# An Introduction to Non-smooth Optimization

## Lecture 03 - Proximal Gradient Descent

**Jingwei LIANG**

Institute of Natural Sciences, Shanghai Jiao Tong University

Email: optimization.sjtu@gmail.com

Office: Room 355, No. 6 Science Building

# Previously

We have seen two problems

- Non-negative least square (NLS)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 \quad \text{such that} \quad x_i \geq 0, i = 1, 2, ..., n.$$

- Sparse logistic regression

$$\min_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\mathbb{R}} \mu\|\boldsymbol{x}\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(1+\mathrm{e}^{-b_i(\boldsymbol{x}^T\boldsymbol{a}_i+y)}\right).$$

# Previously

We have seen two problems

- Non-negative least square (NLS)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 \quad \text{such that} \quad x_i \ge 0, i=1,2,...,n.$$

- Sparse logistic regression

$$\min_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\mathbb{R}} \mu\|\boldsymbol{x}\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(1+\mathrm{e}^{-b_i(\boldsymbol{x}^T\boldsymbol{a}_i+y)}\right).$$

Let $S=\left\{\boldsymbol{x}\in\mathbb{R}^n \,:\, x_i\ge 0,\ i=1,2,...,n\right\}$ and define

$$\iota_S(\boldsymbol{x}) = \begin{cases} 0: & \boldsymbol{x}\in S, \\ +\infty: & \boldsymbol{x}\notin S. \end{cases}$$

The NLS problem can be equivalently written as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \iota_S(\boldsymbol{x}) + \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2.$$

# Previously

We have seen two problems

- Non-negative least square (NLS)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 \quad \text{such that} \quad x_i \geq 0, i = 1, 2, ..., n.$$

- Sparse logistic regression

$$\min_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\mathbb{R}} \mu\|\boldsymbol{x}\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(1+\mathrm{e}^{-b_i(\boldsymbol{x}^T\boldsymbol{a}_i+y)}\right).$$

**Problem - Non-smooth optimization problem**

Let $F, R \in \Gamma_0(\mathbb{R}^n)$, consider

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{\Phi(\boldsymbol{x}) \stackrel{\text{def}}{=} F(\boldsymbol{x}) + R(\boldsymbol{x})\right\},$$

with

$F$: smooth differentiable with $\nabla F$ being $L$-Lipschitz continuous.

$R$: non-smooth with proximal mapping easy to compute.

# Previously

We have seen two problems

- Non-negative least square (NLS)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2 \quad \text{such that} \quad x_i \geq 0, i=1,2,...,n.$$

- Sparse logistic regression

$$\min_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\mathbb{R}} \mu\|\boldsymbol{x}\|_1 + \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\mathrm{e}^{-b_i(\boldsymbol{x}^T\boldsymbol{a}_i+y)}\right).$$

**Proposition - Optimality condition**

Suppose $\mathrm{zer}(\nabla F+\partial R)$ is non-empty, and let $\boldsymbol{x}^\star \in \mathrm{zer}(\nabla F+\partial R)$. Then

$$\mathbf{0} \in \partial F(\boldsymbol{x}^\star) + \partial R(\boldsymbol{x}^\star).$$

# Outline

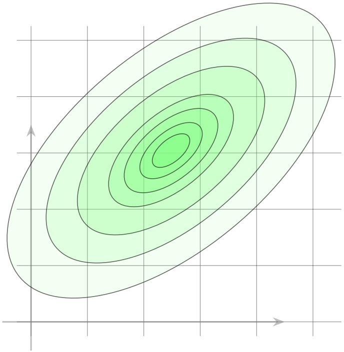
1. Gradient descent
2. Projected gradient descent
3. Proximal gradient descent
4. Convergence analysis

**Problem - Unconstrained smooth optimization**

Consider minimizing

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}),$$

where $F:\mathbb{R}^n \to \mathbb{R}$ is proper convex and smooth differentiable.



Assumptions:

- $\nabla F(\boldsymbol{x})$ is $L$-Lipschitz continuous for some $L > 0$

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

- Set of minimizers is non-empty, *i.e.* $\mathrm{Argmin}(F) \neq \emptyset$.

**NB**: $C_L^1(\mathbb{R}^n)$ — proper convex functions with $L$-Lipschitz (with $0 < L < +\infty$) continuous gradient on $\mathbb{R}^n$.

**Algorithm - Gradient descent**

**initial**: $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$;

**repeat**:

1. Choose step-size $\gamma_k > 0$
2. Update $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \gamma_k \nabla F(\boldsymbol{x}^{(k)})$

**until**: stopping criterion is satisfied.

Stopping criterion: $\epsilon > 0$ is the tolerance,

- Function value: $F(\boldsymbol{x}^{(k+1)}) - F(\boldsymbol{x}^{(k)}) \le \epsilon$.
- Sequence: $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| \le \epsilon$.
- Optimality condition: $\|\nabla F(\boldsymbol{x}^{(k)})\| \le \epsilon$.

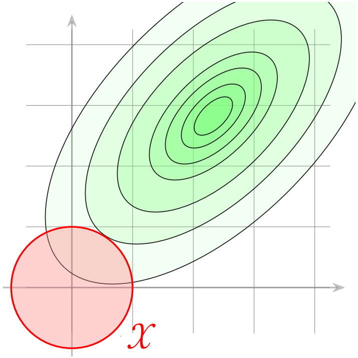
# Projected gradient descent

## Constrained smooth optimization

**Problem - Constrained smooth optimization**

Let $S \subset \mathbb{R}^n$ be closed and convex and $F \in C_L^1(\mathbb{R}^n)$,

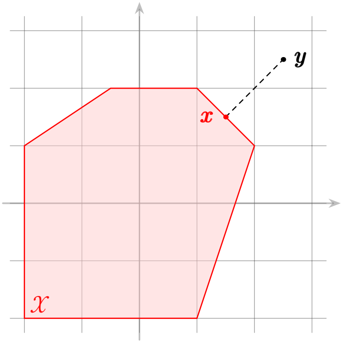
$$\min_{\boldsymbol{x} \in \mathbb{R}^n} F(\boldsymbol{x}) \quad \text{such that} \quad \boldsymbol{x} \in S.$$



# Projection operator

Projection of $\boldsymbol{y}$ onto $S$:

$$\min_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|.$$




**Definition - Projection**

Projection mapping onto a set is defined by

$$\mathcal{P}_S(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

The projection is unique for closed and convex, and $S$

$$\mathcal{P}_S(\boldsymbol{y}) \in S \quad \text{and} \quad \forall \boldsymbol{x} \in S \qquad \langle \boldsymbol{x} - \mathcal{P}_S(\boldsymbol{y}) \mid \boldsymbol{y} - \mathcal{P}_S(\boldsymbol{y}) \rangle \leq 0.$$

# Projected gradient descent

**Algorithm - Projected Gradient descent**

**initial**: $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$;

**repeat**:

1. Choose step-size $\gamma_k > 0$
2. GD: $\boldsymbol{x}^{(k+1/2)} = \boldsymbol{x}^{(k)} - \gamma_k \nabla F(\boldsymbol{x}^{(k)})$
3. Projection: $\boldsymbol{x}^{(k+1)} = \mathcal{P}_S(\boldsymbol{x}^{(k+1/2)})$

**until**: stopping criterion is satisfied.

In a compact form

$$\boldsymbol{x}^{(k+1)} = \mathcal{P}_S\big(\boldsymbol{x}^{(k)} - \gamma_k \nabla F(\boldsymbol{x}^{(k)})\big).$$

- The same as gradient descent, only one parameter which is $\gamma_k$.

