

Chapter 2

Optimization Algorithms

Convex Optimization for Machine Learning & Computer Vision
SS 2017

Tao Wu
Thomas Möllenhoff
Emanuel Laude

Computer Vision Group
Department of Informatics
TU München

Optimization
Algorithms

Tao Wu
Thomas Möllenhoff
Emanuel Laude



Gradient Methods

Proximal Algorithms

Convergence Theory

Acceleration



Gradient-based Methods

Overview of this section

(Unconstrained) differentiable optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1 $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable.
- 2 There exists at least one minimizer u^* (s.t. $\nabla J(u^*) = 0$).



Overview of this section

(Unconstrained) differentiable optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1 $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable.
- 2 There exists at least one minimizer u^* (s.t. $\nabla J(u^*) = 0$).

Methods under investigation:

- 1 (Scaled) gradient descent.
- 2 Line search method.
- 3 Majorize-minimize method.

Remark: They apply to non-convex optimization.





(Unconstrained) differentiable optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1 $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable.
- 2 There exists at least one minimizer u^* (s.t. $\nabla J(u^*) = 0$).

Methods under investigation:

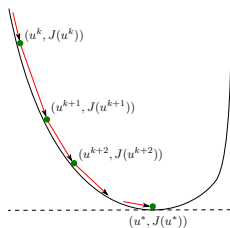
- 1 (Scaled) gradient descent.
- 2 Line search method.
- 3 Majorize-minimize method.

Remark: They apply to non-convex optimization.

General concerns:

- 1 Convergence (or not); global vs. local convergence.
- 2 Convergence rate (in special cases).

Descent algorithm



Descent method

Initialize $u^0 \in \mathbb{E}$. Iterate with $k = 0, 1, 2, \dots$

- 1 If the stopping criteria $\|\nabla J(u^k)\| \leq \epsilon$ is *not* satisfied, then continue; otherwise return u^k and stop.
- 2 Choose a **descent direction** $d^k \in \mathbb{E}$ s.t.

$$\langle \nabla J(u^k), d^k \rangle < 0.$$

- 3 Choose an *appropriate* step size $\tau^k > 0$, and update

$$u^{k+1} = u^k + \tau^k d^k.$$

Theorem

If $\langle \nabla J(u^k), d^k \rangle < 0$, then $J(u^k + \tau d^k) < J(u^k)$ for all sufficiently small $\tau > 0$.



Theorem

If $\langle \nabla J(u^k), d^k \rangle < 0$, then $J(u^k + \tau d^k) < J(u^k)$ for all sufficiently small $\tau > 0$.

Proof: Use the Taylor expansion:

$$\begin{aligned} J(u^k + \tau d^k) &= J(u^k) + \tau \langle \nabla J(u^k), d^k \rangle + o(\tau) \\ &= J(u^k) + \tau \left(\langle \nabla J(u^k), d^k \rangle + o(1) \right) < J(u^k) \quad \text{as } \tau \rightarrow 0^+. \end{aligned}$$



Theorem

If $\langle \nabla J(u^k), d^k \rangle < 0$, then $J(u^k + \tau d^k) < J(u^k)$ for all sufficiently small $\tau > 0$.

Proof: Use the Taylor expansion:

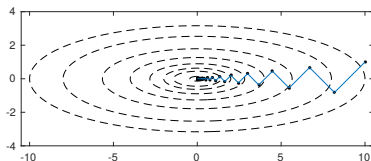
$$\begin{aligned} J(u^k + \tau d^k) &= J(u^k) + \tau \langle \nabla J(u^k), d^k \rangle + o(\tau) \\ &= J(u^k) + \tau \left(\langle \nabla J(u^k), d^k \rangle + o(1) \right) < J(u^k) \quad \text{as } \tau \rightarrow 0^+. \end{aligned}$$

Choices of descent direction

- 1 Scaled gradient: $d^k = -(H^k)^{-1} \nabla J(u^k)$.
- 2 Gradient/Steepest descent: $H^k = I$.
- 3 Newton: $H^k = \nabla^2 J(u^k)$, assuming J is twice continuously differentiable and $\nabla^2 J(u^k) \succ 0$.
- 4 Quasi-Newton: $H^k \approx \nabla^2 J(u^k)$, H^k is spd.



Gradient descent with exact line search



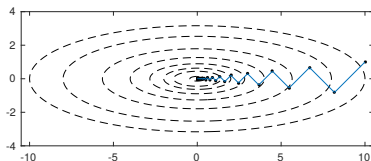
- Gradient descent with *exact* line search:

$$u^{k+1} = u^k - \tau^k \nabla J(u^k),$$

$$\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).$$



Gradient descent with exact line search



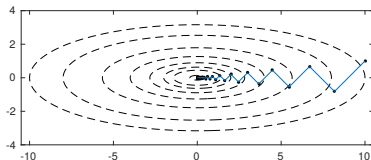
- Gradient descent with *exact* line search:

$$u^{k+1} = u^k - \tau^k \nabla J(u^k),$$
$$\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).$$

- Special case: $J(u) = \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle$, matrix Q is spd.
 - $\nabla J(u) = Qu - b$, $\|\cdot\|_Q^2 \equiv \langle \cdot, Q \cdot \rangle$.



Gradient descent with exact line search



- Gradient descent with *exact* line search:

$$u^{k+1} = u^k - \tau^k \nabla J(u^k),$$
$$\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).$$

- Special case: $J(u) = \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle$, matrix Q is spd.

- $\nabla J(u) = Qu - b$, $\|\cdot\|_Q^2 \equiv \langle \cdot, Q \cdot \rangle$.

- $\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)) = \frac{\|\nabla J(u^k)\|^2}{\|\nabla J(u^k)\|_Q^2} \Rightarrow$

$$\|u^{k+1} - u^*\|_Q^2 = \left(1 - \frac{\|\nabla J(u^k)\|^4}{\|\nabla J(u^k)\|_Q^2 \|\nabla J(u^k)\|_{Q^{-1}}^2} \right) \|u^k - u^*\|_Q^2$$

$$\text{(exercise)} \leq \left(\frac{\lambda_{\max}(Q) - \lambda_{\min}(Q)}{\lambda_{\max}(Q) + \lambda_{\min}(Q)} \right)^2 \|u^k - u^*\|_Q^2.$$



Backtracking line search

- Sufficient decrease condition (let $c_1 \in (0, 1)$):

$$J(u^k + \tau d^k) \leq J(u^k) + c_1 \tau \langle \nabla J(u^k), d^k \rangle. \quad (\text{A})$$

- Curvature condition (let $c_2 \in (c_1, 1)$):

$$\langle \nabla J(u^k + \tau d^k), d^k \rangle \geq c_2 \langle \nabla J(u^k), d^k \rangle. \quad (\text{C})$$



Backtracking line search

- Sufficient decrease condition (let $c_1 \in (0, 1)$):

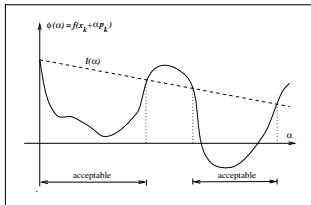
$$J(u^k + \tau d^k) \leq J(u^k) + c_1 \tau \langle \nabla J(u^k), d^k \rangle. \quad (\text{A})$$

- Curvature condition (let $c_2 \in (c_1, 1)$):

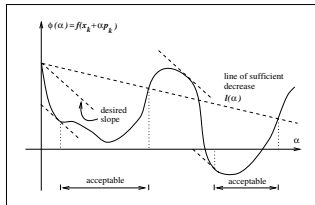
$$\langle \nabla J(u^k + \tau d^k), d^k \rangle \geq c_2 \langle \nabla J(u^k), d^k \rangle. \quad (\text{C})$$

- (A) \rightsquigarrow **Armijo** line search; (A) & (C) \rightsquigarrow **Wolfe-Powell** I.s.

Armijo I.s.



Wolfe-Powell I.s.



Convergence of backtracking line search

Lemma (feasibility of line search)

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable, $\langle \nabla J(u^k), d^k \rangle < 0 \forall k$, and $0 < c_1 < c_2 < 1$. Then there exists an open interval in which the step size τ satisfies (A) and (C).

Proof: on board.



Convergence of backtracking line search

Lemma (feasibility of line search)

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable, $\langle \nabla J(u^k), d^k \rangle < 0 \forall k$, and $0 < c_1 < c_2 < 1$. Then there exists an open interval in which the step size τ satisfies (A) and (C).

Proof: on board.

Theorem (Zoutendijk)

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is continuously differentiable, and (A) and (C) are both satisfied with $0 < c_1 < c_2 < 1$ for each k . In addition, $\nabla J(\cdot)$ is μ -Lipschitz on $\{u \in \mathbb{E} : J(u) \leq J(u^0)\}$. Then we have

$$\sum_{k=0}^{\infty} \frac{|\langle \nabla J(u^k), d^k \rangle|^2}{\|d^k\|^2} < \infty.$$

Proof: on board.

Remark

If $\frac{|\langle \nabla J(u^k), d^k \rangle|}{\|\nabla J(u^k)\| \|d^k\|} \geq C > 0$, then $\lim_{k \rightarrow \infty} \|\nabla J(u^k)\| = 0$.



