Optimization Challenge

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

The challenge

Method:	P1a	P2a	P3a	P4a	P1b	P2b	P3b	P4b	$\sum$
GD	18.8	108.9	35.2	10.4	42.8	245	89.2	70.8	**621.1**
PDHG 1	28.7	29.3	43.1	9.0	30.6	43.5	36.2	13.4	**233.8**
ADMM 1	2.3	10.4	7.2	7.2	3.6	121.4	10.0	7.8	**169.7**
PDHG 2	9.1	12.3	10.6	11	12.5	20.1	17.4	69.5	**162.6**
PDHG 3	12.6	15.7	14	7	19.7	28.3	23.1	12.1	**132.6**
PDHG 4	4.4	5.3	5.7	6.1	6.7	9.7	9.8	38.1	**85.8**
PDHG 5	5.7	7.8	6.9	3.7	9.1	13.3	10.7	6.8	**64.0**
ADMM 2	9.6	4.7	29	23	14.4	34.5	36.2	23.1	**174.6**
ADMM 3	3.5	1.3	10.3	8.9	4.2	6.5	10.9	6.8	**52.6**

- GD: Gradient descent reference implementation
- PDHG 1: First Yumin. ($\|m \cdot (u - f)\|^2$ dualized)
- PDHG 2: $\tau \leftarrow 5\tau$, $\sigma \leftarrow 0.2\sigma$.
- ADMM 1: Zhenzhang. pcg(δ, 30)
- PDHG 3: Adaptive stepsizes (5 applications D, D^T)
- PDHG 4: $\|m \cdot (u - f)\|^2$ primal, no adaptation, (2 D, D^T)
- PDHG 5: PDHG 4 + adaptation + some fixes
- ADMM 2: High accuracy by Zhenzhang. pcg($\delta/10^4$, 100)
- ADMM 3: Moderate accuracy. pcg($\delta/10$, 25)

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Zhenzhang wins!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Chapter 7

Summary Lecture

Convex Optimization for Computer Vision

SS 2016

Michael Moeller
Thomas Möllenhoff
Emanuel Laude
Computer Vision Group
Department of Computer Science
TU München

Convex Analysis Basics

- Convex sets and functions
- Domain, epigraph, proper, closed
- Coercivity, existence and uniqueness of minimizers
- Subdifferential + Optimality condition
- Sum rule + Chain rule
- Convex conjugate
- Fenchel-Young inequality, biconjugate
- Primal, dual and saddle-point formulations

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Example: Rudin-Osher-Fatemi (ROF) model

Primal energy:

$$E_{\text{p}}(u) = \frac{\lambda}{2}\|u - f\|^2 + \sum_{i=1}^{2N} |(Du)_i|$$

Saddle-point formulation:

$$E_{\text{pd}}(u,p) = \langle Du, p\rangle + \frac{\lambda}{2}\|u - f\|^2 - \sum_{i=1}^{2N} \iota_{|\cdot|\leq 1}(p_i)$$

Dual energy:

$$E_{\text{d}}(p) = -\frac{1}{2\lambda}\|D^* p\|^2 + \langle Df, p\rangle - \sum_{i=1}^{2N} \iota_{|\cdot|\leq 1}(p_i)$$

How to recover primal from dual solution?

Gradient Descent

Update equation:

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

Properties:

- $\mathrm{dom} E = \mathbb{R}^n$
- For $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, $\tau = 1/L$, energy convergence in $\mathcal{O}(1/k)$
- For $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$, $\tau = 2/(m+L)$, energy and iterate convergence in $\mathcal{O}(c^k)$

Convergence Proof:

- Beginning of lecture: direct convergence proof by “hand”
- At end of semester: fixed point iteration $u^{k+1} = Fu^k$
- $F = (I - \tau \nabla E)$ is contraction if $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$, $\tau < 2/L$
- $F = (I - \tau \nabla E)$ is averaged if $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, $\tau < 2/L$

Gradient Descent: (smoothed) ROF Implementation

- Primal nonsmooth, dual constrained $\Rightarrow$ gradient descent not applicable
- Optimize smooth approximation instead

$$E_{\text{p-smooth}}(u) = \frac{\lambda}{2}\|u-f\|^2 + \sum_{i=1}^{N} h_\varepsilon((Du)_i)$$

- Infimal convolution $f \Box g$, $f(x) = |x|$ and $g(x) = \frac{x^2}{2\varepsilon}$
- Huber penalty $h_\varepsilon(x) = (f \Box g)(x) = \begin{cases} \frac{x^2}{2\varepsilon} & \text{if } |x| \leq \varepsilon, \\ |x| - \frac{\varepsilon}{2} & \text{otherwise.} \end{cases}$