# Proximal gradient descent

## Proximal mapping and algorithm

# From projection to proximal mapping

Previously

$$\mathcal{P}_S(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x}\in S} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

# From projection to proximal mapping

Previously

$$\mathcal{P}_S(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

The following are equivalent: $\iota_S(\boldsymbol{x}) \in \Gamma_0(\mathbb{R}^n)$ for closed convex $S$

$$\begin{aligned} \min_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \boldsymbol{y}\| \quad &\Longleftrightarrow \quad \min_{\boldsymbol{x} \in S} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ &\Longleftrightarrow \quad \min_{\boldsymbol{x} \in \mathbb{R}^n} \iota_S(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \end{aligned}$$

# From projection to proximal mapping

Previously

$$\mathcal{P}_S(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x}\in S} \|\boldsymbol{x}-\boldsymbol{y}\|.$$

The following are equivalent: $\iota_S(\boldsymbol{x}) \in \Gamma_0(\mathbb{R}^n)$ for closed convex $S$

$$\begin{aligned} \min_{\boldsymbol{x}\in S} \|\boldsymbol{x}-\boldsymbol{y}\| \quad &\Longleftrightarrow \quad \min_{\boldsymbol{x}\in S} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \\ &\Longleftrightarrow \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \iota_S(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2. \end{aligned}$$

Given any $R(\boldsymbol{x}) \in \Gamma_0(\mathbb{R}^n)$

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} R(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2.}$$

**Definition - Proximal mapping**

The proximal mapping (or proximity operator) of a function $R \in \Gamma_0(\mathbb{R}^n)$ is defined by

$$\operatorname{prox}_{\gamma R}(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} \gamma R(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

# Proximal mapping

**Definition - Proximal mapping**

The proximal mapping (or proximity operator) of a function $R \in \Gamma_0(\mathbb{R}^n)$ is defined by

$$\operatorname{prox}_{\gamma R}(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} \gamma R(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

- $\operatorname{prox}_{\gamma R}(\boldsymbol{y})$ is unique for $R \in \Gamma_0(\mathbb{R}^n)$.

**Definition - Proximal mapping**

The proximal mapping (or proximity operator) of a function $R \in \Gamma_0(\mathbb{R}^n)$ is defined by

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) \stackrel{\mathrm{def}}{=} \mathrm{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\ \gamma R(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

- $\mathrm{prox}_{\gamma R}(\boldsymbol{y})$ is unique for $R \in \Gamma_0(\mathbb{R}^n)$.
- Alternative characterization let $\boldsymbol{x} \stackrel{\mathrm{def}}{=} \mathrm{prox}_{\gamma R}(\boldsymbol{y})$,

$$\begin{aligned} \boldsymbol{0} \in \gamma\partial R(\boldsymbol{x}) + \boldsymbol{x} - \boldsymbol{y} \quad &\Longleftrightarrow \quad \boldsymbol{y} - \boldsymbol{x} \in \gamma\partial R(\boldsymbol{x}) \\ &\Longleftrightarrow \quad \boldsymbol{y} \in (\boldsymbol{Id} + \gamma\partial R)(\boldsymbol{x}) \\ &\Longleftrightarrow \quad \boldsymbol{x} = (\boldsymbol{Id} + \gamma\partial R)^{-1}(\boldsymbol{y}). \end{aligned}$$

- $(\boldsymbol{Id} + \gamma\partial R)^{-1}$ is called the **resolvent** of $\gamma\partial R$.

# Examples

**Projection** $R(\boldsymbol{x}) = \iota_S(\boldsymbol{x})$, then

$$\partial \iota_S(\boldsymbol{x}) = \mathcal{N}_S(\boldsymbol{x}) = \left\{\boldsymbol{g} \ : \ \langle \boldsymbol{g} \,|\, \boldsymbol{u} - \boldsymbol{x} \rangle \le 0, \ \ \forall \boldsymbol{u} \in S \right\}$$

and

$$\mathcal{P}_S(\boldsymbol{y}) = (\boldsymbol{Id} + \mathcal{N}_S)^{-1}(\boldsymbol{y}).$$

# Examples

**Projection** $R(\boldsymbol{x}) = \iota_S(\boldsymbol{x})$, then

$$\partial \iota_S(\boldsymbol{x}) = \mathcal{N}_S(\boldsymbol{x}) = \big\{\boldsymbol{g} \;:\; \langle \boldsymbol{g} \,|\, \boldsymbol{u} - \boldsymbol{x} \rangle \leq 0, \;\; \forall \boldsymbol{u} \in S \big\}$$

and

$$\mathcal{P}_S(\boldsymbol{y}) = (\boldsymbol{Id} + \mathcal{N}_S)^{-1}(\boldsymbol{y}).$$