Majorize-minimize method

Majorizing function

A function $\hat{J}(\cdot; u)$ is a **majorizing surrogate** of J at $u \in \mathbb{E}$ if

$$\begin{cases} \hat{J}(u; u) = J(u), \\ \hat{J}(\cdot; u) \geq J(\cdot). \end{cases}$$



Majorize-minimize method

Majorizing function

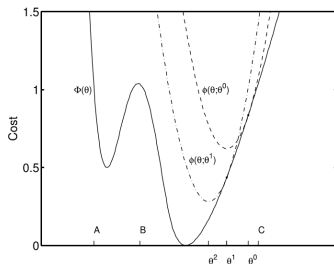
A function $\hat{J}(\cdot; u)$ is a **majorizing surrogate** of J at $u \in \mathbb{E}$ if

$$\begin{cases} \hat{J}(u; u) = J(u), \\ \hat{J}(\cdot; u) \geq J(\cdot). \end{cases}$$

Majorize-minimize (MM) algorithm

Let $\hat{J}(\cdot; u)$ majorize $J \forall u \in \mathbb{E}$. Then the MM iteration reads:

$$u^{k+1} \in \arg \min_u \hat{J}(u; u^k).$$



Remark

- 1 Monotonic decrease of objectives:

$$J(u^{k+1}) \leq \widehat{J}(u^{k+1}; u^k) \leq \widehat{J}(u^k; u^k) = J(u^k).$$

- 2 Efficiency of MM hinges on the choice of the majorizing surrogate $\widehat{J}(\cdot; u)$, i.e. $\widehat{J}(\cdot; u)$ should be easy to minimize.
- 3 Common choices of $\widehat{J}(\cdot; u)$ are convex quadratics.



Remark

- 1 Monotonic decrease of objectives:

$$J(u^{k+1}) \leq \widehat{J}(u^{k+1}; u^k) \leq \widehat{J}(u^k; u^k) = J(u^k).$$

- 2 Efficiency of MM hinges on the choice of the majorizing surrogate $\widehat{J}(\cdot; u)$, i.e. $\widehat{J}(\cdot; u)$ should be easy to minimize.
- 3 Common choices of $\widehat{J}(\cdot; u)$ are convex quadratics.



Gradient Methods

Proximal Algorithms

Convergence Theory

Acceleration

Gradient descent as MM

- Observe that $u^{k+1} = u^k - \tau \nabla J(u^k)$ iff
$$u^{k+1} = \arg \min_u J(u^k) + \langle \nabla J(u^k), u - u^k \rangle + \frac{1}{2\tau} \|u - u^k\|^2.$$
- When does $J(u^k) + \langle \nabla J(u^k), \cdot - u^k \rangle + \frac{1}{2\tau} \|\cdot - u^k\|^2 \geq J(\cdot)$ hold?

Gradient descent as MM

Lemma

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and has μ -Lipschitz gradient.
Then $\forall u, v \in \mathbb{E}$:

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.



Gradient descent as MM

Lemma

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and has μ -Lipschitz gradient. Then $\forall u, v \in \mathbb{E}$:

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.

Theorem (convergence of gradient descent)

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and has μ -Lipschitz gradient. Then the gradient descent iteration

$$u^{k+1} = u^k - \tau \nabla J(u^k)$$

with $\tau \in (0, 1/\mu]$ yields $\lim_{k \rightarrow \infty} \nabla J(u^k) = 0$.

Proof: on board.



Gradient descent as MM

Lemma

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and has μ -Lipschitz gradient. Then $\forall u, v \in \mathbb{E}$:

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.

Theorem (convergence of gradient descent)

Assume that $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and has μ -Lipschitz gradient. Then the gradient descent iteration

$$u^{k+1} = u^k - \tau \nabla J(u^k)$$

with $\tau \in (0, 1/\mu]$ yields $\lim_{k \rightarrow \infty} \nabla J(u^k) = 0$.

Proof: on board.

Recipe of (global) convergence

By solving the surrogate problem in MM, we achieve: (1) sufficient decrease in the objective; (2) inexact optimality condition (OC) which matches the exact OC in the limit.





Proximal Algorithms

Agenda for the rest of the chapter



- (Informal) derivation of proximal algorithms:
 - Forward-backward splitting (FBS), e.g. proximal gradient method.
 - Alternating direction method of multipliers (ADMM).
 - Primal-dual hybrid gradient (PDHG).
 - Douglas-Rachford splitting (DRS), Peaceman-Rachford splitting (PRS).

Agenda for the rest of the chapter



- (Informal) derivation of proximal algorithms:
 - Forward-backward splitting (FBS), e.g. proximal gradient method.
 - Alternating direction method of multipliers (ADMM).
 - Primal-dual hybrid gradient (PDHG).
 - Douglas-Rachford splitting (DRS), Peaceman-Rachford splitting (PRS).
- Application on examples.
- Connection/equivalence between proximal algorithms.
- Unified convergence analysis.
- Acceleration techniques.

Forward-backward splitting

- Consider

$$\min_u F(u) + G(u),$$

whose minimizer exists and satisfies

$$0 \in \partial F(u) + \nabla G(u).$$





- Consider

$$\min_u F(u) + G(u),$$

whose minimizer exists and satisfies

$$0 \in \partial F(u) + \nabla G(u).$$

- **Forward-backward splitting:**

$$\begin{aligned} u^{k+1} &= \text{prox}_{\tau F}(u^k - \tau \nabla G(u^k)) \\ &= (I + \tau \partial F)^{-1}(I - \tau \nabla G)(u^k). \end{aligned}$$

- FBS as *semi-implicit Euler scheme*:

$$-\frac{u^{k+1} - u^k}{\tau} \in \partial F(u^{k+1}) + \nabla G(u^k).$$

Example: Split feasibility problem

Split feasibility problem

Given nonempty, closed, convex sets $C_1 \subset \mathbb{E}_1$, $C_2 \subset \mathbb{E}_2$, and linear operator $K : \mathbb{E}_1 \rightarrow \mathbb{E}_2$, find $u \in \mathbb{E}_1$ s.t. $u \in C_1$, $Ku \in C_2$.

- Variational model:

$$\min_{u \in \mathbb{E}_1} \delta_{C_1}(u) + \frac{1}{2} \|Ku - \text{proj}_{C_2}(Ku)\|^2.$$



Example: Split feasibility problem

Split feasibility problem

Given nonempty, closed, convex sets $C_1 \subset \mathbb{E}_1$, $C_2 \subset \mathbb{E}_2$, and linear operator $K : \mathbb{E}_1 \rightarrow \mathbb{E}_2$, find $u \in \mathbb{E}_1$ s.t. $u \in C_1$, $Ku \in C_2$.

- Variational model:

$$\min_{u \in \mathbb{E}_1} \delta_{C_1}(u) + \frac{1}{2} \|Ku - \text{proj}_{C_2}(Ku)\|^2.$$

- Optimality condition:

$$0 \in \partial \delta_{C_1}(u) + K^\top (I - \text{proj}_{C_2})(Ku).$$



Example: Split feasibility problem

Split feasibility problem

Given nonempty, closed, convex sets $C_1 \subset \mathbb{E}_1$, $C_2 \subset \mathbb{E}_2$, and linear operator $K : \mathbb{E}_1 \rightarrow \mathbb{E}_2$, find $u \in \mathbb{E}_1$ s.t. $u \in C_1$, $Ku \in C_2$.



- Variational model:

$$\min_{u \in \mathbb{E}_1} \delta_{C_1}(u) + \frac{1}{2} \|Ku - \text{proj}_{C_2}(Ku)\|^2.$$

- Optimality condition:

$$0 \in \partial \delta_{C_1}(u) + K^\top (I - \text{proj}_{C_2})(Ku).$$

- Apply FBS \Rightarrow

$$\begin{aligned} u^{k+1} &= (I + \tau \partial \delta_{C_1})^{-1} (u^k - \tau K^\top (I - \text{proj}_{C_2})(Ku^k)) \\ &= \text{proj}_{C_1} (u^k - \tau K^\top (I - \text{proj}_{C_2})(Ku^k)). \end{aligned}$$

Example: Regularized least squares

Regularized least squares

$$\min_u F(u) + \frac{1}{2} \|A(u) - b\|^2,$$

where

- A : differentiable operator modeling forward process.
- b : observation.
- F : regularization/prior term.
 - $\text{prox}_{\tau F}$ is easy to compute.
 - E.g. $F(\cdot) = \|\cdot\|_2^2$, $F(\cdot) = \|\cdot\|_1$, or $F(\cdot) = \|\cdot\|_{\text{nuc}}$.



Example: Regularized least squares

Regularized least squares

$$\min_u F(u) + \frac{1}{2} \|A(u) - b\|^2,$$

where

- A : differentiable operator modeling forward process.
- b : observation.
- F : regularization/prior term.
 - $\text{prox}_{\tau F}$ is easy to compute.
 - E.g. $F(\cdot) = \|\cdot\|_2^2$, $F(\cdot) = \|\cdot\|_1$, or $F(\cdot) = \|\cdot\|_{\text{nuc}}$.

- Optimality condition:

$$0 \in \partial F(u) + \nabla A(u)^\top (A(u) - b).$$

- Apply FBS \Rightarrow

$$u^{k+1} = \text{prox}_{\tau F}(u^k - \tau \nabla A(u^k)^\top (A(u^k) - b)).$$



Alternating direction method of multipliers

- Consider

$$\min_{u,v} J(u, v) = F(v) + G(u) + \delta\{Ku - v = 0\},$$

for given proper, convex, lsc functions F , G and matrix K .



Alternating direction method of multipliers

- Consider

$$\min_{u,v} J(u, v) = F(v) + G(u) + \delta\{Ku - v = 0\},$$

for given proper, convex, lsc functions F , G and matrix K .

- *Augmented Lagrangian* (with $\tau > 0$):

$$\mathcal{L}_\tau(u, v; p) = F(v) + G(u) + \langle p, Ku - v \rangle + \frac{\tau}{2} \|Ku - v\|^2,$$

such that

$$\min_{u,v} J(u, v) = \sup_p \min_{u,v} \mathcal{L}_\tau(u, v; p).$$



Alternating direction method of multipliers

- Consider

$$\min_{u,v} J(u, v) = F(v) + G(u) + \delta\{Ku - v = 0\},$$

for given proper, convex, lsc functions F , G and matrix K .

