An Introduction to Non-smooth Optimization

Lecture 06 - Several Acceleration Schemes

Jingwei LIANG

Institute of Natural Sciences, Shanghai Jiao Tong University

Email: optimization.sjtu@gmail.com
Office: Room 355, No. 6 Science Building

Outline

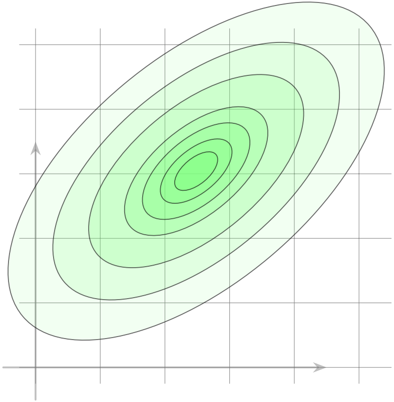
1. Accelerate gradient descent
2. Accelerate proximal gradient descent
3. Accelerate ADMM
4. Accelerate fixed-point iteration

Gradient descent

Problem - Unconstrained smooth optimization

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

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

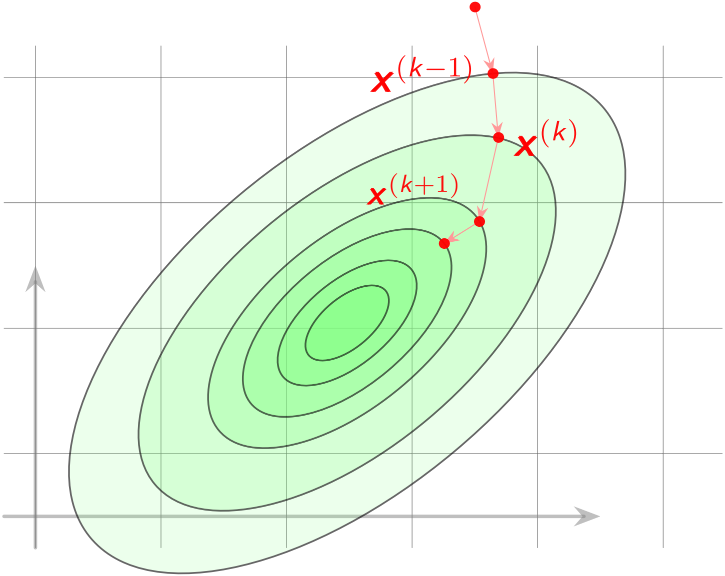


Algorithm - Gradient descent

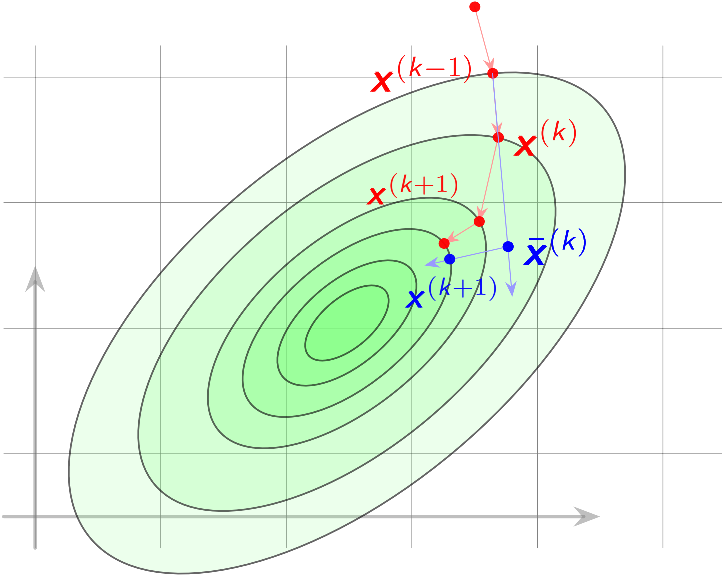
Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\gamma \in]0, 2/L[$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \gamma \nabla F(\boldsymbol{x}^{(k)}).$$

Gradient descent



Gradient descent



Heavy-ball method

Algorithm - Heavy-ball method [Polyak '64]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$, let $a > 0$ and $\gamma \in]0, 2/L[$

$$\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k)} + a(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}),$$
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{y}^{(k)} - \gamma \nabla F(\boldsymbol{x}^{(k)}).$$

Heavy-ball method

Algorithm - Heavy-ball method [Polyak '64]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$, let $a > 0$ and $\gamma \in]0, 2/L[$

$$\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k)} + a(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}),$$
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{y}^{(k)} - \gamma \nabla F(\boldsymbol{x}^{(k)}).$$

- $\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$ is called the inertial term or momentum term.
- a is called the inertial parameter.
- In general, no convergence rate for $F \in C_L^1$. Local rate if moreover F is twice-differentiable and strongly convex.

Theorem - Optimal rate

Let $\boldsymbol{x}^\star$ be a (local) minimizer of F such that $\mu \boldsymbol{Id} \preceq \nabla^2 F(\boldsymbol{x}^\star) \preceq L\boldsymbol{Id}$ and choose a, γ with $a \in [0, 1[, \gamma \in]0, 2(1 + a)/L[$. There exists $\underline{\rho} < 1$ such that if $\underline{\rho} < \rho < 1$ and if $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}$ are close enough to $\boldsymbol{x}^\star$, one has

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\| \leq C\rho^k.$$

Moreover, if

$$a = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2, \quad \gamma = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \quad \text{then} \quad \underline{\rho} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

Theorem - Optimal rate

Let $\boldsymbol{x}^\star$ be a (local) minimizer of F such that $\mu \boldsymbol{Id} \preceq \nabla^2 F(\boldsymbol{x}^\star) \preceq L\boldsymbol{Id}$ and choose a, γ with $a \in [0,1[, \gamma \in]0, 2(1+a)/L[$. There exists $\underline{\rho} < 1$ such that if $\underline{\rho} < \rho < 1$ and if $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}$ are close enough to $\boldsymbol{x}^\star$, one has

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\| \leq C\rho^k.$$

Moreover, if

$$a = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2, \quad \gamma = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} \quad \text{then} \quad \underline{\rho} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

- Starting points need to close enough to $\boldsymbol{x}^\star$
- Almost the optimal rate can be achieve by gradient method (or first-order method)
- Gradient descent

$$\underline{\rho} = \frac{L - \mu}{L + \mu}.$$

Nesterov's acceleration scheme

Algorithm - Nesterov's acceleration scheme [Nesterov '83]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$; Let $\phi_0 \in]0,1[$ and $q = \mu/L$

$$
\begin{aligned}
\phi_{k+1}^2 &= (1-\phi_{k+1})\phi_k^2 + q\phi_{k+1} \\
a_k &= \frac{\phi_k(1-\phi_k)}{\phi_k^2 + \phi_{k+1}} \\
\boldsymbol{y}^{(k)} &= \boldsymbol{x}^{(k)} + a_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\
\boldsymbol{x}^{(k+1)} &= \boldsymbol{y}^{(k)} - \frac{1}{L}\nabla F(\boldsymbol{y}^{(k)})
\end{aligned}
$$

Nesterov's acceleration scheme

Algorithm - Nesterov's acceleration scheme [Nesterov '83]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$; Let $\phi_0 \in]0,1[$ and $q = \mu/L$

$$
\begin{aligned}
\phi_{k+1}^2 &= (1-\phi_{k+1})\phi_k^2 + q\phi_{k+1} \\
a_k &= \frac{\phi_k(1-\phi_k)}{\phi_k^2 + \phi_{k+1}} \\
\boldsymbol{y}^{(k)} &= \boldsymbol{x}^{(k)} + a_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\
\boldsymbol{x}^{(k+1)} &= \boldsymbol{y}^{(k)} - \frac{1}{L}\nabla F(\boldsymbol{y}^{(k)})
\end{aligned}
$$

Theorem - Convergence rate

Let $\phi_0 \geq \sqrt{\mu/L}$, then

$$
F(\boldsymbol{x}^{(k)}) - F(\boldsymbol{x}^\star) \leq \min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L}+k\sqrt{\nu})^2}\right\} \times \left(F(x_0) - F(\boldsymbol{x}^\star) + \frac{\nu}{2}\|x_0 - \boldsymbol{x}^\star\|^2\right),
$$

where $\nu = \frac{\phi_0(\phi_0 L - \mu)}{1-\phi_0}$.

Nesterov's acceleration scheme

Algorithm - Nesterov's acceleration scheme [Nesterov '83]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$; Let $\phi_0 \in]0,1[$ and $q = \mu/L$

$$\phi_{k+1}^2 = (1-\phi_{k+1})\phi_k^2 + q\phi_{k+1}$$
$$a_k = \frac{\phi_k(1-\phi_k)}{\phi_k^2+\phi_{k+1}}$$
$$\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k)} + a_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)})$$
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{y}^{(k)} - \frac{1}{L}\nabla F(\boldsymbol{y}^{(k)})$$

Parameter choices

- $F \in C_L^1$: $\phi_0 = 1$,

$$q = 0, \quad \phi_k \approx \frac{2}{k+1} \to 0 \quad \text{and} \quad a_k \approx \frac{1-\phi_k}{1+\phi_k} \to 1.$$

- $F \in S_{\mu,L}^1$: $\phi_0 = \sqrt{\mu/L}$

$$q = \sqrt{\frac{\mu}{L}}, \quad \phi_k \equiv \sqrt{\frac{\mu}{L}} \quad \text{and} \quad a_k \equiv \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}.$$

Accelerate proximal gradient descent

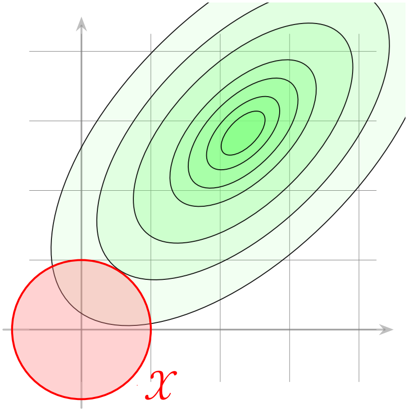
FISTA

Proximal gradient descent

Problem - Unconstrained smooth optimization

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

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



Algorithm - Gradient descent

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\gamma \in]0, 2/L[$

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

FISTA (Fast Iterative Soft-Thresholding Algorithm)

Algorithm - FISTA [Beck & Teboulle '09]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$; Let $t_0 = 1$ and $\gamma = 1/L$

$$
\begin{aligned}
t_k &= \frac{1+\sqrt{1+4t_{k-1}^2}}{2} \\
a_k &= \frac{t_{k-1}-1}{t_k} \\
\boldsymbol{y}^{(k)} &= \boldsymbol{x}^{(k)} + a_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\
\boldsymbol{x}^{(k+1)} &= \mathrm{prox}_{\gamma R}\big(\boldsymbol{y}^{(k)} - \gamma \nabla F(\boldsymbol{y}^{(k)})\big)
\end{aligned}
$$

- A special case of inertial proximal gradient descent.
- Inertial parameters

$$
t_k \approx \frac{k+1}{2} \quad \text{and} \quad a_k \to 1.
$$

FISTA (Fast Iterative Soft-Thresholding Algorithm)

Algorithm - FISTA [Beck & Teboulle '09]

Choose $\boldsymbol{x}^{(0)} \in \mathrm{dom}(F)$ and $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)}$; Let $t_0 = 1$ and $\gamma = 1/L$

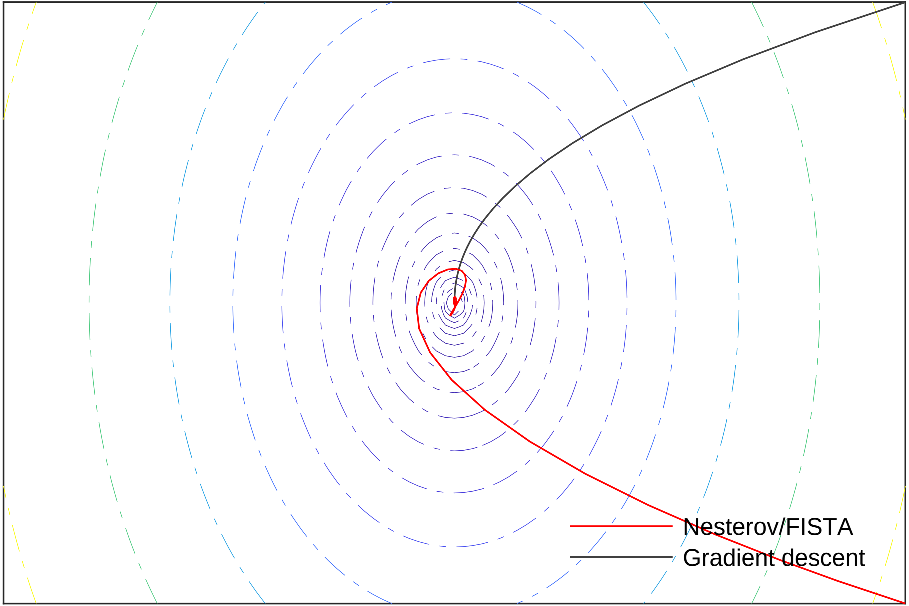
$$\begin{aligned}
t_k &= \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2} \\
a_k &= \frac{t_{k-1} - 1}{t_k} \\
\boldsymbol{y}^{(k)} &= \boldsymbol{x}^{(k)} + a_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\
\boldsymbol{x}^{(k+1)} &= \mathrm{prox}_{\gamma R}\left(\boldsymbol{y}^{(k)} - \gamma \nabla F(\boldsymbol{y}^{(k)})\right)
\end{aligned}$$

Theorem - Convergence rate

Let $\boldsymbol{x}^\star \in \mathrm{Argmin}(F + R)$,

$$(F + R)(\boldsymbol{x}^{(k)}) - (F + R)(\boldsymbol{x}^\star) \leq \frac{L\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|^2}{2(k+1)^2}.$$

Restarting FISTA



Restarting FISTA

Why FISTA oscillates

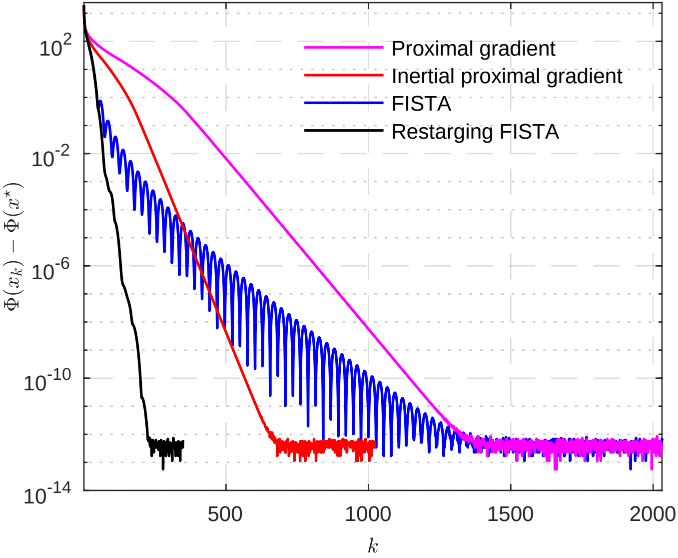
- for LSE, leading eigenvalue of the system is complex.
- over extrapolation, momentum beats gradient.

Algorithm - Restarting FISTA [O'Donoghue & Candés '12]

repeat:

1. Run FISTA iteration
2. If $\langle \mathbf{y}^{(k)} - \mathbf{x}^{(k+1)} \mid \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \rangle > 0$: $t_k = 1, \mathbf{y}^{(k)} = \mathbf{x}^{(k)}$.

Restarting FISTA



Accelerate ADMM

Strong convexity

Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^m} \; F(\boldsymbol{x}) + R(\boldsymbol{y}),$$
$$\text{such that } \boldsymbol{Ax} + \boldsymbol{By} = \boldsymbol{f}.$$

Algorithm - ADMM [Gabay, Mercier, Glowinski, Marrocco '76]

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} &\in \operatorname{Argmin}_{\boldsymbol{x}} F(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{Ax} + \boldsymbol{By}^{(k)} - \boldsymbol{f} + \boldsymbol{u}^{(k)}/\rho\|^2,\\
\boldsymbol{y}^{(k+1)} &\in \operatorname{Argmin}_{\boldsymbol{y}} R(\boldsymbol{y}) + \frac{\rho}{2}\|\boldsymbol{Ax}^{(k+1)} + \boldsymbol{By} - \boldsymbol{f} + \boldsymbol{u}^{(k)}/\rho\|^2,\\
\boldsymbol{u}^{(k+1)} &= \boldsymbol{u}^{(k)} + \rho(\boldsymbol{Ax}^{(k+1)} + \boldsymbol{By}^{(k+1)} - \boldsymbol{f}).
\end{aligned}$$

Fast ADMM under strong convexity

Assumption

- Both F and R are strongly convex.

Algorithm - Fast ADMM [Goldstein et al '14]

Let $\boldsymbol{y}^{(0)} \in \mathbb{R}^n, \bar{\boldsymbol{y}}^{(0)} = \boldsymbol{y}^{(0)}$ and $\boldsymbol{u}^{(0)} \in \mathbb{R}^p, \bar{\boldsymbol{u}}^{(0)} = \boldsymbol{u}^{(0)}$; Let $\rho > 0$ and $t_0 = 1$:

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} &\in \operatorname{Argmin}_{\boldsymbol{x}} F(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\bar{\boldsymbol{y}}^{(k)} - \boldsymbol{f} + \bar{\boldsymbol{u}}^{(k)}/\rho\|^2, \\
\boldsymbol{y}^{(k+1)} &\in \operatorname{Argmin}_{\boldsymbol{y}} R(\boldsymbol{y}) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x}^{(k+1)} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{f} + \bar{\boldsymbol{u}}^{(k)}/\rho\|^2, \\
\boldsymbol{u}^{(k+1)} &= \bar{\boldsymbol{u}}^{(k)} + \rho(\boldsymbol{A}\boldsymbol{x}^{(k+1)} + \boldsymbol{B}\boldsymbol{y}^{(k+1)} - \boldsymbol{f}) \\
t_k &= \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2} \\
\bar{\boldsymbol{y}}^{(k+1)} &= \boldsymbol{y}^{(k)} + \frac{t_{k-1} - 1}{t_k}(\boldsymbol{y}^{(k)} - \boldsymbol{y}^{(k-1)}) \\
\bar{\boldsymbol{u}}^{(k+1)} &= \boldsymbol{u}^{(k)} + \frac{t_{k-1} - 1}{t_k}(\boldsymbol{u}^{(k)} - \boldsymbol{u}^{(k-1)})
\end{aligned}$$

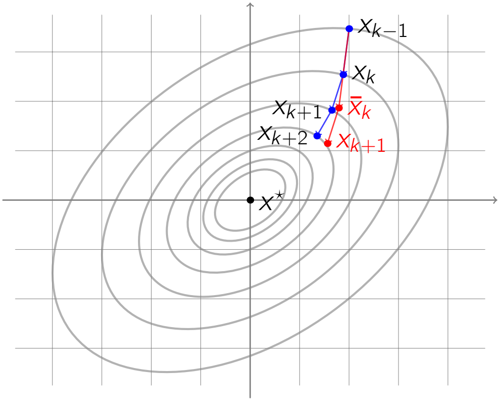
Accelerate fixed-point iteration

Inertial and over-relaxation

Accelerations: two approaches

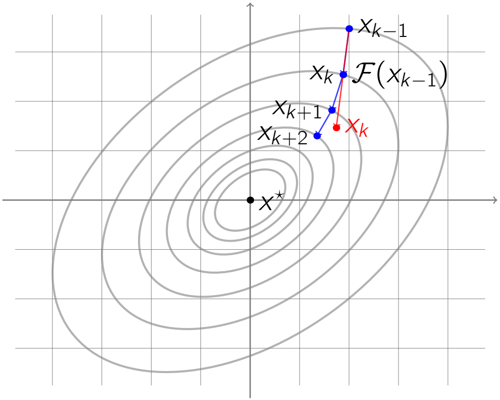
Algorithm - Inertial technique [Polyak '64, Nesterov '83, Beck & Teboulle '09]

$$
\left\lfloor
\begin{aligned}
\bar{\mathbf{x}}^{(k)} &= \mathbf{x}^{(k)} + a_k\big(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\big),\\
\mathbf{x}^{(k+1)} &= \mathcal{F}\big(\bar{\mathbf{x}}^{(k)}\big).
\end{aligned}
\right.
$$



Algorithm - Successive over-relaxation [Richardson '1911, Young '50]

$$\mathbf{x}^{(k+1)} = (1-\lambda_k)\mathbf{x}^{(k)} + \lambda_k \mathcal{F}(\mathbf{x}^{(k)}) \xRightarrow{a_k = \lambda_k - 1} \left\lfloor \begin{aligned} \bar{\mathbf{x}}^{(k)} &= \mathbf{x}^{(k)} + a_k(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k-1)}), \\ \mathbf{x}^{(k+1)} &= \mathcal{F}(\bar{\mathbf{x}}^{(k)}). \end{aligned} \right.$$



Alert: acceleration NOT guaranteed!

Problem - Sum of two Γ_0 functions

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

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \stackrel{\mathrm{def}}{=} \frac{1}{2}\big(\boldsymbol{Id} + (2\mathrm{prox}_{\gamma R} - \boldsymbol{Id})(2\mathrm{prox}_{\gamma F} - \boldsymbol{Id})\big).$$

Alert: acceleration NOT guaranteed!

Problem - Sum of two Γ_0 functions

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

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \stackrel{\mathrm{def}}{=} \frac{1}{2}\big(\mathbf{Id} + (2\mathrm{prox}_{\gamma R} - \mathbf{Id})(2\mathrm{prox}_{\gamma F} - \mathbf{Id})\big).$$

Douglas–Rachford splitting [Douglas & Rachford ’56]

$$\mathbf{z}^{(k+1)} = \mathcal{F}_{\mathrm{DR}}(\mathbf{z}^{(k)}),$$

- Sequence $o(1/\sqrt{k})$, objective **NA**.

Alert: acceleration NOT guaranteed!

Problem - Sum of two Γ_0 functions

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

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \stackrel{\text{def}}{=} \frac{1}{2}\Big(\boldsymbol{Id} + (2\mathrm{prox}_{\gamma R} - \boldsymbol{Id})(2\mathrm{prox}_{\gamma F} - \boldsymbol{Id})\Big).$$

Inertial Douglas–Rachford [Boţ, Csetnek & Hendrich ’15]

$$\begin{aligned} \bar{\boldsymbol{z}}_k &= \boldsymbol{z}^{(k)} + a_k(\boldsymbol{z}^{(k)} - \boldsymbol{z}^{(k-1)}), \\ \boldsymbol{z}^{(k+1)} &= \mathcal{F}_{\mathrm{DR}}(\bar{\boldsymbol{z}}_k). \end{aligned}$$

- No rates available, **may fail to provide acceleration.**

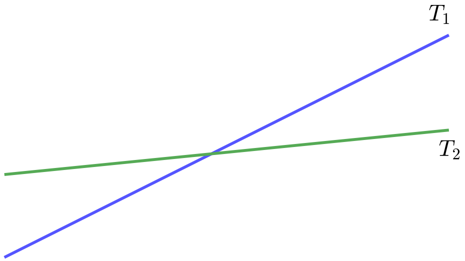
Problem - Feasibility problem in $\mathbb{R}^2$

Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

$$\text{Find } x \in \mathbb{R}^2 \quad \text{such that} \quad x \in T_1 \cap T_2.$$

Define

$$\mathcal{F}_{\text{DR}} \stackrel{\text{def}}{=} \frac{1}{2}\big(\mathbf{Id} + (2\mathcal{P}_{T_1} - \mathbf{Id})(2\mathcal{P}_{T_2} - \mathbf{Id})\big).$$



Alert: acceleration NOT guaranteed!

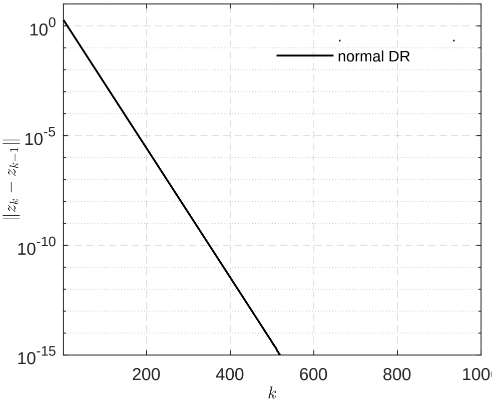
Problem - Feasibility problem in $\mathbb{R}^2$

Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

$$\text{Find } x \in \mathbb{R}^2 \ \text{ such that } \ x \in T_1 \cap T_2.$$

Douglas–Rachford:

$$\bar{z}_k = z^{(k)},$$
$$z^{(k+1)} = \mathcal{F}_{\mathrm{DR}}(\bar{z}_k).$$



Alert: acceleration NOT guaranteed!

Problem - Feasibility problem in $\mathbb{R}^2$

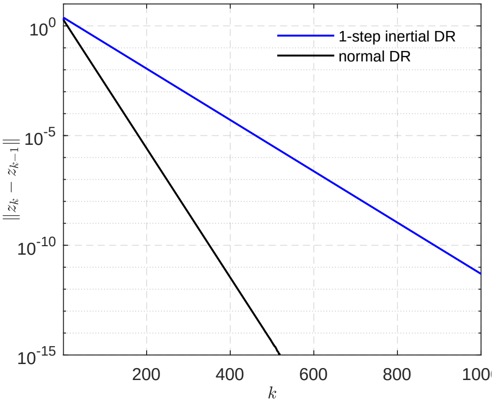
Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

$$\text{Find } x \in \mathbb{R}^2 \ \text{ such that } \ x \in T_1 \cap T_2.$$

Inertial Douglas–Rachford:

$$\bar{z}_k = z^{(k)} + a(z^{(k)} - z^{(k-1)}),$$
$$z^{(k+1)} = \mathcal{F}_{\mathrm{DR}}(\bar{z}_k).$$

- 1-step inertial: $a = 0.3$.



References

- B. Polyak. “Introduction to optimization”. Optimization Software, 1987.
- Y. Nesterov. “Introductory lectures on convex optimization: A basic course”. Vol. 87. Springer Science & Business Media, 2013.
- A. Beck and M. Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.
- B. O’Donoghue and E. J. Candés. “Adaptive restart for accelerated gradient schemes”. Foundations of Computational Mathematics, pages 1–18, 2012.
- T. Goldstein, B. O’Donoghue, S. Setzer, R. Baraniuk. “Fast alternating direction optimization methods”. SIAM Journal on Imaging Sciences, pages 1588-1623, 2014.
- 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.