**Examples**

- Hyperplane: $S = \{\boldsymbol{x} : \boldsymbol{a}^T\boldsymbol{x} = b\}$, $\boldsymbol{a} \neq 0$

$$\mathcal{P}_S(\boldsymbol{y}) = \boldsymbol{y} + \frac{b - \boldsymbol{a}^T\boldsymbol{y}}{\|\boldsymbol{a}\|^2}\boldsymbol{a}.$$

- Affine subspace: $S = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(\boldsymbol{A}) = m < n$

$$\mathcal{P}_S(\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}).$$

- Nonnegative orthant: $S = \mathbb{R}^n_+$

$$\mathcal{P}_S(\boldsymbol{y}) = \big(\max\{0, y_i\}\big)_i.$$

**Quadratic function** $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ being symmetric and positive semi-definite

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = (\boldsymbol{Id} + \gamma\boldsymbol{A})^{-1}(\boldsymbol{y} - \gamma\boldsymbol{b}).$$

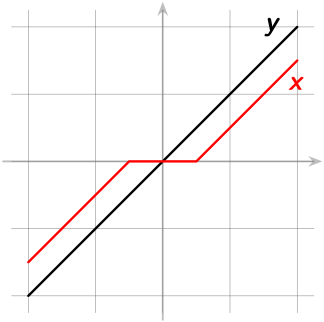
# Examples

**Quadratic function** $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ being symmetric and positive semi-definite

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = (\boldsymbol{Id} + \gamma\boldsymbol{A})^{-1}(\boldsymbol{y} - \gamma\boldsymbol{b}).$$

**Soft-threshold:** $R(x) = |x|$,

$$\mathrm{prox}_{\gamma R}(y) = \mathcal{T}_\gamma(y) = \begin{cases} y - \gamma : y > \gamma, \\ \quad\quad 0 : y \in [-\gamma, \gamma], \\ y + \gamma : y < -\gamma. \end{cases}$$



# Examples

**Quadratic function** $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ being symmetric and positive semi-definite

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = (\boldsymbol{Id} + \gamma\boldsymbol{A})^{-1}(\boldsymbol{y} - \gamma\boldsymbol{b}).$$

**Soft-threshold**: $R(x) = |x|$,

$$\mathrm{prox}_{\gamma R}(y) = \mathcal{T}_\gamma(y) = \begin{cases} y - \gamma : y > \gamma, \\ 0 : y \in [-\gamma, \gamma], \\ y + \gamma : y < -\gamma. \end{cases}$$

**Euclidean norm** $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = \begin{cases} (1 - \frac{\gamma}{\|\boldsymbol{y}\|})\boldsymbol{y} : \|\boldsymbol{y}\| > \gamma, \\ \boldsymbol{0} : \text{o.w.} \end{cases}$$

# Examples

**Quadratic function** $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ being symmetric and positive semi-definite

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = (\boldsymbol{Id} + \gamma\boldsymbol{A})^{-1}(\boldsymbol{y} - \gamma\boldsymbol{b}).$$

**Soft-threshold**: $R(x) = |x|$,

$$\mathrm{prox}_{\gamma R}(y) = \mathcal{T}_\gamma(y) = \begin{cases} y - \gamma : y > \gamma, \\ \quad 0 : y \in [-\gamma, \gamma], \\ y + \gamma : y < -\gamma. \end{cases}$$

**Euclidean norm** $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = \begin{cases} (1 - \frac{\gamma}{\|\boldsymbol{y}\|})\boldsymbol{y} : \|\boldsymbol{y}\| > \gamma, \\ \quad \boldsymbol{0} : \text{o.w.} \end{cases}$$

**Nuclear norm** $R(\boldsymbol{x}) = \sum_i \boldsymbol{S}_i$, let $\boldsymbol{y} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \in \mathbb{R}^{m\times n}$

$$\mathrm{prox}_{\gamma R}(\boldsymbol{y}) = \boldsymbol{U}\mathcal{T}_\gamma(\boldsymbol{S})\boldsymbol{V}^T.$$