- Augmented Lagrangian* (with $\tau > 0$):

$$\mathcal{L}_\tau(u, v; p) = F(v) + G(u) + \langle p, Ku - v \rangle + \frac{\tau}{2} \|Ku - v\|^2,$$

such that

$$\min_{u,v} J(u, v) = \sup_p \min_{u,v} \mathcal{L}_\tau(u, v; p).$$

- Alternating direction method of multipliers (ADMM):**

$$\begin{cases} u^{k+1} \in \arg \min_u G(u) + \langle p^k, Ku \rangle + \frac{\tau}{2} \|Ku - v^k\|^2, \\ v^{k+1} \in \arg \min_v F(v) - \langle p^k, v \rangle + \frac{\tau}{2} \|Ku^{k+1} - v\|^2, \\ p^{k+1} = p^k + \tau(Ku^{k+1} - v^{k+1}). \end{cases}$$



Primal-dual hybrid gradient

- In view of the duality theorem, reformulate

$$\min_u F(Ku) + G(u)$$

as the saddle-point problem:

$$\max_p \min_u \langle p, Ku \rangle + G(u) - F^*(p).$$



Primal-dual hybrid gradient



- In view of the duality theorem, reformulate

$$\min_u F(Ku) + G(u)$$

as the saddle-point problem:

$$\max_p \min_u \langle p, Ku \rangle + G(u) - F^*(p).$$

- Primal-dual hybrid gradient (PDHG)** with $st > \|K\|^2$:

$$u^{k+1} = \arg \min_u \langle u, K^\top p^k \rangle + G(u) + \frac{s}{2} \|u - u^k\|^2,$$

$$p^{k+1} = \arg \min_p - \langle K(2u^{k+1} - u^k), p \rangle + F^*(p) + \frac{t}{2} \|p - p^k\|^2.$$

- Optimality conditions for the updates:

$$0 \in \partial G(u^{k+1}) + K^\top p^k + s(u^{k+1} - u^k),$$

$$0 \in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + t(p^{k+1} - p^k).$$

Scaled primal-dual hybrid gradient

- Recall PDGH:

$$\begin{aligned}0 &\in \partial G(u^{k+1}) + K^\top p^k + s(u^{k+1} - u^k), \\0 &\in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + t(p^{k+1} - p^k).\end{aligned}$$

- Replace s, t by spd matrices $S, T \rightsquigarrow$ Scaled PDHG:

$$\begin{aligned}0 &\in \partial G(u^{k+1}) + K^\top p^k + S(u^{k+1} - u^k), \\0 &\in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + T(p^{k+1} - p^k).\end{aligned}$$

- Scaled PDHG in compact form:

$$0 \in \begin{bmatrix} S & -K^\top \\ -K & T \end{bmatrix} \left(\begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} - \begin{bmatrix} u^k \\ p^k \end{bmatrix} \right) + \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix}.$$



Scaled primal-dual hybrid gradient

- Recall PDGH:

$$\begin{aligned}0 &\in \partial G(u^{k+1}) + K^\top p^k + s(u^{k+1} - u^k), \\0 &\in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + t(p^{k+1} - p^k).\end{aligned}$$

- Replace s, t by spd matrices $S, T \rightsquigarrow$ Scaled PDHG:

$$\begin{aligned}0 &\in \partial G(u^{k+1}) + K^\top p^k + S(u^{k+1} - u^k), \\0 &\in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + T(p^{k+1} - p^k).\end{aligned}$$

- Scaled PDHG in compact form:

$$0 \in \begin{bmatrix} S & -K^\top \\ -K & T \end{bmatrix} \left(\begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} - \begin{bmatrix} u^k \\ p^k \end{bmatrix} \right) + \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix}.$$

- Scaled PDHG is a **generalized proximal iteration**:

$$\boxed{0 \in M(\xi^{k+1} - \xi^k) + R(\xi^{k+1})} \Leftrightarrow \boxed{\xi^{k+1} = (M + R)^{-1} M \xi^k}$$

- Sufficient conditions for convergence:

(1) M is spd matrix; (2) R is maximal monotone operator.



Interpret ADMM as generalized proximal iteration

- Recall ADMM (with reordered updates):

$$v^{k+1} \in \arg \min_v F(v) - \langle p^k, v \rangle + \frac{\tau}{2} \|Ku^k - v\|^2, \quad (1)$$

$$p^{k+1} = p^k + \tau(Ku^k - v^{k+1}), \quad (2)$$

$$u^{k+1} \in \arg \min_u G(u) + \langle p^{k+1}, Ku \rangle + \frac{\tau}{2} \|Ku - v^{k+1}\|^2. \quad (3)$$



Interpret ADMM as generalized proximal iteration



- Recall ADMM (with reordered updates):

$$v^{k+1} \in \arg \min_v F(v) - \langle p^k, v \rangle + \frac{\tau}{2} \|Ku^k - v\|^2, \quad (1)$$

$$p^{k+1} = p^k + \tau(Ku^k - v^{k+1}), \quad (2)$$

$$u^{k+1} \in \arg \min_u G(u) + \langle p^{k+1}, Ku \rangle + \frac{\tau}{2} \|Ku - v^{k+1}\|^2. \quad (3)$$

- ADMM as generalized proximal iteration:

$$(1) \Rightarrow 0 \in \partial F(v^{k+1}) - p^k + \tau(v^{k+1} - Ku^k), \quad (4)$$

$$(3) \Rightarrow 0 \in \partial G(u^{k+1}) + K^\top p^{k+1} + \tau K^\top (Ku^{k+1} - v^{k+1}), \quad (5)$$

$$(2), (4) \Rightarrow p^{k+1} \in \partial F(v^{k+1}) \Leftrightarrow v^{k+1} \in \partial F^*(p^{k+1}), \quad (6)$$

$$(2), (5) \Rightarrow 0 \in \partial G(u^{k+1}) + K^\top (2p^{k+1} - p^k) + \tau K^\top K(u^{k+1} - u^k), \quad (7)$$

$$(2), (6) \Rightarrow 0 \in -Ku^k + \frac{1}{\tau}(p^{k+1} - p^k) + \partial F^*(p^{k+1}), \quad (8)$$

$$(7), (8) \Rightarrow 0 \in \begin{bmatrix} \tau K^\top K & K^\top \\ K & \frac{1}{\tau} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix} + \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix}.$$

Reflection operator

- Given a proper, convex, lsc function $J : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ and $\tau > 0$, we call

$$\text{refl}_{\tau J} = 2 \text{prox}_{\tau J} - I = 2(I + \tau \partial J)^{-1} - I$$

the **reflection operator** on ∂J .

- In a more general definition for refl , ∂J is replaced by a *maximal monotone operator*.
 - We don't formally introduce maximal monotone operators in this lecture.
 - Fact: For any proper, convex, lsc function J , ∂J is a maximal monotone operator.



Reflection operator

- Given a proper, convex, lsc function $J : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ and $\tau > 0$, we call

$$\text{refl}_{\tau J} = 2 \text{prox}_{\tau J} - I = 2(I + \tau \partial J)^{-1} - I$$

the **reflection operator** on ∂J .

- In a more general definition for refl , ∂J is replaced by a *maximal monotone operator*.
 - We don't formally introduce maximal monotone operators in this lecture.
 - Fact: For any proper, convex, lsc function J , ∂J is a maximal monotone operator.
- Fixed points of $\text{refl}_{\tau J}$:

$$\begin{aligned} u &= \text{refl}_{\tau J}(u) \\ \Leftrightarrow u &= 2 \text{prox}_{\tau J}(u) - u \\ \Leftrightarrow u &= \text{prox}_{\tau J}(u) \\ \Leftrightarrow 0 &\in \partial J(u). \end{aligned}$$



Douglas-Rachford- & Peaceman-Rachford splitting

- Consider the monotone inclusion problem:

$$0 \in \partial F(u) + \partial G(u).$$



Douglas-Rachford- & Peaceman-Rachford splitting

- Consider the monotone inclusion problem:

$$0 \in \partial F(u) + \partial G(u).$$

- Douglas-Rachford splitting (DRS):**

$$\begin{cases} u^{k+1} = \text{prox}_{\tau G}(v^k), \\ v^{k+1} = v^k - u^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - v^k). \end{cases} \quad (\text{DRS})$$



Douglas-Rachford- & Peaceman-Rachford splitting

- Consider the monotone inclusion problem:

$$0 \in \partial F(u) + \partial G(u).$$

- Douglas-Rachford splitting (DRS):**

$$\begin{cases} u^{k+1} = \text{prox}_{\tau G}(v^k), \\ v^{k+1} = v^k - u^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - v^k). \end{cases} \quad (\text{DRS})$$

- Peaceman-Rachford splitting (PRS):**

$$\begin{cases} u^{k+1} = \text{prox}_{\tau G}(v^k), \\ v^{k+1} = v^k - 2u^{k+1} + 2 \text{prox}_{\tau F}(2u^{k+1} - v^k). \end{cases} \quad (\text{PRS})$$



Douglas-Rachford- & Peaceman-Rachford splitting

- Consider the monotone inclusion problem:

$$0 \in \partial F(u) + \partial G(u).$$

- Douglas-Rachford splitting (DRS):**

$$\begin{cases} u^{k+1} = \text{prox}_{\tau G}(v^k), \\ v^{k+1} = v^k - u^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - v^k). \end{cases} \quad (\text{DRS})$$

- Peaceman-Rachford splitting (PRS):**

$$\begin{cases} u^{k+1} = \text{prox}_{\tau G}(v^k), \\ v^{k+1} = v^k - 2u^{k+1} + 2 \text{prox}_{\tau F}(2u^{k+1} - v^k). \end{cases} \quad (\text{PRS})$$

- DRS & PRS in compact form:

$$v^{k+1} = \left(\frac{1}{2}I + \frac{1}{2} \text{refl}_{\tau F} \circ \text{refl}_{\tau G} \right) (v^k), \quad (\text{DRS}')$$

$$v^{k+1} = (\text{refl}_{\tau F} \circ \text{refl}_{\tau G}) (v^k). \quad (\text{PRS}')$$



Fixed points of DRS & PRS:

$$v = \text{refl}_{\tau F}(\text{refl}_{\tau G}(v)) = 2 \text{prox}_{\tau F}(\text{refl}_{\tau G}(v)) - \text{refl}_{\tau G}(v)$$

$$\Leftrightarrow \text{prox}_{\tau F}(\text{refl}_{\tau G}(v)) = \text{prox}_{\tau G}(v)$$

$$\Leftrightarrow \text{refl}_{\tau G}(v) \in (I + \tau \partial F)(\text{prox}_{\tau G}(v))$$

$$\Leftrightarrow 2 \text{prox}_{\tau G}(v) - v \in \text{prox}_{\tau G}(v) + \tau \partial F(\text{prox}_{\tau G}(v))$$

$$\Leftrightarrow \text{prox}_{\tau G}(v) - v \in \tau \partial F(\text{prox}_{\tau G}(v))$$

$$\Leftrightarrow u = \text{prox}_{\tau G}(v) \wedge u - v \in \tau \partial F(u)$$

$$\Leftrightarrow v \in u + \tau \partial G(u) \wedge u - v \in \tau \partial F(u)$$

$$\Leftrightarrow 0 \in \partial F(u) + \partial G(u).$$



Interpret DRS as generalized proximal iteration

- Apply DRS to: $\min_u F(u) + G(u)$. \Rightarrow

$$u^{k+1} = \text{prox}_{\tau G}(v^k), \quad (1)$$

$$v^{k+1} = v^k - u^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - v^k). \quad (2)$$



Interpret DRS as generalized proximal iteration



- Apply DRS to: $\min_u F(u) + G(u)$. \Rightarrow

$$u^{k+1} = \text{prox}_{\tau G}(v^k), \quad (1)$$

$$v^{k+1} = v^k - u^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - v^k). \quad (2)$$

- DRS as generalized proximal iteration ($w^k \equiv (u^k - v^k)/\tau$):

$$\begin{aligned} (1) &\Leftrightarrow u^{k+1} = \text{prox}_{\tau G}(u^k - \tau w^k) \Leftrightarrow u^k - \tau w^k \in (I + \tau \partial G)u^{k+1} \\ &\Leftrightarrow 0 \in (u^{k+1} - u^k)/\tau + w^k + \partial G(u^{k+1}), \end{aligned} \quad (3)$$

$$\begin{aligned} (2) &\Leftrightarrow 2u^{k+1} - u^k + \tau w^k = \tau w^{k+1} + \text{prox}_{\tau F}(2u^{k+1} - u^k + \tau w^k) \\ &\Rightarrow \tau w^{k+1} = (I - \text{prox}_{\tau F})(2u^{k+1} - u^k + \tau w^k) \\ &\Leftrightarrow w^{k+1} = \text{prox}_{\frac{1}{\tau} F^*}((2u^{k+1} - u^k)/\tau + w^k) \text{ by Moreau's identity} \\ &\Leftrightarrow (2u^{k+1} - u^k)/\tau + w^k \in \left(I + \frac{1}{\tau} \partial F^*\right)(w^{k+1}) \\ &\Leftrightarrow 0 \in \tau(w^{k+1} - w^k) + \partial F^*(w^{k+1}) - (2u^{k+1} - u^k), \end{aligned} \quad (4)$$

$$(3), (4) \Rightarrow 0 \in \begin{bmatrix} \frac{1}{\tau} I & -I \\ -I & \tau I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ w^{k+1} - w^k \end{bmatrix} + \begin{bmatrix} \partial G & I \\ -I & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ w^{k+1} \end{bmatrix}.$$



- Total-variation denoising:

$$\min_{u:\Omega\rightarrow\mathbb{R}^3} \alpha\|\nabla u\|_{1,1,1} + \frac{1}{2}\|u - f\|^2.$$

- Example codes will be posted on course webpage.
- Acknowledgement to Thomas Möllenhoff.



Convergence Theory

Contraction, nonexpansive operators

Definition

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $\Phi : C \rightarrow \mathbb{E}$. Then Φ is:

- 1 μ -Lipschitz with $\mu \geq 0$ if

$$\forall u, v \in C : \|\Phi(u) - \Phi(v)\| \leq \mu \|u - v\|.$$

- 2 **contractive** if Φ is μ -Lipschitz with $\mu \in [0, 1)$.
- 3 **nonexpansive** if Φ is 1-Lipschitz.



Contraction, nonexpansive operators

Definition

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $\Phi : C \rightarrow \mathbb{E}$. Then Φ is:

- 1 μ -Lipschitz with $\mu \geq 0$ if

$$\forall u, v \in C : \|\Phi(u) - \Phi(v)\| \leq \mu \|u - v\|.$$

- 2 **contractive** if Φ is μ -Lipschitz with $\mu \in [0, 1)$.
- 3 **nonexpansive** if Φ is 1-Lipschitz.

Remark

- 1 If Φ is contractive (mod $\mu \in [0, 1)$), then by **Banach fixed point theorem** the iteration $u^{k+1} = \Phi(u^k)$ converges to the unique fixed point u^* linearly: $\|u^k - u^*\| \leq \mu^k \|u^0 - u^*\|$.
- 2 Unfortunately, most proximal algorithms consist of nonexpansive operators; e.g. proj, prox, and refl. Hence, Banach fixed point theorem does not apply.



Averaged operator

Definition

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $\Phi : C \rightarrow \mathbb{E}$. Then Φ is α -**averaged** with $\alpha \in (0, 1)$ if there exists a nonexpansive operator $\Psi : C \rightarrow \mathbb{E}$ such that

$$\Phi = (1 - \alpha)I + \alpha\Psi.$$

In particular, “ $\frac{1}{2}$ -averaged” is also called **firmly nonexpansive**.



Averaged operator

Definition

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $\Phi : C \rightarrow \mathbb{E}$. Then Φ is α -**averaged** with $\alpha \in (0, 1)$ if there exists a nonexpansive operator $\Psi : C \rightarrow \mathbb{E}$ such that

$$\Phi = (1 - \alpha)I + \alpha\Psi.$$

In particular, “ $\frac{1}{2}$ -averaged” is also called **firmly nonexpansive**.

Proposition

Let C be a nonempty, closed, convex subset of \mathbb{E} , $\Phi : C \rightarrow \mathbb{E}$, and $\alpha \in (0, 1)$. Then the following statements are equivalent:

- 1 Φ is α -averaged.
- 2 $(1 - \frac{1}{\alpha})I + \frac{1}{\alpha}\Phi$ is nonexpansive.
- 3 $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 \leq \|u - v\|^2 - \frac{1-\alpha}{\alpha} \|(I - \Phi)(u) - (I - \Phi)(v)\|^2$.
- 4 $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 + (1 - 2\alpha)\|u - v\|^2 \leq 2(1 - \alpha) \langle u - v, \Phi(u) - \Phi(v) \rangle$.

Proof: on board.



Averaged operator in proximal algorithms

- Recall the generalized proximal iteration:

$$u^{k+1} = \Phi^{(\text{gpi})}(u^k), \quad \Phi^{(\text{gpi})} = (M + R)^{-1}M,$$

for given spd matrix M and monotone operator R .

- One can verify that $\Phi^{(\text{gpi})}$ is firmly nonexpansive under the scaled norm $\|\cdot\|_M = \sqrt{\langle \cdot, M \cdot \rangle}$.



Averaged operator in proximal algorithms

- Recall the generalized proximal iteration:

$$u^{k+1} = \Phi^{(\text{gpi})}(u^k), \quad \Phi^{(\text{gpi})} = (M + R)^{-1}M,$$

for given spd matrix M and monotone operator R .

- One can verify that $\Phi^{(\text{gpi})}$ is firmly nonexpansive under the scaled norm $\|\cdot\|_M = \sqrt{\langle \cdot, M \cdot \rangle}$.
- Recall Douglas-Rachford splitting (in compact form):

$$v^{k+1} = \Phi^{(\text{drs})}(v^k), \quad \Phi^{(\text{drs})} = \frac{1}{2}I + \frac{1}{2}\text{refl}_{\tau F} \circ \text{refl}_{\tau G},$$

for some proper, convex, lsc functions $F, G : \mathbb{E} \rightarrow \overline{\mathbb{R}}$.

- Since $\text{refl}_{\tau F} = 2\text{prox}_{\tau F} - I$ is nonexpansive and $\text{refl}_{\tau G}$ as well, $\Phi^{(\text{drs})}$ is firmly nonexpansive.



Averaged operator in proximal algorithms

- Recall the generalized proximal iteration:

$$u^{k+1} = \Phi^{(\text{gpi})}(u^k), \quad \Phi^{(\text{gpi})} = (M + R)^{-1}M,$$

for given spd matrix M and monotone operator R .

- One can verify that $\Phi^{(\text{gpi})}$ is firmly nonexpansive under the scaled norm $\|\cdot\|_M = \sqrt{\langle \cdot, M \cdot \rangle}$.
- Recall Douglas-Rachford splitting (in compact form):

$$v^{k+1} = \Phi^{(\text{drs})}(v^k), \quad \Phi^{(\text{drs})} = \frac{1}{2}I + \frac{1}{2}\text{refl}_{\tau F} \circ \text{refl}_{\tau G},$$

for some proper, convex, lsc functions $F, G : \mathbb{E} \rightarrow \overline{\mathbb{R}}$.

- Since $\text{refl}_{\tau F} = 2\text{prox}_{\tau F} - I$ is nonexpansive and $\text{refl}_{\tau G}$ as well, $\Phi^{(\text{drs})}$ is firmly nonexpansive.
- Recall forward-backward splitting:

$$u^{k+1} = \Phi^{(\text{fbs})}(u^k), \quad \Phi^{(\text{fbs})} = \text{prox}_{\tau F} \circ (I - \tau \nabla G),$$

where ∇G is μ -Lipschitz and $\tau \in (0, 2/\mu)$.

- As a consequence of the Baillon-Haddad Theorem (next slide), $I - \tau \nabla G$ is an averaged operator. Hence, $\Phi^{(\text{fbs})}$ is a composition of two averaged operators (again averaged).



Averaged operator in proximal algorithms

Theorem (Baillon-Haddad)

Let $J : \mathbb{E} \rightarrow \mathbb{R}$ be a convex, continuously differentiable function. Then ∇J is a nonexpansive operator iff ∇J is firmly nonexpansive.

Proof: on board.



Averaged operator in proximal algorithms

Theorem (Baillon-Haddad)

Let $J : \mathbb{E} \rightarrow \mathbb{R}$ be a convex, continuously differentiable function. Then ∇J is a nonexpansive operator iff ∇J is firmly nonexpansive.

Proof: on board.



Corollary

Assume that $G : \mathbb{E} \rightarrow \mathbb{R}$ is a convex, differentiable function with μ -Lipschitz ∇G , $\tau = 2\alpha/\mu$, and $\alpha \in (0, 1)$, then $I - \tau\nabla G$ is α -averaged.

Proof: Since $\frac{1}{\mu}\nabla G$ is nonexpansive, by the Baillon-Haddad theorem, $\frac{1}{\mu}\nabla G$ is firmly nonexpansive, i.e. $\exists \Psi : \mathbb{E} \rightarrow \mathbb{E}$ nonexpansive s.t. $\frac{1}{\mu}\nabla G = \frac{1}{2}I + \frac{1}{2}\Psi$. Hence,

$$I - \tau\nabla G = \left(1 - \frac{\tau\mu}{2}\right)I - \frac{\tau\mu}{2}\Psi = (1 - \alpha)I + \alpha(-\Psi),$$

i.e. $I - \tau\nabla G$ is α -averaged.

Composition of averaged operators

In forward-backward splitting,

$$\text{prox}_{\tau F} \circ \left(I - \frac{2\alpha}{\mu} \nabla G \right)$$

appears as the composition of a $\frac{1}{2}$ -averaged operator $\text{prox}_{\tau F}$ and an α -averaged operator $I - \frac{2\alpha}{\mu} \nabla G$ with $\alpha \in (0, 1)$.



Composition of averaged operators

In forward-backward splitting,

$$\text{prox}_{\tau F} \circ \left(I - \frac{2\alpha}{\mu} \nabla G \right)$$

appears as the composition of a $\frac{1}{2}$ -averaged operator $\text{prox}_{\tau F}$ and an α -averaged operator $I - \frac{2\alpha}{\mu} \nabla G$ with $\alpha \in (0, 1)$.

Theorem (composition of averaged operators)

Let C be a nonempty, closed, convex subset of \mathbb{E} . For each $i \in \{1, \dots, m\}$, let $\alpha_i \in (0, 1)$ and $\Phi_i : C \rightarrow C$ be an α_i -averaged operator. Then

$$\Phi = \Phi_m \circ \dots \circ \Phi_1$$

is α -averaged with

$$\alpha = \frac{m}{m-1 + \frac{1}{\max_{1 \leq i \leq m} \alpha_i}}.$$

Proof: on board.





Theorem (convex combination of averaged operators)

Let C be a nonempty, closed, convex subset of \mathbb{E} . For each $i \in \{1, \dots, m\}$, let $\alpha_i \in (0, 1)$, $\omega_i \in (0, 1)$ and $\Phi_i : C \rightarrow \mathbb{E}$ be an α_i -averaged operator. If $\sum_{i=1}^m \omega_i = 1$ and $\alpha = \max_{1 \leq i \leq m} \alpha_i$, then

$$\Phi = \sum_{i=1}^m \omega_i \Phi_i$$

is α -averaged.

Proof: as exercise.



Theorem (Krasnoselskii)

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $u^{k+1} = \Phi(u^k)$ for $k = 0, 1, 2, \dots$ where $\Phi : C \rightarrow C$ satisfies:

- 1 Φ is α -averaged for some $\alpha \in (0, 1)$.
- 2 Φ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Φ .

Proof: on board.

Convergence of averaged-operator iterations

Theorem (Krasnoselskii-Mann)

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $u^{k+1} = (1 - \tau^k)u^k + \tau^k\Psi(u^k)$ for $k = 0, 1, 2, \dots$ where $\{\tau^k\} \subset [0, 1]$ s.t.

$$\sum_{k=0}^{\infty} \tau^k(1 - \tau^k) = \infty,$$

and $\Psi : C \rightarrow C$ satisfies:

- 1 Ψ is nonexpansive.
- 2 Ψ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Ψ .

Proof: on board.



Convergence of averaged-operator iterations

Theorem (Krasnoselskii-Mann)

Let C be a nonempty, closed, convex subset of \mathbb{E} , and $u^{k+1} = (1 - \tau^k)u^k + \tau^k\Psi(u^k)$ for $k = 0, 1, 2, \dots$ where $\{\tau^k\} \subset [0, 1]$ s.t.

$$\sum_{k=0}^{\infty} \tau^k(1 - \tau^k) = \infty,$$

and $\Psi : C \rightarrow C$ satisfies:

- 1 Ψ is nonexpansive.
- 2 Ψ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Ψ .

Proof: on board.



Remark

- 1 Condition $\sum_{k=0}^{\infty} \tau^k(1 - \tau^k) = \infty$ is fulfilled if $\{\tau^k\} \subset [\epsilon, 1 - \epsilon]$ for some $\epsilon \in (0, 1/2)$.
- 2 Decay rate of fixed-point residual: $\|u^{k+1} - u^k\| = o(1/\sqrt{k})$.

Convergence in infinite dimensional space

Theorem (Krasnoselskii in infinite dimensions)

Let C be a nonempty, closed, convex subset of a **real Hilbert space** \mathbb{H} , and $u^{k+1} = \Phi(u^k)$ for $k = 0, 1, 2, \dots$ where $\Phi : C \rightarrow C$ satisfies:

- 1 Φ is α -averaged for some $\alpha \in (0, 1)$.
- 2 Φ has at least one fixed point.

Then $\{u^k\}$ converges **weakly** to a fixed point of Φ .



Convergence in infinite dimensional space

Theorem (Krasnoselskii in infinite dimensions)

Let C be a nonempty, closed, convex subset of a **real Hilbert space** \mathbb{H} , and $u^{k+1} = \Phi(u^k)$ for $k = 0, 1, 2, \dots$ where $\Phi : C \rightarrow C$ satisfies:

- 1 Φ is α -averaged for some $\alpha \in (0, 1)$.
- 2 Φ has at least one fixed point.

Then $\{u^k\}$ converges **weakly** to a fixed point of Φ .

Proof: ... $\Rightarrow \|u^{k+1} - \bar{u}\|^2 \leq \|u^0 - \bar{u}\|^2 - \frac{1-\alpha}{\alpha} \sum_{l=0}^k \|(I - \Phi)(u^l)\|^2$
 \Rightarrow (i) $\|u^k - \bar{u}\| \searrow c \geq 0$; (ii) $\sum_{k=0}^{\infty} \|(I - \Phi)(u^k)\|^2 < \infty$.

- (i) $\Rightarrow \{u^k\}$ converges weakly to $u^* \in C$ along a subsequence;
(ii) & “demiclosedness principle” $\Rightarrow u^* - \Phi(u^*) = 0$. $\Rightarrow \dots$ \square

Lemma (demiclosedness principle)

Let C be a nonempty, closed, convex subset of a real Hilbert space \mathbb{H} , and $\Phi : C \rightarrow \mathbb{H}$ be nonexpansive. For any sequence $\{u^k\} \subset C$ s.t. $\{u^k\}$ weakly converges to $u \in C$ and $u^k - \Phi(u^k)$ strongly converges to $v \in \mathbb{H}$, we have $u - \Phi(u) = v$.



Linear convergence under strong monotonicity

- Recall the generalized proximal iteration:

$$0 \in M(u^{k+1} - u^k) + R(u^{k+1}),$$

where M is spd matrix, R is (maximal) monotone operator.



Linear convergence under strong monotonicity

- Recall the generalized proximal iteration:

$$0 \in M(u^{k+1} - u^k) + R(u^{k+1}),$$

where M is spd matrix, R is (maximal) monotone operator.

- Let $u^* = \lim_{k \rightarrow \infty} u^k$, $0 \in R(u^*)$, and $\xi^{k+1} \in R(u^{k+1})$ s.t.

$$\begin{aligned} 0 &= \langle u^{k+1} - u^*, u^{k+1} - u^k \rangle_M + \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \\ &= \frac{1}{2} \|u^{k+1} - u^*\|_M^2 - \frac{1}{2} \|u^k - u^*\|_M^2 + \frac{1}{2} \|u^{k+1} - u^k\|_M^2 \\ &\quad + \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle. \end{aligned}$$



Linear convergence under strong monotonicity

- Recall the generalized proximal iteration:

$$0 \in M(u^{k+1} - u^k) + R(u^{k+1}),$$

where M is spd matrix, R is (maximal) monotone operator.

- Let $u^* = \lim_{k \rightarrow \infty} u^k$, $0 \in R(u^*)$, and $\xi^{k+1} \in R(u^{k+1})$ s.t.

$$\begin{aligned} 0 &= \langle u^{k+1} - u^*, u^{k+1} - u^k \rangle_M + \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \\ &= \frac{1}{2} \|u^{k+1} - u^*\|_M^2 - \frac{1}{2} \|u^k - u^*\|_M^2 + \frac{1}{2} \|u^{k+1} - u^k\|_M^2 \\ &\quad + \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle. \end{aligned}$$

- Previously, R is monotone

$$\Rightarrow \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \geq 0$$

$$\Rightarrow \frac{1}{2} \|u^{k+1} - u^*\|_M^2 \leq \frac{1}{2} \|u^k - u^*\|_M^2 - \frac{1}{2} \|u^{k+1} - u^k\|_M^2$$

$\Rightarrow \dots$



Linear convergence under strong monotonicity

Strongly monotone operator

- ▶ R is said **μ -strongly monotone** if $R - \mu I$ is monotone.
- ▶ For a proper, convex, lsc function J , ∂J is μ -strongly monotone iff J is μ -strongly convex, i.e. $J - \frac{\mu}{2} \|\cdot\|^2$ is convex.



Linear convergence under strong monotonicity

Strongly monotone operator

- ▶ R is said μ -strongly monotone if $R - \mu I$ is monotone.
- ▶ For a proper, convex, lsc function J , ∂J is μ -strongly monotone iff J is μ -strongly convex, i.e. $J - \frac{\mu}{2} \|\cdot\|^2$ is convex.

- Assume R is μ -strongly monotone

$$\begin{aligned} \Rightarrow & \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \geq \mu \|u^{k+1} - u^*\|^2 \\ \Rightarrow & \left(\frac{1}{2} + \frac{\mu}{\lambda_{\max}(M)} \right) \|u^{k+1} - u^*\|_M^2 \\ & \leq \frac{1}{2} \|u^{k+1} - u^*\|_M^2 + \mu \|u^{k+1} - u^*\|^2 \leq \frac{1}{2} \|u^k - u^*\|_M^2 \\ \Rightarrow & \|u^{k+1} - u^*\|_M \leq \sqrt{\frac{1}{1 + 2\mu/\lambda_{\max}(M)}}} \|u^k - u^*\|_M. \end{aligned}$$



Linear convergence under strong monotonicity

Strongly monotone operator

- ▶ R is said μ -strongly monotone if $R - \mu I$ is monotone.
- ▶ For a proper, convex, lsc function J , ∂J is μ -strongly monotone iff J is μ -strongly convex, i.e. $J - \frac{\mu}{2} \|\cdot\|^2$ is convex.

- Assume R is μ -strongly monotone

$$\begin{aligned} \Rightarrow & \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \geq \mu \|u^{k+1} - u^*\|^2 \\ \Rightarrow & \left(\frac{1}{2} + \frac{\mu}{\lambda_{\max}(M)} \right) \|u^{k+1} - u^*\|_M^2 \\ & \leq \frac{1}{2} \|u^{k+1} - u^*\|_M^2 + \mu \|u^{k+1} - u^*\|^2 \leq \frac{1}{2} \|u^k - u^*\|_M^2 \\ \Rightarrow & \|u^{k+1} - u^*\|_M \leq \sqrt{\frac{1}{1 + 2\mu/\lambda_{\max}(M)}}} \|u^k - u^*\|_M. \end{aligned}$$

- Recall in PDHG:

$$R = \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix}.$$

R is μ -strongly monotone $\Leftrightarrow G, F^*$ are μ -strongly convex;
 F^* is μ -strongly convex $\Leftrightarrow F$ has $(1/\mu)$ -Lipschitz gradient.





Acceleration Techniques



- Accelerating gradient step:
 - Second-order method (Newton).
 - Multistep method.
 - Heavy-ball method (Polyak).
 - Extragradient method (Nesterov).
 - Embedding into proximal algorithms.
- Preconditioning proximal algorithms:
 - Preconditioning PDHG algorithm.
 - Diagonal preconditioners (Pock/Chambolle).
 - Application to problems on weighted graphs.

Proximal Newton method

- Consider $\min_u J$ where $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and twice continuously differentiable.
- Classical Newton method:

$$d^k = -[\nabla^2 J(u^k)]^{-1} \nabla J(u^k), \quad u^{k+1} = u^k + d^k.$$

- Minimizing local quadratic model:

$$d^k = \arg \min_d J(u^k) + \langle \nabla J(u^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 J(u^k) d \rangle.$$



Proximal Newton method

- Consider $\min_u J$ where $J : \mathbb{E} \rightarrow \mathbb{R}$ is convex and twice continuously differentiable.
- Classical Newton method:

$$d^k = -[\nabla^2 J(u^k)]^{-1} \nabla J(u^k), \quad u^{k+1} = u^k + d^k.$$

- Minimizing local quadratic model:

$$d^k = \arg \min_d J(u^k) + \langle \nabla J(u^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 J(u^k) d \rangle.$$

- Local quadratic convergence near u^* , where $\nabla J(u^*) = 0$, $\nabla^2 J(u^*)$ is spd:

$$\begin{aligned} \|u^{k+1} - u^*\| &= \|u^k - u^* - [\nabla^2 J(u^k)]^{-1} \nabla J(u^k)\| \\ &\leq \|[\nabla^2 J(u^k)]^{-1}\| \|\nabla^2 J(u^k)(u^k - u^*) - (\nabla J(u^k) - \nabla J(u^*))\| \\ &= O(\|u^k - u^*\|^2). \end{aligned}$$

- Can we use Newton step in proximal gradient method?



Proximal Newton method

$$\min_{u \in \mathbb{E}} F(u) + G(u),$$

where F is convex (possibly non-differentiable), G is convex and twice continuously differentiable.

Proximal Newton method

Initialize $u^0 \in \mathbb{E}$. Iterate with $k = 0, 1, 2, \dots$

- 1 $d^k = \arg \min_d F(u^k + d) + \langle \nabla G(u^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 G(u^k) d \rangle$.
- 2 $u^{k+1} = u^k + d^k$.



Proximal Newton method

$$\min_{u \in \mathbb{E}} F(u) + G(u),$$

where F is convex (possibly non-differentiable), G is convex and twice continuously differentiable.

Proximal Newton method

Initialize $u^0 \in \mathbb{E}$. Iterate with $k = 0, 1, 2, \dots$

- 1 $d^k = \arg \min_d F(u^k + d) + \langle \nabla G(u^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 G(u^k) d \rangle$.
- 2 $u^{k+1} = u^k + d^k$.

Theorem (local convergence of proximal Newton)

The proximal Newton method converges locally quadratically to the (global) minimizer u^* if $\nabla^2 G(u^*)$ is spd.

Proof: on board.



Proximal Newton method

$$\min_{u \in \mathbb{E}} F(u) + G(u),$$

where F is convex (possibly non-differentiable), G is convex and twice continuously differentiable.

Proximal Newton method

Initialize $u^0 \in \mathbb{E}$. Iterate with $k = 0, 1, 2, \dots$

- 1 $d^k = \arg \min_d F(u^k + d) + \langle \nabla G(u^k), d \rangle + \frac{1}{2} \langle d, \nabla^2 G(u^k) d \rangle$.
- 2 $u^{k+1} = u^k + d^k$.

Theorem (local convergence of proximal Newton)

The proximal Newton method converges locally quadratically to the (global) minimizer u^* if $\nabla^2 G(u^*)$ is spd.

Proof: on board.

Remark

- 1 Global convergence via backtracking line search.
- 2 Computation of d^k can be involved even if prox_F is easy.



Heavy-ball method

Minimize J that is convex and twice continuously differentiable.

Heavy-ball method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$. Iterate with $k = 0, 1, 2, \dots$

$$u^{k+1} = u^k - \tau \nabla J(u^k) + \theta(u^k - u^{k-1}),$$

where $\tau, \theta > 0$ are step sizes.



Heavy-ball method

Minimize J that is convex and twice continuously differentiable.

Heavy-ball method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$. Iterate with $k = 0, 1, 2, \dots$

$$u^{k+1} = u^k - \tau \nabla J(u^k) + \theta(u^k - u^{k-1}),$$

where $\tau, \theta > 0$ are step sizes.

- Originated from [Polyak, 1964].
- The term $u^k - u^{k-1}$ is referred to as *momentum*.
- Naturally related to the second-order ODE:

$$\ddot{u} + a\dot{u} + b\nabla J(u) = 0.$$

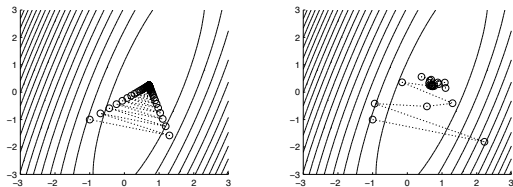


Figure: gradient descent (left) vs. heavy ball (right).



Heavy-ball method

- Quantitative analysis of heavy-ball method:

$$u^{k+1} = u^k - \tau \nabla J(u^k) + \theta(u^k - u^{k-1}).$$

$$\begin{aligned} \begin{bmatrix} u^{k+1} - u^* \\ u^k - u^* \end{bmatrix} &= \begin{bmatrix} u^k + \theta(u^k - u^{k-1}) - u^* - \tau \nabla J(u^k) \\ u^k - u^* \end{bmatrix} \\ &= \begin{bmatrix} u^k + \theta(u^k - u^{k-1}) - u^* - \tau \nabla^2 J(\tilde{u}^k)(u^k - u^*) \\ u^k - u^* \end{bmatrix} \quad (\tilde{u}^k \in [u^k, u^*]) \\ &= \begin{bmatrix} (1 + \theta)I - \tau \nabla^2 J(\tilde{u}^k) & -\theta I \\ I & 0 \end{bmatrix} \begin{bmatrix} u^k - u^* \\ u^{k-1} - u^* \end{bmatrix} =: A^k \begin{bmatrix} u^k - u^* \\ u^{k-1} - u^* \end{bmatrix}. \end{aligned}$$



Heavy-ball method

- Quantitative analysis of heavy-ball method:

$$u^{k+1} = u^k - \tau \nabla J(u^k) + \theta(u^k - u^{k-1}).$$

$$\begin{aligned} \begin{bmatrix} u^{k+1} - u^* \\ u^k - u^* \end{bmatrix} &= \begin{bmatrix} u^k + \theta(u^k - u^{k-1}) - u^* - \tau \nabla J(u^k) \\ u^k - u^* \end{bmatrix} \\ &= \begin{bmatrix} u^k + \theta(u^k - u^{k-1}) - u^* - \tau \nabla^2 J(\tilde{u}^k)(u^k - u^*) \\ u^k - u^* \end{bmatrix} \quad (\tilde{u}^k \in [u^k, u^*]) \\ &= \begin{bmatrix} (1 + \theta)I - \tau \nabla^2 J(\tilde{u}^k) & -\theta I \\ I & 0 \end{bmatrix} \begin{bmatrix} u^k - u^* \\ u^{k-1} - u^* \end{bmatrix} =: A^k \begin{bmatrix} u^k - u^* \\ u^{k-1} - u^* \end{bmatrix}. \end{aligned}$$

- Lemma: Assume $\forall k : \text{sr}(A^k) \leq \rho$, then $\exists \epsilon_k \rightarrow 0^+$ s.t. $\|A^k \dots A^0\| \leq (\rho + \epsilon_k)^k \forall k$.

Theorem

Assume $\forall k : \ell I \preceq \nabla^2 J(\tilde{u}^k) \preceq LI$ for some constants $\ell, L > 0$. If $\theta \geq \max\{|1 - \sqrt{\tau\ell}|, |1 - \sqrt{\tau L}|\}^2$, then $\text{sr}(A^k) = \sqrt{\theta} \forall k$.

Proof: on board.

- $\tau = \frac{4}{(\sqrt{L} + \sqrt{\ell})^2}$, $\theta = \left(\frac{\sqrt{L/\ell} - 1}{\sqrt{L/\ell} + 1}\right)^2 \Rightarrow \text{optimal rate } \rho = \frac{\sqrt{L/\ell} - 1}{\sqrt{L/\ell} + 1}$.



Extragradient method

Minimize J that is convex and continuously differentiable.
Assume ∇J is L -Lipschitz continuous.

Extragradient method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$, $\beta^0 = 1$, $0 < \tau \leq 1/L$.
Iterate with $k = 0, 1, 2, \dots$

- 1 $\beta^{k+1} = (1 + \sqrt{1 + 4(\beta^k)^2})/2$, $\theta^k = (\beta^k - 1)/\beta^{k+1}$.
- 2 $v^k = u^k + \theta^k(u^k - u^{k-1})$.
- 3 $u^{k+1} = v^k - \tau \nabla J(v^k)$.

- Originated from [Nesterov, 1983].
- The gradient is evaluated at the *extrapolated* point v^k .
- The analysis of this scheme is very technical.



Multistep proximal gradient method

We embed multistep acceleration into proximal gradient for:

$$\min_{u \in \mathbb{E}} F(u) + G(u),$$

where F is convex (possibly non-differentiable), G is convex and twice continuously differentiable, and $\ell I \preceq \nabla^2 G(\cdot) \preceq LI$.

Proximal heavy-ball method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$, $\tau = \frac{4}{(\sqrt{L} + \sqrt{\ell})^2}$, $\theta = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}$.
Iterate with $k = 0, 1, 2, \dots$

$$u^{k+1} = \text{prox}_{\tau F}(u^k - \tau \nabla G(u^k) + \theta(u^k - u^{k-1})).$$





Multistep proximal gradient method

We embed multistep acceleration into proximal gradient for:

$$\min_{u \in \mathbb{E}} F(u) + G(u),$$

where F is convex (possibly non-differentiable), G is convex and twice continuously differentiable, and $\ell I \preceq \nabla^2 G(\cdot) \preceq LI$.

Proximal heavy-ball method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$, $\tau = \frac{4}{(\sqrt{L} + \sqrt{\ell})^2}$, $\theta = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}$.
Iterate with $k = 0, 1, 2, \dots$

$$u^{k+1} = \text{prox}_{\tau F}(u^k - \tau \nabla G(u^k) + \theta(u^k - u^{k-1})).$$

Proximal extragradient method

Initialize $u^0 \in \mathbb{E}$, and set $u^{-1} = u^0$, $\beta^0 = 1$, $0 < \tau \leq 1/L$.
Iterate with $k = 0, 1, 2, \dots$

- ① $\beta^{k+1} = (1 + \sqrt{1 + 4(\beta^k)^2})/2$, $\theta^k = (\beta^k - 1)/\beta^{k+1}$.
- ② $v^k = u^k + \theta^k(u^k - u^{k-1})$.
- ③ $u^{k+1} = \text{prox}_{\tau F}(v^k - \tau \nabla G(v^k))$.

Preconditioning iterative linear solvers

- Consider solving the linear system

$$Qu = b \Leftrightarrow \min_u \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle,$$

where $b \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n \times n}$ is spd.

- Define the *condition number* $\kappa_Q = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$, then
 - Convergence rate for steepest descent: $\frac{\kappa_Q - 1}{\kappa_Q + 1}$.
 - Convergence rate for conjugate gradient: $\frac{\sqrt{\kappa_Q} - 1}{\sqrt{\kappa_Q} + 1}$.



Preconditioning iterative linear solvers

- Consider solving the linear system

$$Qu = b \Leftrightarrow \min_u \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle,$$

where $b \in \mathbb{R}^n$, $Q \in \mathbb{R}^{n \times n}$ is spd.

- Define the *condition number* $\kappa_Q = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$, then

- Convergence rate for steepest descent: $\frac{\kappa_Q - 1}{\kappa_Q + 1}$.
- Convergence rate for conjugate gradient: $\frac{\sqrt{\kappa_Q} - 1}{\sqrt{\kappa_Q} + 1}$.
- Preconditioning (or rescaling) with spd $M \in \mathbb{R}^{n \times n}$:

$$\begin{cases} \hat{Q} = M^{-1/2}QM^{-1/2}, \hat{u} = M^{1/2}u, \hat{b} = M^{-1/2}b, \\ \text{Solve: } \min_{\hat{u}} \frac{1}{2} \langle \hat{u}, \hat{Q}\hat{u} \rangle - \langle \hat{b}, \hat{u} \rangle, \text{ where } \kappa_{\hat{Q}} \ll \kappa_Q. \end{cases}$$

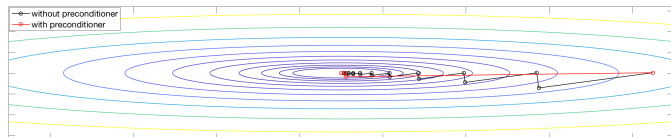


Figure: Steepest descent without precondition. vs. with precondition.



Preconditioning PDHG

- Recall the saddle-point problem:

$$\max_p \min_u \langle p, Ku \rangle + G(u) - F^*(p).$$

- Recall the scaled PDHG:

$$0 \in \partial G(u^{k+1}) + K^\top p^k + \mathbf{S}(u^{k+1} - u^k), \quad \{\text{primal update}\}$$

$$0 \in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + \mathbf{T}(p^{k+1} - p^k). \quad \{\text{dual update}\}$$

- Compact-form PDHG:

$$0 \in \begin{bmatrix} \mathbf{S} & -K^\top \\ -K & \mathbf{T} \end{bmatrix} \left(\begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} - \begin{bmatrix} u^k \\ p^k \end{bmatrix} \right) + \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix}.$$



Preconditioning PDHG

- Recall the saddle-point problem:

$$\max_p \min_u \langle p, Ku \rangle + G(u) - F^*(p).$$

- Recall the scaled PDHG:

$$0 \in \partial G(u^{k+1}) + K^\top p^k + S(u^{k+1} - u^k), \quad \{\text{primal update}\}$$

$$0 \in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + T(p^{k+1} - p^k). \quad \{\text{dual update}\}$$

- Compact-form PDHG:

$$0 \in \begin{bmatrix} S & -K^\top \\ -K & T \end{bmatrix} \left(\begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} - \begin{bmatrix} u^k \\ p^k \end{bmatrix} \right) + \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix}.$$

- Here S is primal preconditioner, T is dual preconditioner:

$$\begin{cases} \hat{u} = S^{1/2}u, \hat{p} = T^{1/2}p, \hat{K} = T^{-1/2}KS^{-1/2}, \\ \hat{G} = G \circ S^{-1/2}, \hat{F} = F \circ T^{1/2}. \\ \text{Solve: } \max_{\hat{p}} \min_{\hat{u}} \langle \hat{p}, \hat{K}\hat{u} \rangle + \hat{G}(\hat{u}) - \hat{F}^*(\hat{p}). \end{cases}$$





- Here S is primal preconditioner, T is dual preconditioner:

$$\begin{cases} \hat{u} = S^{1/2}u, \hat{p} = T^{1/2}p, \hat{K} = T^{-1/2}KS^{-1/2}, \\ \hat{G} = G \circ S^{-1/2}, \hat{F} = F \circ T^{1/2}. \\ \text{Solve: } \max_{\hat{p}} \min_{\hat{u}} \langle \hat{p}, \hat{K}\hat{u} \rangle + \hat{G}(\hat{u}) - \hat{F}^*(\hat{p}). \end{cases}$$

- PDHG in (\hat{u}, \hat{p}) :

$$\begin{aligned} 0 &\in \partial \hat{G}(\hat{u}^{k+1}) + \hat{K}^\top \hat{p}^k + (\hat{u}^{k+1} - \hat{u}^k), \\ 0 &\in \partial \hat{F}^*(\hat{p}^{k+1}) - \hat{K}(2\hat{u}^{k+1} - \hat{u}^k) + (\hat{p}^{k+1} - \hat{p}^k). \end{aligned}$$

- Compact-form PDHG in (\hat{u}, \hat{p}) :

$$0 \in \begin{bmatrix} I & -\hat{K}^\top \\ -\hat{K} & I \end{bmatrix} \left(\begin{bmatrix} \hat{u}^{k+1} \\ \hat{p}^{k+1} \end{bmatrix} - \begin{bmatrix} \hat{u}^k \\ \hat{p}^k \end{bmatrix} \right) + \begin{bmatrix} \partial \hat{G} & \hat{K}^\top \\ -\hat{K} & \partial \hat{F}^* \end{bmatrix} \begin{bmatrix} \hat{u}^{k+1} \\ \hat{p}^{k+1} \end{bmatrix}.$$

Preconditioning PDHG

- Here S is primal preconditioner, T is dual preconditioner:

$$\begin{cases} \hat{u} = S^{1/2}u, \hat{p} = T^{1/2}p, \hat{K} = T^{-1/2}KS^{-1/2}, \\ \hat{G} = G \circ S^{-1/2}, \hat{F} = F \circ T^{1/2}. \\ \text{Solve: } \max_{\hat{p}} \min_{\hat{u}} \langle \hat{p}, \hat{K}\hat{u} \rangle + \hat{G}(\hat{u}) - \hat{F}^*(\hat{p}). \end{cases}$$

- Compact-form PDHG in (\hat{u}, \hat{p}) :

$$0 \in \begin{bmatrix} I & -\hat{K}^\top \\ -\hat{K} & I \end{bmatrix} \left(\begin{bmatrix} \hat{u}^{k+1} \\ \hat{p}^{k+1} \end{bmatrix} - \begin{bmatrix} \hat{u}^k \\ \hat{p}^k \end{bmatrix} \right) + \begin{bmatrix} \partial\hat{G} & \hat{K}^\top \\ -\hat{K} & \partial\hat{F}^* \end{bmatrix} \begin{bmatrix} \hat{u}^{k+1} \\ \hat{p}^{k+1} \end{bmatrix}.$$

Proposition

Assume S, T are spd matrices. Then

$$M_{S,T} = \begin{bmatrix} S & -K^\top \\ -K & T \end{bmatrix} \succ 0 \Leftrightarrow \begin{bmatrix} I & -\hat{K}^\top \\ -\hat{K} & I \end{bmatrix} \succ 0 \\ \Leftrightarrow \|T^{-1/2}KS^{-1/2}\| < 1.$$

Proof: Argue with *Schur complement*.





- Scaled PDHG:

$$\begin{cases} 0 \in \partial G(u^{k+1}) + K^\top p^k + S(u^{k+1} - u^k), \\ 0 \in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + T(p^{k+1} - p^k). \end{cases}$$

- Expectations on S and T :

- 1 S and T shall fulfill $M_{S,T} \succ 0$.
 - 2 (Scaled) resolvents $(S + \partial G)^{-1}$ and $(T + \partial F^*)^{-1}$ are easy to compute.
 - 3 $\hat{K} = T^{-1/2} K S^{-1/2}$ has smaller condition number than K .
 - The theory on why this accelerates convergence is open.
 - Plenty of empirical evidences are observed in experiments.
- Design of S and T that balance (1), (2), (3) is challenging.

Diagonal preconditioner

- Diagonal preconditioners [Pock/Chambolle, 2011]:

$$S = \text{diag}(\{s_j\}), \quad s_j = \sum_i |K_{ij}|^{2-\theta},$$

$$T = \text{diag}(\{t_i\}), \quad t_i = \sum_j |K_{ij}|^\theta,$$

where $\theta \in [0, 2]$.

- $\hat{K} = T^{-1/2}KS^{-1/2}$ suggests that S (resp. T) normalizes columns (resp. rows) of K by row (resp. column) sums.





- Diagonal preconditioners [Pock/Chambolle, 2011]:

$$S = \text{diag}(\{s_j\}), \quad s_j = \sum_i |K_{ij}|^{2-\theta},$$

$$T = \text{diag}(\{t_i\}), \quad t_i = \sum_j |K_{ij}|^\theta,$$

where $\theta \in [0, 2]$.

- $\hat{K} = T^{-1/2}KS^{-1/2}$ suggests that S (resp. T) normalizes columns (resp. rows) of K by row (resp. column) sums.
- Convergence is (almost) justified by the following result:

Proposition

Given matrix K , the diagonal preconditioners S and T above satisfy $M_{S,T} \succeq 0$.

Proof: on board.



- Diagonal preconditioners [Pock/Chambolle, 2011]:

$$S = \text{diag}(\{s_j\}), \quad s_j = \sum_i |K_{ij}|^{2-\theta},$$

$$T = \text{diag}(\{t_i\}), \quad t_i = \sum_j |K_{ij}|^\theta,$$

where $\theta \in [0, 2]$.

- $\hat{K} = T^{-1/2}KS^{-1/2}$ suggests that S (resp. T) normalizes columns (resp. rows) of K by row (resp. column) sums.
- Convergence is (almost) justified by the following result:

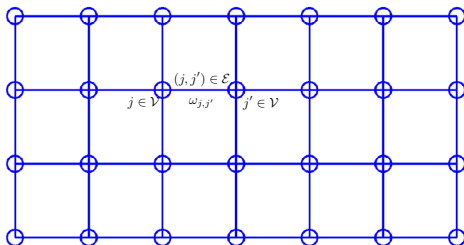
Proposition

Given matrix K , the diagonal preconditioners S and T above satisfy $M_{S,T} \succeq 0$.

Proof: on board.

- Particularly interesting for problems on *weighted graphs*.

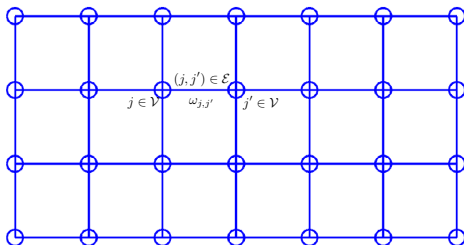
Convex optimization on weighted graphs



- Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$ be a weighted graph, with \mathcal{V} set of vertices, \mathcal{E} set of edges, $\omega : \mathcal{E} \rightarrow \mathbb{R}_+$ weight for edges.
- $\nabla \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{V}|}$ is the *incidence matrix* s.t. for each $(j, j') \in \mathcal{E}$:
 $\nabla_{(j, j'), j} = 1$, $\nabla_{(j, j'), j'} = -1$, $\nabla_{(j, j'), j''} = 0$ whenever $j'' \notin \{j, j'\}$.



Convex optimization on weighted graphs



- Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$ be a weighted graph, with \mathcal{V} set of vertices, \mathcal{E} set of edges, $\omega : \mathcal{E} \rightarrow \mathbb{R}_+$ weight for edges.
- $\nabla \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{V}|}$ is the *incidence matrix* s.t. for each $(j, j') \in \mathcal{E}$:
 $\nabla_{(j, j'), j} = 1$, $\nabla_{(j, j'), j'} = -1$, $\nabla_{(j, j'), j''} = 0$ whenever $j'' \notin \{j, j'\}$.
- Convex optimization on weighted graphs:

$$\min_{u: \mathcal{V} \rightarrow \mathbb{R}} F(Ku) + G(u).$$

where $F : \mathbb{R}^{\mathcal{E}} \rightarrow \mathbb{R}$, $G : \mathbb{R}^{\mathcal{V}} \rightarrow \mathbb{R}$ are convex functions, and $K = \text{diag}(\omega)\nabla$.



Example: image segmentation

- Image segmentation (into $L = 4$ classes):



$$\min_{u: \mathcal{V} \rightarrow \mathbb{R}^L} \underbrace{\sum_{j \in \mathcal{V}} \left(\delta\{u_j \in \Delta^L\} + \langle u_j, f_j \rangle \right)}_{G(u)} + \alpha \underbrace{\sum_{l=1}^L \sum_{(j,j') \in \mathcal{E}} \omega_{j,j'} |u_j^l - u_{j'}^l|}_{F(Ku)}$$

- \mathcal{V} contains image pixels; \mathcal{E} , ω are model-dependent.
- Pointwise constraint: Δ^L is the unit $(L - 1)$ -simplex.
- Unary term: $f : \mathcal{V} \rightarrow \mathbb{R}^L$ is the pixelwise prediction.
- Pairwise term: $\omega_{j,j'}$ models pairwise similarities, e.g.
 - Edges are forged among spatially neighbored pixels; or
 - Use Gaussian similarity measure: $\omega_{j,j'} = \exp\left(-\frac{|j - j'|^2}{\sigma^2}\right)$.



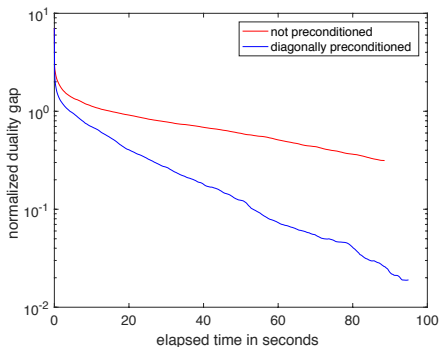
Experimental comparison

On the image segmentation example, we compare PDHG

$$\begin{cases} 0 \in \partial G(u^{k+1}) + K^\top p^k + S(u^{k+1} - u^k), \\ 0 \in \partial F^*(p^{k+1}) - K(2u^{k+1} - u^k) + T(p^{k+1} - p^k). \end{cases}$$

with:

- 1 $S = sI$, $T = tI$, $s = t = \|K\|$.
- 2 $S = \text{diag}(\{s_j\})$, $T = \text{diag}(\{t_i\})$, $s_j = \sum_i |K_{ij}|$, $t_i = \sum_j |K_{ij}|$.



What you should know after this chapter

- Gradient methods:
 - What is a descent method? (descent direction & step size)
 - How to guarantee convergence with properly chosen step sizes? (exact/inexact line search, majorizing surrogate)
- Proximal algorithms:
 - How to derive proximal algorithms (FBS, ADMM, PDHG, DRS) on model problems?
 - How to apply proximal algorithms to specific examples?
 - What is an averaged operator?
 - How to interpret proximal algorithms as averaged-operator iterations?
 - How to prove convergence of averaged-operator iterations? (under general / special assumptions)
- Acceleration techniques:
 - How to accelerate gradient steps in proximal algorithms? (Second-order, multistep)
 - How to precondition PDHG?
 - Some intuitions on why such acceleration techniques work.

