Fixed-Point Interpretations of Large-Scale Convex Optimization Algorithms

Pontus Giselsson

Algorithm types and problem dimensions

Problem dimension	Algorithm type
small to medium scale (up to 1'000 variables)	Second-order methods (Newton's method, interior point)
large-scale (up to 100'000 variables)	First-order methods
huge-scale (more than 100'000 variables)	Stochastic, coordinate, parallel asynchronous first-order methods

In data rich fields, problems usually large to huge scale

Large-and huge scale algorithms

Will present unified view of:

- Projected gradient methods
- Proximal gradient methods
- Forward-backward splitting
- Douglas-Rachford splitting
- The alternating direction method of multipliers
- SAGA
- Finito/MISO
- SVRG
- Block-coordinate (proximal) gradient descent
- Block-coordinate consensus optimization
- (Three operator splitting methods)
- (Chambolle-Pock and Primal-dual methods)

First-order method building blocks

- (Sub-)gradients:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

- Projections onto a sets C:

$$\Pi_C(z) = \underset{x}{\operatorname{argmin}}(\|x - z\|_2 : x \in C)$$

- Proximal operators:

$$\operatorname{prox}_{\gamma g}(z) = \underset{x}{\operatorname{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

where $\gamma > 0$ is a parameter.

Prox is generalization of projection

- Introduce the indicator function of a set C

$$\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$$

(this is an extended valued function, i.e., $\iota_C : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$)

- Then

$$\begin{aligned} \Pi_C(z) &= \operatorname*{argmin}_x(\|x - z\|_2 : x \in C) \\ &= \operatorname*{argmin}_x(\tfrac{1}{2}\|x - z\|_2^2 : x \in C) \\ &= \operatorname*{argmin}_x(\tfrac{1}{2}\|x - z\|_2^2 + \iota_C(x)) \\ &= \operatorname{prox}_{\iota_C}(z) \end{aligned}$$

(projection onto C equals prox of indicator function of C)

Prox as resolvent

- The proximal operator satisfies

$$\mathrm{prox}_{\gamma g} = (I + \gamma \partial g)^{-1}$$

  where
  - ∂g is the subdifferential operator
  - $(\cdot)^{-1}$ is the inverse operator
  - $(I + \gamma \partial g)^{-1}$ is called the *resolvent*
- Reason: optimality condition for the prox-computation:

$$\begin{aligned}
& x = \mathrm{prox}_{\gamma g}(z) && \Leftrightarrow \\
& x = \operatorname*{argmin}_x \{g(x) + \tfrac{1}{2\gamma}\|x - z\|^2\} && \Leftrightarrow \\
& 0 \in \gamma \partial g(x) + x - z && \Leftrightarrow \\
& z \in (I + \gamma \partial g)x && \Leftrightarrow \\
& x = (I + \gamma \partial g)^{-1} z
\end{aligned}$$

Problem formulations

- Most algorithms solve problems of the form

$$\text{minimize } f(x) + g(x)$$

  where f, g may be extended-valued: $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$
- Models e.g., constrained problems through

$$\text{minimize } f(x) + \iota_C(x)$$

  where ι_C is indicator function for set C

Consenus formulation

- What if we want to solve problems of the form

$$\text{minimize } \tfrac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- One approach is to use consensus formulation:

$$\text{minimize } \underbrace{\tfrac{1}{n}\sum_{i=1}^{n} f_i(x_i)}_{f(\mathbf{x})} + \underbrace{\iota_C(x_1,\ldots,x_n)}_{g(\mathbf{x})}$$

with individual x_i for each f_i and a consensus constraint

$$C := \{(x_1,\ldots,x_n) : x_1 = \cdots = x_n\}$$

- Problem reduces to two function problem from before
- (Also called divide and concur)

Algorithms – An abstract view

- Most algorithms translate problem to fixed-point problem:

$$\text{find } x^\star \text{ such that } Tx^\star = x^\star$$

  where T is referred to as fixed-point operator (mapping)
- Fixed-points of T have close relationship to solution of problem
- Most algorithms are based on one of the following:
  - The forward-backward map
  - The Douglas-Rachford map