**Quadratic perturbation** $H(\boldsymbol{x}) = R(\boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{x}\|^2 + \langle \boldsymbol{x} \mid \boldsymbol{u} \rangle + b,\ \alpha \geq 0$

$$\operatorname{prox}_H(\boldsymbol{y}) = \operatorname{prox}_{R/(\alpha+1)}\left(\frac{\boldsymbol{y}-\boldsymbol{u}}{\alpha+1}\right).$$

# Calculus

**Quadratic perturbation** $H(\boldsymbol{x}) = R(\boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{x}\|^2 + \langle \boldsymbol{x} \mid \boldsymbol{u} \rangle + b,\ \alpha \geq 0$

$$\operatorname{prox}_H(\boldsymbol{y}) = \operatorname{prox}_{R/(\alpha+1)}\left(\frac{\boldsymbol{y}-\boldsymbol{u}}{\alpha+1}\right).$$

**Translation** $H(\boldsymbol{x}) = R(\boldsymbol{x} - \boldsymbol{z})$

$$\operatorname{prox}_H(\boldsymbol{y}) = \boldsymbol{z} + \operatorname{prox}_R(\boldsymbol{y} - \boldsymbol{z}).$$

**Quadratic perturbation** $H(\mathbf{x}) = R(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|^2 + \langle \mathbf{x} \mid \mathbf{u} \rangle + b, \ \alpha \geq 0$

$$\mathrm{prox}_H(\mathbf{y}) = \mathrm{prox}_{R/(\alpha+1)}\left(\frac{\mathbf{y}-\mathbf{u}}{\alpha+1}\right).$$

**Translation** $H(\mathbf{x}) = R(\mathbf{x} - \mathbf{z})$

$$\mathrm{prox}_H(\mathbf{y}) = \mathbf{z} + \mathrm{prox}_R(\mathbf{y} - \mathbf{z}).$$

**Scaling** $H(\mathbf{x}) = R(\mathbf{x}/\rho)$

$$\mathrm{prox}_H(\mathbf{y}) = \rho\, \mathrm{prox}_{R/\rho^2}\left(\mathbf{y}/\rho\right).$$

# Calculus

**Quadratic perturbation** $H(\boldsymbol{x}) = R(\boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{x}\|^2 + \langle \boldsymbol{x} \mid \boldsymbol{u} \rangle + b,\ \alpha \geq 0$

$$\mathrm{prox}_H(\boldsymbol{y}) = \mathrm{prox}_{R/(\alpha+1)}\left(\frac{\boldsymbol{y}-\boldsymbol{u}}{\alpha+1}\right).$$

**Translation** $H(\boldsymbol{x}) = R(\boldsymbol{x} - \boldsymbol{z})$

$$\mathrm{prox}_H(\boldsymbol{y}) = \boldsymbol{z} + \mathrm{prox}_R(\boldsymbol{y} - \boldsymbol{z}).$$

**Scaling** $H(\boldsymbol{x}) = R(\boldsymbol{x}/\rho)$

$$\mathrm{prox}_H(\boldsymbol{y}) = \rho\, \mathrm{prox}_{R/\rho^2}(\boldsymbol{y}/\rho).$$

**Reflection** $H(\boldsymbol{x}) = R(-\boldsymbol{x})$

$$\mathrm{prox}_H(\boldsymbol{y}) = -\mathrm{prox}_R(-\boldsymbol{y}).$$

# Calculus

**Quadratic perturbation** $H(\mathbf{x}) = R(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x}\|^2 + \langle \mathbf{x} \mid \mathbf{u} \rangle + b,\ \alpha \geq 0$

$$\mathrm{prox}_H(\mathbf{y}) = \mathrm{prox}_{R/(\alpha+1)}\left(\frac{\mathbf{y}-\mathbf{u}}{\alpha+1}\right).$$

**Translation** $H(\mathbf{x}) = R(\mathbf{x} - \mathbf{z})$

$$\mathrm{prox}_H(\mathbf{y}) = \mathbf{z} + \mathrm{prox}_R(\mathbf{y} - \mathbf{z}).$$