Update:

$$u^{k+1} = u^k - \tau(\lambda(u^k - f) + D^* h'_\varepsilon(Du^k))$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Subgradient Descent

Update equation:

$$u^{k+1} = u^k - \tau_k p^k, p^k \in \partial E(u^k)$$

Properties:

- $\mathrm{dom}E = \mathbb{R}^n$
- Applicable to any Lipschitz-continuous convex energy
- Step size $\tau_k \to 0$ with $\sum_{k=1}^{\infty} \tau_k = \infty$
- Energy convergence in $\mathcal{O}(1/\sqrt{k})$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Subgradient Descent: ROF Implementation

Apply to primal energy

$$E_{\mathrm{p}}(u) = \frac{\lambda}{2}\|u - f\|^2 + \sum_{i=1}^{2N} |(Du)_i|$$

Update:

$$\begin{aligned}
\tau_k &= 1/k, \\
p^{k+1} &= \lambda(u^k - f) + D^* \begin{cases} \operatorname{sgn}((Du)_i), & \text{if} (Du)_i \neq 0 \\ 0 & \text{else} \end{cases}, \\
u^{k+1} &= u^k - \tau_k p^{k+1}.
\end{aligned}$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Proximal Mapping

Most important ingredient for rest of algorithms:

$$\operatorname{prox}_{\tau f}(v) = \underset{u}{\operatorname{argmin}} \; f(u) + \frac{1}{2\tau} \|u - v\|^2$$

- For closed, proper, convex f, $\operatorname{prox}_{\tau f}$ is single-valued and well-defined
- Generalization of projection: $f = \iota_C$, $\operatorname{prox}_{\tau f}(v) = \Pi_C(v)$
- Moreau's identity: $\operatorname{prox}_f(v) + \operatorname{prox}_{f^*}(v) = v$
- Later: resolvent of subdifferential operator

$$\operatorname{prox}_{\tau f}(v) = (I + \tau \partial f)^{-1}(v)$$

- Implicit vs explicit (sub)gradient step

$$(I - \tau \partial f) \qquad \text{vs} \qquad (I + \tau \partial f)^{-1}$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Proximal Gradient / Gradient Projection

Structured problem:

$$E(u) = G(u) + F(u)$$

Update equation:

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

Properties:

- G closed, convex proper, F is L-smooth
- G should be simple, i.e., $\text{prox}_{\tau G}$ easy to evaluate
- Energy convergence $\mathcal{O}(1/k)$ for $\tau \leq 2/L$

Convergence Proof:

- Can write proximal gradient as $u^{k+1} = u^k - \tau \varphi_\tau(u^k)$, $\varphi_\tau(u) = \frac{1}{\tau}(u - \text{prox}_{\tau G}(u - \tau \nabla F(u)))$, similar proof as gradient descent
- Different proof: $(I + \tau \partial G)^{-1}$ averaged, $(I - \tau \nabla F)$ averaged for $\tau < 2/L$
- The composition is averaged, hence proximal gradient is fixed point iteration of averaged operator

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Proximal Gradient / Gradient Projection: ROF Implementation

Dual energy has correct structure for proximal gradient **(why not apply to primal?)**

$$E_{\mathrm{d}}(p) = \underbrace{\frac{1}{2\lambda}\|D^* p\|^2 - \langle Df, p\rangle}_{F(p)} + \underbrace{\sum_{i=1}^{2N} \iota_{|\cdot|\leq 1}(p_i)}_{G(p)}$$

G is simple, F is L-smooth

Update:

$$p^{k+1} = \Pi_{[-1,1]^{2N}}\left(p^k - \tau\left(\frac{1}{\lambda}DD^* p^k - Df\right)\right)$$

Recover:

$$u^* = f - \frac{1}{\lambda}D^* p^*$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

PDHG

Structured problem:

$$E(u) = G(u) + F(Ku)$$

Update equation:

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

Properties:

- G and F closed, proper, convex
- Both can be possibly **non-smooth**!
- G and F simple, i.e., $\text{prox}_{\tau F}$ and $\text{prox}_{\tau G}$ easy to evaluate

Convergence proof:

- One way: rather technical and long calculation
- Later: just a PPA $z^{k+1} = (I + T)^{-1} z^k$ for some T
- PPA is fixed point iteration of averaged operator

