathbf{A})^{-1}(\mathbf{x} - \eta\mathbf{b}),$$

where we used that $\mathbf{A} \succcurlyeq 0,\ \eta > 0$ and therefore $(\mathbf{I} + \eta\mathbf{A})^{-1}$ exists, because the sum of SPD matrices is again SPD.

### Homework 2 (Projected Gradient Descent):

We want to minimize a function $f$ over a convex set $\mathcal{X}$. To do so, we use projected gradient descent that, starting from $\mathbf{x_0} \in \mathcal{X}$, performs the following updates:

$$\begin{aligned}\mathbf{y}_{k+1} &= \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k) \\ \mathbf{x}_{k+1} &= \Pi_{\mathcal{X}}(\mathbf{y}_{k+1}).\end{aligned}$$

1. Prove that for all $\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathbb{R}^d$:

$$H_{\mathbf{x}}(\mathbf{z}) := (\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{z}))^\top(\mathbf{z} - \Pi_{\mathcal{X}}(\mathbf{z})) \leq 0.$$

Proof: We notice that $\Pi_{\mathcal{X}}(\mathbf{z}) = \operatorname{argmin}_{u\in\mathcal{X}} \|u - z\|$. If $\mathbf{z} \in \mathcal{X}$ we have $\Pi_{\mathcal{X}}(\mathbf{z}) = z$ and therefore $H_{\mathbf{x}}(\mathbf{z}) = 0$ and the statement holds. Therefore we now assume $\mathbf{z} \in \mathbb{R}^d\backslash\mathcal{X}$.

Let $\mathbf{x} \in \mathcal{X}$. Since $\mathcal{X}$ is convex and $\Pi_{\mathcal{X}}(\mathbf{z}) \in \mathcal{X}$, the points on the line segment $\lambda\mathbf{x} + (1-\lambda)\Pi_{\mathcal{X}}(\mathbf{z})$ for $\lambda \in [0,1]$ are again in $\mathcal{X}$. We now use the fact that $\Pi_{\mathcal{X}}(\mathbf{z})$ is the closest point to $\mathbf{z}$ in $\mathcal{X}$.

$$\begin{aligned}\left\|\mathbf{z} - \Pi_{\mathcal{X}}(\mathbf{z})\right\|^2 &\leq \left\|\mathbf{z} - (\lambda\mathbf{x} + (1-\lambda)\Pi_{\mathcal{X}}(\mathbf{z}))\right\|^2 \\ &= \left\|\mathbf{z} - \Pi_{\mathcal{X}}(\mathbf{z})\right\|^2 - 2\,\underbrace{\lambda(\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{z}))^\top(\mathbf{z} - \Pi_{\mathcal{X}}(\mathbf{z}))}_{H_{\mathbf{x}}(\mathbf{z})} + \lambda^2\left\|\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{z})\right\|^2.\end{aligned}$$

Subtracting the term on the left on both sides yields

$$2\lambda H_{\mathbf{x}}(\mathbf{z}) \leq \lambda^2\left\|\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{z})\right\|^2.$$

Asuming $\lambda \neq 0$, dividing by $2\lambda$, and taking the limit as $\lambda$ goes to zero we get the result.

2. If $\mathbf{x}_{k+1} = \mathbf{x}_k$ after the projected gradient descent update, then $\mathbf{x}_k$ is a minimizer of $f$ over $\mathcal{X}$.

   Proof: Using the result from 1. and $\mathbf{z} = \mathbf{y}_{k+1}$ we get for every $\mathbf{x} \in \mathcal{X}$

$$(\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{y}_{k+1}))^\top (\mathbf{y}_{k+1} - \Pi_{\mathcal{X}}(\mathbf{y}_{k+1})) \leq 0$$

   Taking a look at the update rules of the projected gradient descent we see that this is the same as

$$(\mathbf{x} - \mathbf{x}_{k+1})^\top (\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k) - \mathbf{x}_{k+1}) \leq 0.$$

   Now using the fact that $\mathbf{x}_k = \mathbf{x}_{k+1}$ this becomes

$$(\mathbf{x} - \mathbf{x}_k)^\top (-\eta \nabla f(\mathbf{x}_k)) \leq 0.$$

   This means that $-\nabla f(\mathbf{x}_k) \in N_{\mathcal{X}}(\mathbf{x}_k)$. Therefore $x_k$ satisfies the first-order optimality condition and is a minimizer of $f$ over $\mathcal{X}$.

### Homework 3 (Proximal Gradient Descent: Convergence analysis):

Assume that $f$ is $\beta$-smooth and $\mu$ strongly-convex, then

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq (1 - \eta\mu)^k \|\mathbf{x}_1 - \mathbf{x}^*\|^2 . \tag{2}$$

Proof: We assume that $\eta \leq \frac{1}{\beta}$. By Lemma 3 (in lecture notes of proximal gradient lecture) we have for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$f(\mathbf{x} - \eta G_\eta(\mathbf{x})) \leq f(\mathbf{z}) + \langle G_\eta(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle - \frac{\eta}{2} \|G_\eta(\mathbf{x})\|^2 - \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2 .$$

With $\mathbf{z} = \mathbf{x}^*,\ \mathbf{x} = \mathbf{x}_k$, we have

$$\begin{aligned}
f(\mathbf{x}_k - \eta G_\eta(\mathbf{x}_k)) - f(\mathbf{x}^*) &\leq \langle G_\eta(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle - \frac{\eta}{2} \|G_\eta(\mathbf{x}_k)\|^2 - \frac{\mu}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&= \frac{1}{2\eta} \left( 2(\eta G_\eta(\mathbf{x}_k))^\top (\mathbf{x}_k - \mathbf{x}^*) - \|\eta G_\eta(\mathbf{x}_k)\|^2 \right) - \frac{\mu}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&= \frac{1}{2\eta} \left( \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^* - \eta G_\eta(\mathbf{x}_k)\|^2 \right) - \frac{\mu}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&= \frac{1}{2\eta} \left( \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \right) - \frac{\mu}{2} \|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&= \frac{1}{2\eta} \left( (1 - \eta\mu) \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \right)
\end{aligned}$$

Since $\mathbf{x}^*$ is the minimizer of $f$, the left side is bounded from below by 0 and we get

$$0 \leq \frac{1}{2\eta} \left( (1 - \eta\mu) \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \right)$$

Rearranging yields

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq (1 - \eta\mu) \|\mathbf{x}_k - \mathbf{x}^*\|^2 .$$

Applying this inequality to the norm term on the right hand side $k - 1$ times yields the final inequality

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq (1 - \eta\mu)^k \|\mathbf{x}_1 - \mathbf{x}^*\|^2 .$$