**Scaling** $H(\mathbf{x}) = R(\mathbf{x}/\rho)$

$$\mathrm{prox}_H(\mathbf{y}) = \rho\,\mathrm{prox}_{R/\rho^2}(\mathbf{y}/\rho).$$

**Reflection** $H(\mathbf{x}) = R(-\mathbf{x})$

$$\mathrm{prox}_H(\mathbf{y}) = -\mathrm{prox}_R(-\mathbf{y}).$$

**Composition** $H = R \circ \mathbf{K}$ with $\mathbf{K}$ being bijective bounded linear mapping such that $\mathbf{K}^{-1} = \mathbf{K}^*$,

$$\mathrm{prox}_H(\mathbf{y}) = \mathbf{K}^*\mathrm{prox}_R(\mathbf{K}\mathbf{y}).$$

# Proximal gradient descent

**Problem - Non-smooth optimization**

Let $R \in \Gamma_0(\mathbb{R}^n)$ and $F \in C_L^1(\mathbb{R}^n)$,

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} R(\boldsymbol{x}) + F(\boldsymbol{x}).$$

# Proximal gradient descent

**Problem - Non-smooth optimization**

Let $R \in \Gamma_0(\mathbb{R}^n)$ and $F \in C_L^1(\mathbb{R}^n)$,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} R(\boldsymbol{x}) + F(\boldsymbol{x}).$$

**Algorithm - Proximal Gradient descent**

**initial**: $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$;

**repeat**:

1. Choose step-size $\gamma_k > 0$
2. GD: $\boldsymbol{x}^{(k+1/2)} = \boldsymbol{x}^{(k)} - \gamma_k \nabla F(\boldsymbol{x}^{(k)})$
3. Projection: $\boldsymbol{x}^{(k+1)} = \mathrm{prox}_{\gamma_k R}(\boldsymbol{x}^{(k+1/2)})$

**until**: stopping criterion is satisfied.

In a compact form

$$\boldsymbol{x}^{(k+1)} = \mathrm{prox}_{\gamma_k R}\big(\boldsymbol{x}^{(k)} - \gamma_k \nabla F(\boldsymbol{x}^{(k)})\big).$$

# Convergence analysis

## Fixed-point iteration perspective

# Optimality condition

**Problem - Non-smooth optimization problem**

Let $F, R \in \Gamma_0(\mathbb{R}^n)$, consider

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \Phi(\boldsymbol{x}) \stackrel{\text{def}}{=} F(\boldsymbol{x}) + R(\boldsymbol{x}) \right\},$$

with

- $F$: smooth differentiable with $\nabla F$ being $L$-Lipschitz continuous.
- $R$: non-smooth with proximal mapping easy to compute.

# Optimality condition

**Problem - Non-smooth optimization problem**

Let $F, R \in \Gamma_0(\mathbb{R}^n)$, consider

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ \Phi(\boldsymbol{x}) \stackrel{\text{def}}{=} F(\boldsymbol{x}) + R(\boldsymbol{x}) \right\},$$

with

- $F$: smooth differentiable with $\nabla F$ being $L$-Lipschitz continuous.
- $R$: non-smooth with proximal mapping easy to compute.

**Proposition - Optimality condition**

Suppose $\mathrm{zer}(\nabla F + \partial R)$ is non-empty, and let $\boldsymbol{x}^\star \in \mathrm{zer}(\nabla F + \partial R)$. Then

$$\mathbf{0} \in \nabla F(\boldsymbol{x}^\star) + \partial R(\boldsymbol{x}^\star).$$

# Cocoercive operators

**Definition - Cocoercive operator**

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. Then $\mathcal{B}$ is $\beta$-cocoercive if

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \langle \mathbf{x} - \mathbf{y} \mid \mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y}) \rangle \geq \beta \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\|^2.$$

# Cocoercive operators

## Definition - Cocoercive operator

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. Then $\mathcal{B}$ is $\beta$-cocoercive if

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \langle \mathbf{x} - \mathbf{y} \mid \mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y}) \rangle \geq \beta \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\|^2.$$

## Theorem - Cocoercivity and Lipschitz continuity

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. If $\mathcal{B}$ is $\beta$-cocoercive, then

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\| \leq \frac{1}{\beta} \|\mathbf{x} - \mathbf{y}\|.$$

- Cocoercivity implies Lipschitz continuity, the reverse in general is not true.

