

CS794/CO673: Optimization for Data Science

Lec 04: Proximal Gradient

Yaoliang Yu



UNIVERSITY OF
WATERLOO

FACULTY OF MATHEMATICS
**DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE**

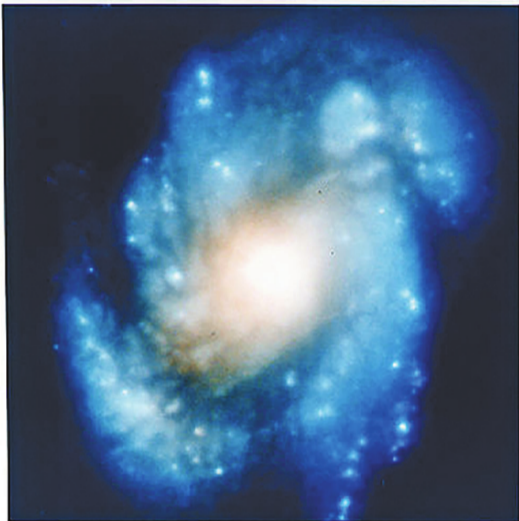
September 23, 2022

Problem

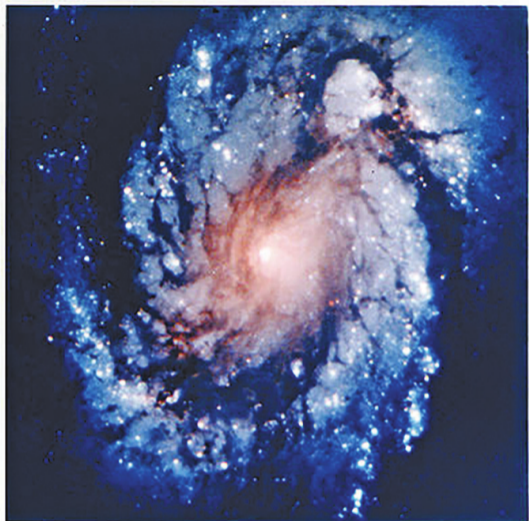
Composite smooth minimization:

$$f_{\star} = \inf_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = \ell(\mathbf{w}) + r(\mathbf{w})$$

- ℓ : smooth and possibly nonconvex
- r : nonsmooth and possibly nonconvex
- The sum $f = \ell + r$ may not be smooth or convex
- Minimizer may or may not be attained
- Maximization is just negation



Wide Field Planetary Camera 1



Wide Field Planetary Camera 2

<https://www.ams.org/journals/notices/202208/noti2534/>

Sparsity

$$\min_{\mathbf{w}} \underbrace{\frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2}_{\ell} + \lambda \cdot \underbrace{\|\mathbf{w}\|_0}_r$$

- Balancing square error with sparsity
- ℓ is convex and L-smooth, r is nonsmooth and nonconvex

$$\min_{\mathbf{w}} \underbrace{\frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2}_{\ell} + \lambda \cdot \underbrace{\|\mathbf{w}\|_1}_r$$

- Convex relaxation: r is now convex but remains nonsmooth (crucial)

Proximal Map and Moreau Envelope

$$P_f^\eta(\mathbf{w}) := \operatorname{argmin}_{\mathbf{z}} \frac{1}{2\eta} \|\mathbf{w} - \mathbf{z}\|_2^2 + f(\mathbf{z})$$

$$M_f^\eta(\mathbf{w}) := \min_{\mathbf{z}} \frac{1}{2\eta} \|\mathbf{w} - \mathbf{z}\|_2^2 + f(\mathbf{z})$$

- $P_f^\eta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ while $M_f^\eta : \mathbb{R}^d \rightarrow \mathbb{R}$
- Under mild conditions, P_f^η is always nonempty and compact
- P_f^η is unique if f is convex while M_f^η is always unique
- M_f^η is a nicer version of f :
 - $M_f^\eta \leq f$, $\inf M_f^\eta = \inf f$, $\operatorname{argmin} M_f^\eta = \operatorname{argmin} f$
 - $M_f^\eta \rightarrow f$ if $\eta \rightarrow 0$, and M_f^η is “smoother” than f

Notation

- We allow functions to take value ∞ (but not $-\infty$ since we are minimizing).
- $\text{dom } f := \{\mathbf{w} : f(\mathbf{w}) < \infty\}$
- Identify a set $C \subseteq \mathbb{R}^d$ with an indicator function

$$\iota_C(\mathbf{w}) = \begin{cases} 0, & \text{if } \mathbf{w} \in C \\ \infty, & \text{if } \mathbf{w} \notin C \end{cases}$$

- Can rewrite constrained problem as a “seemingly” unconstrained one:

$$\inf_{\mathbf{w} \in C} \ell(\mathbf{w}) = \inf_{\mathbf{w} \in \mathbb{R}^d} \ell(\mathbf{w}) + \iota_C(\mathbf{w})$$

- Hence the generality of our composite minimization problem

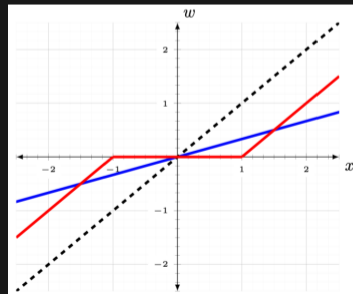
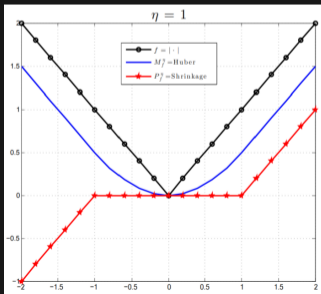
Example: Euclidean projection is a proximal map

$$P_C(\mathbf{w}) = P_{\iota_C}^\eta(\mathbf{w}) \text{ for any } \eta > 0.$$

Example: Soft-shrinkage

Let $r(\mathbf{w}) = \|\mathbf{w}\|_1$, i.e. the sum of absolute values in \mathbf{w} . We have

$$P_r^\eta(\mathbf{w}) = \text{sign}(\mathbf{w}) \odot (|\mathbf{w}| - \eta)_+$$



Algorithm 1: Proximal point algorithm for minimization

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, function $f : \mathbb{R}^d \rightarrow \mathbb{R}$

```
1 for  $t = 0, 1, \dots$  do
2    $\mathbf{w}_{t+1} \leftarrow P_f^{\eta_t}(\mathbf{w}_t)$  //  $\eta_t$  is the step size
```

- $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot \nabla f(\mathbf{w}_{t+1})$, i.e. $\mathbf{w}_t = \mathbf{w}_{t+1} + \eta_t \cdot \nabla f(\mathbf{w}_{t+1})$
- Gradient descent descends from \mathbf{w}_t to \mathbf{w}_{t+1}
- Time flows backwards in PPA: it ascends from \mathbf{w}_{t+1} to \mathbf{w}_t
- