Quenstions (and answers!)

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

What is the relation to between ”implicit” gradient descent and proximity operators?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

What is the relation to between ”implicit” gradient descent and proximity operators?

- Consider
$$\partial_t u(t) = -\nabla E(u(t))$$
and think about possible discretizations.
- Compute the optimality conditions for a prox-operator with τE.
- Show the implicit gradient descent is unconditionally stable.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

Why did we look at the gradient map

$$\phi_r(u) = \frac{1}{\tau}(u - \text{prox}_{\tau G}(u - \tau \nabla F(u))))$$

in the convergence proof of the proximal gradient method?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

Why did we look at the gradient map

$$\phi_r(u) = \frac{1}{\tau}(u - \text{prox}_{\tau G}(u - \tau \nabla F(u)))$$

in the convergence proof of the proximal gradient method?

- Remember $u^{k+1} - u^k = -\tau \nabla E$ in the gradient descent case, and $u^{k+1} - u^k = -\tau \phi_r(u^k)$ in the proximal gradient case.
- We were able to carry out the convergence analysis of the proximal gradient method in full analogy to the gradient descent method using ϕ.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

What are different ways to compute

$$\mathrm{prox}_{\alpha\|\cdot - f\|_1}(v)$$

with or without duality and with or without substitution?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

What are different ways to compute

$$\text{prox}_{\alpha\|\cdot - f\|_1}(v)$$

with or without duality and with or without substitution?

- $\text{prox}_{\alpha\|\cdot - f\|_1}(v) = \arg\min_u \frac{1}{2}\|u - v\|^2 + \alpha\|u - f\|_1$
  substitution + shrinkage
- Moreaus identity and projection on convex conjugate.
- Substitutions are always good if they simplify your problem!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

In chapter 5 we derived a fixed point iteration of the form

$$v^{k+1} = C_A C_B v^k$$

for C_A and C_B being the Caley operators of maximally monotone operators A and B. Then we replaced this by

$$v^{k+1} = \left(\frac{1}{2} I + \frac{1}{2} C_A C_B \right) v^k.$$

Why are we allowed to do this? Why does it make sense?

Question

In chapter 5 we derived a fixed point iteration of the form

$$v^{k+1} = C_A C_B v^k$$

for C_A and C_B being the Caley operators of maximally monotone operators A and B. Then we replaced this by

$$v^{k+1} = \left(\frac{1}{2} I + \frac{1}{2} C_A C_B\right) v^k.$$

Why are we allowed to do this? Why does it make sense?

- Fixed point iteration with averaged operator $\rightarrow$ convergence!
- The fixed point remains the same!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

In chapter 5 slide 35 we showed that applying DRS on the primal problem $\min_u G(u) + F(u)$ is equivalent to PDHG. Does it also apply to $\min_u G(u) + F(Ku)$?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

In chapter 5 slide 35 we showed that applying DRS on the primal problem $\min_u G(u) + F(u)$ is equivalent to PDHG. Does it also apply to $\min_u G(u) + F(Ku)$?

- No, consider that DRS applied to our standard minimization problem was the same as ADMM.
- Recall the customized proximal point formulations of ADMM and PDHG, e.g.

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

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

- For $KK^T = c\,I$, $\lambda = \tau$, $\sigma = \frac{1}{c\tau}$ the algorithms are the same. Otherwise they are not.

Question

We applied algorithms like PDHG, ADMM or DRS sometimes on the primal and sometimes on the dual problem. Why? What is the influence? Will a sometimes get a wrong solution if I use one or the other?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

We applied algorithms like PDHG, ADMM or DRS sometimes on the primal and sometimes on the dual problem. Why? What is the influence? Will a sometimes get a wrong solution if I use one or the other?

- Why? $\rightarrow$ Increase the number of options we have.
- Influence? $\rightarrow$ Hard to say in general. Problem specific.
- Wrong solutions? $\rightarrow$ Not if you didn't mess up the derivation! :-)

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

Why may we formulate our problem as

$$\min_{u,d} \max_{p} G(u) + F(d) + \langle Du - d, p \rangle ?$$

There seems to be a strong relation between this Lagrangian form and the primal-dual saddle point form.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

Why may we formulate our problem as

$$\min_{u,d} \max_{p} G(u) + F(d) + \langle Du - d, p \rangle ?$$

There seems to be a strong relation between this Lagrangian form and the primal-dual saddle point form.

- It actually holds that $\delta_{(D\ -I)\cdot=0}(u,d) = (\delta_{(D\ -I)\cdot=0})^{**}(u,d) = \sup_p \langle Du - d, p \rangle$.
- Furthermore, after exchanging $\min_d \max_p = \max_p \min_d$ we arrive at the saddle point form.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

In the script the graph-projection ADMM algorithm first applies a prox operator and then a projection. On the optimization challenge slides there is a graph projection PDHG method which does not even project. Why? What is their relation?

Question

In the script the graph-projection ADMM algorithm first applies a prox operator and then a projection. On the optimization challenge slides there is a graph projection PDHG method which does not even project. Why? What is their relation? Moreover the PDHG projection method does not even have an indicator function, but a Lagrange multiplier instead. Is

$$G(u) + F(d) + \langle Du - d, p \rangle$$

really the right form for calling it a graph-projection?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

In the script the graph-projection ADMM algorithm first applies a prox operator and then a projection. On the optimization challenge slides there is a graph projection PDHG method which does not even project. Why? What is their relation? Moreover the PDHG projection method does not even have an indicator function, but a Lagrange multiplier instead. Is

$$G(u) + F(d) + \langle Du - d, p \rangle$$

really the right form for calling it a graph-projection?

- Our problem is equivalent to

$$\min_{u,d} \underbrace{G(u) + F(d)}_{=\tilde{G}(u,d)} + \underbrace{\delta_{(D\ -I)\cdot=0}(u,d)}_{\tilde{F}(K(u,d))}$$

Applying ADMM yields the graph projection method of the lecture, appying PDHG yields the one of the challenge.

Question

When commenting on the challenge, Michael said that gradient descent on L-smooth, m-strongly convex problems has a linear convergence rate, which is the fastest asymptotic rate we discussed. But isn't quadratic convergence - by which I mean $\mathcal{O}(1/k^2)$ - faster than linear convergence?

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

When commenting on the challenge, Michael said that gradient descent on L-smooth, m-strongly convex problems has a linear convergence rate, which is the fastest asymptotic rate we discussed. But isn't quadratic convergence - by which I mean $\mathcal{O}(1/k^2)$ - faster than linear convergence?

- Linear convergence means $\mathcal{O}(c^k)$ for $c < 1$.
- For every $c < 1$ there exists a K such that $c^k < 1/k^2$ for all $k \geq K$.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

When we stated customized proximal point algorithms we always had some operator of the form

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix}.$$

However, if I consider the optimality condition of the saddle-point formulation $G(u) + \langle Ku, p\rangle - F^*(p)$ I get

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial G & K^T \\ K & -\partial F^* \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix}.$$

Why did we multiply the second part with -1? Why is it more convenient?

Question

When we stated customized proximal point algorithms we always had some operator of the form

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix}.$$

However, if I consider the optimality condition of the saddle-point formulation $G(u) + \langle Ku, p \rangle - F^*(p)$ I get

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \partial G & K^T \\ K & -\partial F^* \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix}.$$

Why did we multiply the second part with -1? Why is it more convenient?

To get a maximally monotone operator!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Question

Can you explain (again) the figure from the Ecksten's dissertation addressing the intuition behind the proximal point algorithm?