# Cocoercive operators

**Definition - Cocoercive operator**

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. Then $\mathcal{B}$ is $\beta$-cocoercive if

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \langle \mathbf{x} - \mathbf{y} \mid \mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y}) \rangle \geq \beta \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\|^2.$$

**Theorem - Cocoercivity and Lipschitz continuity**

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. If $\mathcal{B}$ is $\beta$-cocoercive, then

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\| \leq \frac{1}{\beta} \|\mathbf{x} - \mathbf{y}\|.$$

- Cocoercivity implies Lipschitz continuity, the reverse in general is not true.

**Theorem - [Baillon-Haddad '77]**

For $F \in C_L^1(\mathbb{R}^n)$, its gradient $\nabla F$ is $\frac{1}{L}$-cocoercive.

# Cocoercive operators

**Definition - Cocoercive operator**

Let $S$ be a non-empty subset of $\mathbb{R}^n$, $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta > 0$. Then $\mathcal{B}$ is $\beta$-cocoercive if

$$(\forall \mathbf{x} \in S)(\mathbf{y} \in S) \quad \langle \mathbf{x} - \mathbf{y} \mid \mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y}) \rangle \geq \beta \|\mathcal{B}(\mathbf{x}) - \mathcal{B}(\mathbf{y})\|^2.$$

**Proposition - Cocoercivity and non-expansiveness**

Let $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ be $\beta$-cocoercive for some $\beta > 0$, then

- $\beta\mathcal{B}$ is firmly non-expansive.
- $\boldsymbol{Id} - \gamma\mathcal{B}$ is $\frac{\gamma}{2\beta}$-averaged non-expansive for $\gamma \in ]0, 2\beta[$.

# Resolvent of maximal monotone operator

**Definition - Resolvent**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone. The resolvent of $\mathcal{A}$ is

$$\mathcal{J}_{\mathcal{A}} = (\textbf{\textit{Id}} + \mathcal{A})^{-1}.$$

# Resolvent of maximal monotone operator

**Definition - Resolvent**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone. The resolvent of $\mathcal{A}$ is

$$\mathcal{J}_{\mathcal{A}} = (\textbf{\textit{Id}} + \mathcal{A})^{-1}.$$

**Example - Proximal mapping**

Let $R \in \Gamma_0(\mathbb{R}^n)$ and $\gamma > 0$. Then

$$\mathcal{J}_{\gamma \partial R} = \mathrm{prox}_{\gamma R}.$$

**Definition - Resolvent**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone. The resolvent of $\mathcal{A}$ is

$$\mathcal{J}_{\mathcal{A}} = (\textbf{\textit{Id}} + \mathcal{A})^{-1}.$$

**Theorem - Monotonicity and firmly non-expansiveness**

Let $S$ be a nonempty subset of $\mathbb{R}^n$, let $\mathcal{F} : S \to \mathbb{R}^n$ and set $\mathcal{A} = \mathcal{F}^{-1} - \textbf{\textit{Id}}$. Then the following holds

- $\mathcal{F} = \mathcal{J}_{\mathcal{A}}$.
- $\mathcal{F}$ is *firmly non-expansive* if and only if $\mathcal{A}$ is monotone.
- $\mathcal{F}$ is *firmly non-expansive* and $S = \mathbb{R}^n$ if and only if $\mathcal{A}$ is *maximally monotone*.

# Resolvent of maximal monotone operator

**Definition - Resolvent**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone. The resolvent of $\mathcal{A}$ is

$$\mathcal{J}_{\mathcal{A}} = (\textbf{\textit{Id}} + \mathcal{A})^{-1}.$$

**Corollary**

Let $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^n$. Then $\mathcal{F}$ is firmly non-expansive if and only if it is the resolvent of a maximally monotone operator $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

## Fixed-point characterization

**Definition - Monotone inclusion problem**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal monotone and $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ be $\beta$-cocoercive for some $\beta > 0$. Then monotone inclusion problem associated to $\mathcal{A} + \mathcal{B}$ reads

$$\text{find} \quad \boldsymbol{x} \in \mathbb{R}^n \quad \text{such that} \quad \mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x}).$$

# Fixed-point characterization

## Definition - Monotone inclusion problem

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal monotone and $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ be $\beta$-cocoercive for some $\beta > 0$. Then monotone inclusion problem associated to $\mathcal{A} + \mathcal{B}$ reads

$$\text{find} \quad \boldsymbol{x} \in \mathbb{R}^n \quad \text{such that} \quad \mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x}).$$

## Definition - Forward–Backward splitting

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal monotone and $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$. Let $\boldsymbol{x}$ be such that $\mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x})$ and $\gamma > 0$. Then

$$\begin{aligned}
\mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x}) \quad &\Longleftrightarrow \quad \mathbf{0} \in \gamma\mathcal{A}(\boldsymbol{x}) + \gamma\mathcal{B}(\boldsymbol{x}) \\
&\Longleftrightarrow \quad -\gamma\mathcal{B}(\boldsymbol{x}) \in \gamma\mathcal{A}(\boldsymbol{x}) \\
&\Longleftrightarrow \quad \boldsymbol{x} - \gamma\mathcal{B}(\boldsymbol{x}) \in \boldsymbol{x} + \gamma\mathcal{A}(\boldsymbol{x}) \\
&\Longleftrightarrow \quad \boldsymbol{x} = (\boldsymbol{Id} + \gamma\mathcal{A})^{-1}(\boldsymbol{Id} - \gamma\mathcal{B})(\boldsymbol{x})
\end{aligned}$$

**Definition - Monotone inclusion problem**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal monotone and $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$ be $\beta$-cocoercive for some $\beta > 0$. Then monotone inclusion problem associated to $\mathcal{A} + \mathcal{B}$ reads

$$\text{find} \quad \boldsymbol{x} \in \mathbb{R}^n \quad \text{such that} \quad \mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x}).$$

**Proposition - Fixed-point operator**

Let $\mathcal{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal monotone and $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$. Set

$$\mathcal{F}_{\mathcal{A},\mathcal{B}} = \mathcal{J}_{\mathcal{A}} \circ (\boldsymbol{Id} - \mathcal{B}).$$

Suppose $\mathcal{B}$ is $\beta$-cocoercive for some $\beta > 0$ and that $\gamma \in ]0, 2\beta[$. Let

$$\alpha = \frac{2\beta}{4\beta - \gamma},$$

then $\mathcal{F}_{\gamma\mathcal{A},\gamma\mathcal{B}}$ is $\alpha$-averaged non-expansive.

**Theorem - Convergence with constant step-size**

For proximal gradient descent, let $R \in \Gamma_0(\mathbb{R}^n)$ and $F \in C_L^1(\mathbb{R}^n)$. Let

$$\gamma_k \equiv \gamma \in ]0, 2/L[.$$

Then

- $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ converges to a point $\boldsymbol{x}^\star$ in $\mathrm{zer}(\partial R + \nabla F)$.

**Theorem - Convergence speed**

With the above convergence result,

- Sequence

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| = o\left(\frac{1}{\sqrt{k}}\right).$$

- Objective function

$$(R+F)(\boldsymbol{x}^{(k)}) - (R+F)(\boldsymbol{x}^\star) = o\left(\frac{1}{k}\right).$$

# References


- Patrick L. Combettes, and Valérie R. Wajs. "Signal recovery by proximal forward-backward splitting." Multiscale modeling & simulation 4.4 (2005): 1168-1200.
- Cesare Molinari, Jingwei Liang, and Jalal Fadili. "Convergence rates of Forward–Douglas–Rachford splitting method." Journal of Optimization Theory and Applications 182.2 (2019): 606-639.
- Jingwei Liang, Jalal Fadili, and Gabriel Peyré. "Convergence rates with inexact non-expansive operators." Mathematical Programming 159.1 (2016): 403-434.
- A. Beck: First-order methods in optimization, Vol. 25. SIAM, 2017.
- H. H. Bauschke and P. L. Combettes: Convex analysis and monotone operator theory in Hilbert spaces, Vol. 408. New York: Springer, 2011.