The forward-backward map

- Assume ∇f is Lipschitz and f is convex, g is convex, then (CQ)

$$\begin{aligned} x \in \operatorname{argmin}\{f(x)+g(x)\} &\Leftrightarrow 0 \in \nabla f(x) + \partial g(x) \\ &\Leftrightarrow -\gamma \nabla f(x) \in \gamma \partial g(x) \\ &\Leftrightarrow (I-\gamma\nabla f)x \in (I+\gamma\partial g)x \\ &\Leftrightarrow (I+\gamma\partial g)^{-1}(I-\gamma\nabla f)x \ni x \\ &\Leftrightarrow \operatorname{prox}_{\gamma g}(I-\gamma\nabla f)x = x \end{aligned}$$

- The map $\operatorname{prox}_{\gamma g}(I-\gamma\nabla f)$ is the FB map
- Its fixed-points coincide with solutions to optimization problem
- Reverse order gives backward-forward operator $(I-\gamma\nabla f)\operatorname{prox}_{\gamma g}$:

$$\operatorname{Argmin}\{f(x)+g(x)\} = \operatorname{prox}_{\gamma g}\left(\operatorname{Fix}\left((I-\gamma\nabla f)\operatorname{prox}_{\gamma g}\right)\right)$$

where $\operatorname{Fix}T = \{x : x = Tx\}$

The Douglas-Rachford map

- Let $R_{\gamma f} = 2\text{prox}_{\gamma f} - I$ be the *reflector* or *reflected resolvent*
- It can be shown that

$$\underset{x}{\text{Argmin}}\{f(x) + g(x)\} = \text{prox}_{\gamma g}(\text{Fix} R_{\gamma f} R_{\gamma g})$$

- The composition of reflected resolvents $R_{\gamma f} R_{\gamma g}$ is DR map
- Fixed-point solves optimization problem after prox-step

Why these mappings?

- They have the favorable property of being nonexpansive
- Forward-backward operator
  - Assume f, g convex, ∇f L-Lipschitz, and $\gamma \in (0, \frac{2}{L})$
  - Then $\mathrm{prox}_{\gamma}(I - \gamma \nabla f)$ is nonexpansive
- Douglas-Rachford operator
  - Assume f, g convex and $\gamma \in (0, \infty)$
  - Then $R_{\gamma f} R_{\gamma g}$ is nonexpansive
- Reason, building blocks have similar favorable properties

Nonexpansive

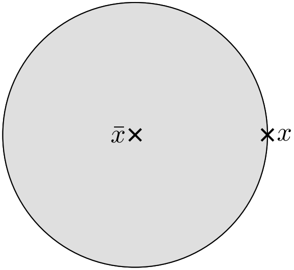
- The operators T are nonexpansive: for all x, y:

$$\|Tx - Ty\| \leq \|x - y\|$$

- Let $y = \bar{x}$ where $\bar{x} = T\bar{x}$ is a fixed-point to T, then

$$\|Tx - \bar{x}\| \leq \|x - \bar{x}\|$$

- 2D graphical representation



Tx in gray area (distance to fixed-point not increased)

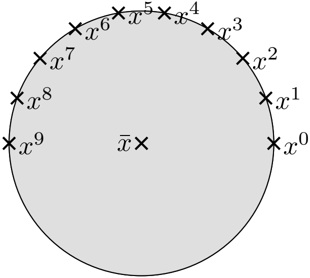
Iterating T

- The iteration

$$x^{k+1} = Tx^k$$

  is not guaranteed to converge to a fixed-point

- Example: T is a rotation



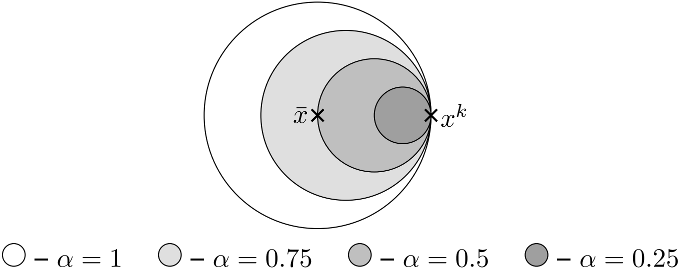
- Why is nonexpansiveness a useful property?

The role of α-averaging

- We consider averaged iteration of the nonexpansive mapping T:

$$x^{k+1} = (1-\alpha)x^k + \alpha T x^k$$

  where $\alpha \in (0,1)$
- 2D example on where x^{k+1} can end up for different α ($\bar{x} \in \mathrm{Fix} T$):



- Distance to fixed-points decreased if $\alpha \in (0,1)$ and $Tx^k \neq x^k$

Property of α-averaged operator

- Let $S = (1-\alpha)I + \alpha T$ and $x^{k+1} = Sx^k$, then it can be shown

$$\|x^{k+1} - z\|^2 \leq \|x^k - z\|^2 - \beta\|x^k - Sx^k\|^2$$

  for all $z \in \mathrm{Fix}S = \mathrm{Fix}T$ and some $\beta > 0$
- $\|x^k - z\|^2$ is Lyapunov function and $\|x^k - Sx^k\|$ gives decrease
- Consequence:
  - $(\|x^k - z\|)_{k\geq 0}$ converges for all $z \in \mathrm{Fix}T$
  - $\|x^k - Sx^k\| = \alpha\|x^k - Tx^k\| \to 0$ as $k \to \infty$

  which is sufficient to show convergence towards a fixed-point

Many different ways to find fixed-point

- Many algorithms for large-scale optimization are of the form:

$$z^{k+1} := (1-\alpha)z^k + \alpha\hat{T}_k z^k = z^k - \alpha(z^k - \hat{T}_k z^k)$$

  where $\alpha \in (0, 1)$ and $\hat{T}_k$ is either:
  - The full operator T (large-scale)
  - A randomized coordinate block update operator of T (huge-scale)
  - A stochastic approximation of T (huge-scale)
- The expected z^{k+1} given z^k for both stochastic methods satisfy:

$$\mathbb{E}_k z^{k+1} = z^k - \alpha(z^k - Tz^k)$$

  they are unbiased stochastic versions of the full operator method

Finding fixed-point of nonexpansive mapping

- The sufficient conditions:
    1. $(\|z - x^k\|)_{k \geq 0}$ converges for all $z \in \mathrm{Fix} T$
    2. $\|Tx^k - x^k\| \to 0$ as $k \to \infty$

  are also necessary conditions
- All orbits from algorithms that find fixed-point satisfy these

How to guarantee conditions – Deterministic case

- Typically, by constructing Lyapunov inequality of the form

$$\|z^{k+1} - z^\star\|_2^2 + \kappa_{k+1} \leq \|z^k - z^\star\|_2^2 + \kappa_k - \gamma_k$$

  where $\gamma_k \geq 0$ and $\kappa_k \geq 0$ satisfy
  - $\gamma_k \to 0$ implies $\|Tx^k - x^k\| \to 0$
  - $\|Tx^k - x^k\| \to 0$ implies $\kappa_k \to 0$
- Easy to verify that necessay and sufficient assumptions hold

How to guarantee conditions – Stochastic case

- Typically by a stochastic Lyapunov inequality of the form

$$\mathbb{E}_k\|z^{k+1} - z^\star\|_2^2 + \kappa_{k+1} \leq \|z^k - z^\star\|_2^2 + \kappa_k - \gamma_k$$

  where $\gamma_k \geq 0$ and $\kappa_k \geq 0$ as before
- The Robbins-Siegmund supermartingale theorem show that conditions for convergence hold a.s.

The only thing left is to find κ_k and γ_k for your algorithm ;)

Thank you

Questions?