Chapter 4
Summary

Convex Optimization for Computer Vision
SS 2017

Michael Moeller
Visual Scene Analysis
Department of Computer Science
and Electrical Engineering
University of Siegen

Summary Lecture

Convexity

Convexity of $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$: For all $u, v \in \mathbb{R}^n$ and all $\theta \in [0, 1]$ it holds that

$$E(\theta u + (1-\theta)v) \leq \theta E(u) + (1-\theta)E(v) \tag{c}$$

We call E **strictly convex**, if the inequality (c) is strict for all $\theta \in]0, 1[$, and $v \neq u$.

We call E **m-strongly convex** if $G(u) = E(u) - \frac{m}{2}\|u\|^2$ is a convex function.

Existence+uniqueness

The **domain** of E is

$$\text{dom}(E) := \{u \in \mathbb{R}^n \mid E(u) < \infty\}.$$

We call E **proper** if $\text{dom}(E) \neq \emptyset$.

The **epigraph** of E is defined as

$$\text{epi}(E) := \{(u, \alpha) \mid E(u) \leq \alpha\}.$$

A function is called **closed** if its epigraph is a closed set.

If E is closed and there exists a nonempty and bounded sublevelset

$$\{u \in \mathbb{R}^n \mid E(u) \leq \alpha\},$$

then ***E* has a minimizer**.

The subdifferential

The **subdifferential** of a convex function E is

$$\partial E(u) = \{p \in \mathbb{R}^n \mid E(v) - E(u) - \langle p, v - u \rangle \geq 0 \quad \forall v \in \mathbb{R}^n\}$$

If E is differentiable at u then

$$\partial E(u) = \{\nabla E(u)\}.$$

For convex functions, any local minimizer is a global minimizer. The **optimality condition** is

$$\hat{u} \in \arg\min_u E(u) \Leftrightarrow 0 \in \partial E(\hat{u})$$

If E has a minimizer and is strictly convex, then the minimizer of E is unique.

The subdifferential

The **relative interior** of a convex set M is defined as

$$\mathrm{ri}(M) := \{x \in M \mid \forall y \in M,\ \exists \lambda > 1,\ \text{s.t. } \lambda x + (1-\lambda)y \in M\}.$$

If E is a proper convex function and $u \in \mathrm{ri}(\mathrm{dom}(E))$, then $\partial E(u)$ **is non-empty**.

Sum rule – Let E_1, E_2 be convex functions such that $\mathrm{ri}(\mathrm{dom}(E_1)) \cap \mathrm{ri}(\mathrm{dom}(E_2)) \neq \emptyset$, then it holds that

$$\partial(E_1 + E_2)(u) = \{p_1 + p_2 \mid p_1 \in \partial E_1(u),\ p_2 \in \partial E_2(u)\}.$$

Chain rule – If $A \in \mathbb{R}^{m \times n}$, $E : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is convex, and $\mathrm{ri}(\mathrm{dom}(E)) \cap \mathrm{range}(A) \neq \emptyset$, then it holds that

$$\partial(E \circ A)(u) = \{A^T p \mid p \in \partial E(Au)\}.$$

Contractions

Question: When does the **fixed-point iteration**

$$u^{k+1} = G(u^k) \tag{fp}$$

converge?

We call $G : \mathbb{R}^n \to \mathbb{R}^n$ a **contraction** if it is Lipschitz-continuous with constant $L < 1$, i.e. if there exists a $L < 1$ such that

$$\|G(u) - G(v)\|_2 \leq L\|u - v\|_2$$

holds for all $u, \ v \in \mathbb{R}^n$.

If G is a contraction, it has a **unique fixed-point** $\hat{u}$ and (fp) **converges linearly** to $\hat{u}$.

Averaged operators

An operator $H : \mathbb{R}^n \to \mathbb{R}^n$ is called **non-expasive** if it is Lipschitz-continuous with constant 1, i.e. if

$$\|H(u) - H(v)\|_2 \leq \|u - v\|_2$$

holds for all $u,\ v \in \mathbb{R}^n$.

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **averaged** if there exists a non-expansive mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ and a constant $\alpha \in]0, 1[$ such that

$$G = \alpha I + (1 - \alpha)H.$$

If the operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is averaged and has a fixed-point, then the iteration

$$u^{k+1} = G(u^k)$$

converges to a fixed point of G for any starting point $u^0 \in \mathbb{R}^n$.

Averaged operators

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **firmly nonexpansive**, if for all $u, v \in \mathbb{R}^n$ it holds that

$$\|G(u) - G(v)\|_2^2 \leq \langle G(u) - G(v), u - v \rangle.$$

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is **firmly nonexpansive** if and only if G is **averaged with** $\alpha = \frac{1}{2}$.

Compositions of averaged operators are averaged.

Gradient descent

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

is called the **gradient descent iteration**.

An energy $E : \mathbb{R}^n \to \mathbb{R}$ is called **L-smooth** if E is differentiable and ∇E **is L-Lipschitz continuous**.

Baillon-Haddad Theorem: A continuously differentiable convex function $E : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if and only if $\frac{1}{L}\nabla E$ is firmly nonexpansive.

If E is L-smooth then $\frac{1}{L}\nabla E = \frac{1}{2}(I + T)$ for some non-expansive operator T. It follows that

$$G(u) = u - \tau L \frac{1}{L} \nabla E(u) = \left(1 - \frac{L\tau}{2}\right) I + \frac{L\tau}{2}(-T)$$

is **averaged for** $\tau \in]0, 2/L[$.

Gradient projection

$$u^{k+1} = \text{proj}_C(u^k - \tau \nabla E(u^k))$$

is called the **gradient projection iteration** for a nonempty closed convex set C.

The projection onto a nonempty closed convex set is **firmly nonexpansive** and therefore **averaged with $\alpha = 1/2$**.

Since the composition of averaged operators is averaged, we conclude: If E is L-smooth, $\tau \in]0, 2/L[$, and **there exists a minimizer** of E over the set C, then the gradient projection algorithm **converges**.

Proximal gradient

The mapping $\text{prox}_E : \mathbb{R}^n \to \mathbb{R}^n$ defined as

$$\text{prox}_E(v) := \underset{u \in \mathbb{R}^n}{\text{argmin}} \; E(u) + \frac{1}{2}\|u - v\|^2$$

for a closed, proper, convex function $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called the **proximal operator** or **proximal mapping** of E.

The proximal operator prox_E for a closed, proper, convex function E is **firmly nonexpansive** and therefore **averaged with $\alpha = 1/2$**.

Proximal gradient

The iteration

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

is called the **proximal gradient method**.

Let $E(u) = F(u) + G(u)$ **have a minimizer**, let G be L-smooth, and let $\tau \in]0, 2/L[$. Then the proximal gradient method **converges** to a minimizer of E.

The **convergence rates** of the gradient descent, gradient projection, and proximal gradient method are **suboptimal**. They can be **accelerated** by using certain extrapolation schemes.

Accelerated proximal gradient

Pick some starting point $v^0 = u^0$, set $t_0 = 1$, and iterate

1. Compute
$$u^{k+1} = \text{prox}_{\frac{1}{L}G}\left(v^k - \frac{1}{L}\nabla F(v^k)\right)$$
2. Determine
$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},$$
3. Compute the extrapolation of u^{k+1} via
$$v^{k+1} = u^{k+1} + \frac{t_k - 1}{t_{k+1}}(u^{k+1} - u^k)$$

Accelerated gradient projection with line search

Pick $v^0 = u^0$, set $t_0 = 1$, choose $\beta < 1$, $\tau_0 > 0$, and define $Q_\tau(u, v) = F(v) + \langle u - v, \nabla F(v)\rangle + \frac{1}{2\tau}\|u - v\|^2 + G(u)$.

1. Find a suitable step size $\tau_k \leq \tau_{k-1}$ via

$$\tau_k = \tau_{k-1}, \quad u^{k+1} = \text{prox}_{\tau_k G}\left(v^k - \tau_k \nabla F(v^k)\right)$$

while $E(u^{k+1}) > Q_\tau(u^{k+1}, v^k)$

$$\tau_k \leftarrow \beta\tau_k, \quad u^{k+1} \leftarrow \text{prox}_{\tau_k G}\left(v^k - \tau_k \nabla F(v^k)\right)$$

end

2. Determine

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$

3. Compute the extrapolation of u^{k+1} via

$$v^{k+1} = u^{k+1} + \frac{t_k - 1}{t_{k+1}}(u^{k+1} - u^k)$$

Convex conjugation

The **convex conjugate** of a proper function $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is

$$E^*(p) = \sup_u \langle u, p\rangle - E(u).$$

It is always convex and closed.

The **Fenchel-Young inequality** states that

$$E(u) + E^*(p) \geq \langle u, p\rangle,$$

and that equality holds if and only if $p \in \partial E(u)$.

For a proper, closed convex function E it holds that ***E* coincides with its bicocnjugate**,

$$E = E^{**}.$$

For a proper, closed convex function E it holds that

$$p \in \partial E(u) \quad \Leftrightarrow \quad u \in \partial E^*(p).$$

Let $E(u) = G(u) + F(Ku)$ have a minimizer, and let G and F be closed and convex. Let there exist a $u \in \text{ri}(\text{dom}(G))$ such that $Ku \in \text{ri}(\text{dom}(F))$. Then we can reformulate

	$\min_u$	$G(u) + F(Ku)$	**Primal**
$=$	$\min_u \max_q$	$G(u) + \langle q, Ku \rangle - F^*(q)$	**Saddle point**
$=$	$\max_q \min_u$	$G(u) + \langle q, Ku \rangle - F^*(q)$	
$=$	$\max_q$	$-G^*(-K^* q) - F^*(q)$	**Dual**

We are therefore looking for a **saddle point** (u, q) such that

$$-K^T q \in \partial G(u), \quad Ku \in \partial F^*(q).$$

PDHG

The primal-dual point of view motivates the definition of an iterative method to find

$$-K^T q \in \partial G(u), \quad Ku \in \partial F^*(q).$$

The **primal-dual hybrid gradient (PDHG)** method computes

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

or in the **algorithmic-friendly form** of (PDHG)

$$\begin{aligned} p^{k+1} &= \text{prox}_{\sigma F^*}(p^k + \sigma K(2u^k - u^{k-1})), \\ u^{k+1} &= \text{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \end{aligned} \tag{PDHG}$$

Convergence analysis

A set-valued operator T is called **monotone** if

$$\langle p - q, u - v \rangle \geq 0 \qquad \forall u, v, p \in T(u), q \in T(v).$$

The **resolvent** $(I + T)^{-1}$ of a maximally monotone operator is firmly non-expansive, i.e. **averaged with** $\alpha = 1/2$.

Let T be maximally monotone and let there exist a z such that $0 \in T(z)$. Then the **proximal point algorithm**

$$0 \in T(z^{k+1}) + z^{k+1} - z^k$$

converges to a $\tilde{z}$ with $0 \in T(\tilde{z})$.

Convergence of PDHG

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

The operator T is **maximally monotone**. The matrix M is **positive definite** for $\tau\sigma < \frac{1}{\|K\|^2_{S\infty}}$.

Let $M = M^{1/2}M^{1/2}$. Then $M^{-1/2}TM^{-1/2}$ is still maximally montone, and the **PDHG algorithm becomes a proximal point algorithm** in the new variable $z = M^{1/2}(u; p)$.

If the saddle-point problem has a solution and $\tau\sigma < \frac{1}{\|K\|^2_{S\infty}}$, then **PDHG converges**!

PDHG

There are **variants of PDHG** for one of the functions F^* or G being **strongly convex**, or both of them being convex. These variants **converge faster**.

Considering

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

we define the **residuals**

$$r_p^{k+1} = \frac{1}{\sigma}(p^{k+1} - p^k) - K(u^{k+1} - u^k)$$
$$r_d^{k+1} = \frac{1}{\tau}(u^{k+1} - u^k) - K^T(p^{k+1} - p^k)$$

Stopping criteria

Idea: stop the PDHG algorithm if

$$\|r_p^{k+1}\| \leq \sqrt{m}\,\epsilon^{abs} + \|z^{k+1}\| \cdot \epsilon^{rel},$$
$$\|r_d^{k+1}\| \leq \sqrt{n}\,\epsilon^{abs} + \|v^{k+1}\| \cdot \epsilon^{rel}.$$

for $v^{k+1} \in \partial G(u^{k+1})$, $z^{k+1} \in \partial F^*(p^{k+1})$, $u^{k+1} \in \mathbb{R}^n$, $p^{k+1} \in \mathbb{R}^m$.

Residual balancing

Consider the (PDHG) algorithm

$$\begin{aligned} u^{k+1} &= \text{prox}_{\tau G}(u^k - \tau K^* p^k), \\ p^{k+1} &= \text{prox}_{\sigma F^*}(p^k + \sigma K(2u^{k+1} - u^k)), \end{aligned} \tag{PDHG}$$

for two cases

- τ very large, σ very small
  $\rightarrow$ Almost $-K^T p^{k+1} \in \partial G(u^{k+1})$, i.e. $\|r_d\|$ is small.
- τ very small, σ very large
  $\rightarrow$ Almost $K u^{k+1} \in \partial F^*(p^{k+1})$, i.e. $\|r_p\|$ is small.

Idea:

- If $\|r_p\| << \|r_d\|$, increase τ and decrease σ
- If $\|r_p\| >> \|r_d\|$, decrease τ and increase σ

General methods: Make sure the updates decouple, are easy, and M is positive (semi-)definite, e.g.

- **PDHG, overrelaxation on primal**

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

- **PDHG, overrelaxation on dual**

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} I & K^T \\ K & \frac{1}{\sigma} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

- **Primal ADMM**

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \lambda K^T K & K^T \\ K & \frac{1}{\lambda} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

- **Corresponding dual ADMM**

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\lambda} I & -K^T \\ -K & \lambda K K^T \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

Questions to discuss

- Why does the follwing hold

$$\|g\|_{2,1} = \sum_i \|g_i\| = \sum_i \max_{|q_i| \leq 1} \langle q_i, g_i \rangle$$

  Can you repeat the notation of $\|Du\|_{2,1}$?

- PDHG Stopping Criteria

- PDHG Adaptive Stepsize