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

PDHG: ROF Implementation

Consider saddle-point formulation of ROF:

$$E_{\text{pd}}(u,p) = \langle Du, p\rangle + \underbrace{\frac{\lambda}{2}\|u-f\|^2}_{G(u)} - \underbrace{\sum_{i=1}^{2N} \iota_{|\cdot|\leq 1}(p_i)}_{F^*(p)}$$

Update:

$$\begin{aligned}
p^{k+1} &= \Pi_{[-1,1]^{2N}}(p^k + \sigma D\bar{u}^k),\\
u^{k+1} &= \frac{u^k - \tau D^* p^{k+1} + \tau\lambda f}{1+\tau\lambda},\\
\bar{u}^{k+1} &= 2u^{k+1} - u^k.
\end{aligned}$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Fixed Point and Monotone Operator Perspective

Fixed point iteration for appropriate $F : \mathbb{R}^n \to \mathbb{R}^n$:

$$x^{k+1} = Fx^k.$$

Results:

- F is a contraction $\to$ converges linearly
- F is averaged ($F = \theta T + (1-\theta)I$ for $\theta \in]0,1[$ and T being non-expansive) $\to$ converges

Find mappings $F : \mathbb{R}^n \to \mathbb{R}^n$ whose fixed points are solutions to convex optimization problems.

Important operators:

- The subdifferential ∂E is a maximally monotone operator.
- Resolvent operator $J_A = (I + A)^{-1}$ of maximally monotone operator A is nonexpansive and averaged
- Cayley operator $C_A = 2J_A - I$ of maximally monotone operator A is nonexpansive

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Fixed Point Algorithms

$$x^{k+1} = Fx^k.$$

- Proximal Point:
$$F = (I + A)^{-1},$$
  x^* will meet $0 \in A(x^*)$
- Customized Proximal Point, M s.p.d.:
$$F = (M + A)^{-1} M,$$
  x^* will meet $0 \in A(x^*)$
- Douglas-Rachford splitting (DRS):
$$F = \frac{1}{2} I + \frac{1}{2} C_A C_B,$$
  $u^* = J_B x^*$ will meet $0 \in A(u^*) + B(u^*)$

These methods converge as they are fixed point iterations of averaged operators.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

ADMM

Structured problem:

$$E(u) = G(u) + F(Ku)$$

Augmented Lagrangian

$$L_{\text{aug}}^{\tau}(u, w, p) = G(u) + F(w) + \langle p, Ku - w \rangle + \frac{\tau}{2} \|Ku - w\|^2$$

Update equation:

$$\begin{aligned} u^{k+1} &= \underset{u}{\text{argmin}}\, L_{\text{aug}}^{\tau}(u, w^k, p^k), \\ w^{k+1} &= \underset{w}{\text{argmin}}\, L_{\text{aug}}^{\tau}(u^{k+1}, w, p^k), \\ p^{k+1} &= p^k + \tau \left(Ku^{k+1} - w^{l+1}\right) \end{aligned} \qquad \text{(ADMM)}$$

Convergence:

Special case of DRS applied to dual $A = \partial G^* \circ -K^T$, $B = \partial F^*$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

ADMM: ROF Implementation

Augmented Lagrangian for ROF:

$$L^{\tau}_{\text{aug}}(u, w, p) = \frac{\lambda}{2}\|u - f\|^2 + \|w\|_1 + \langle p, Du - w\rangle + \frac{\tau}{2}\|Du - w\|^2$$

Update equations:

$$u^{k+1} = \operatorname*{argmin}_u \frac{\lambda}{2}\|u - f\|^2 + \langle p^k, Du - w^k\rangle + \frac{\tau}{2}\left\|Du - w^k\right\|^2,$$

$$w^{k+1} = \operatorname*{argmin}_w \|w\|_1 + \langle p^k, Du^k - w\rangle + \frac{\tau}{2}\left\|Du^k - w\right\|^2,$$

$$p^{k+1} = p^k + \tau\,(Du^{k+1} - w^{k+1}).$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Outlook

Rule of thumb:

- Smooth energy $E(u)$: (accelerated) **gradient descent**
- Non-smooth energy $E(u)$: **subgradient descent**
- Sum of smooth and simple (possibly) nonsmooth term $E(u) = F(u) + G(u)$: **proximal gradient method**
- Sum of two simple (possibly) nonsmooth terms, composition with linear operator $E(u) = G(u) + F(Ku)$: **PDHG/DRS/ADMM**

We considered applications in image processing/computer vision but convex optimization is everywhere.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude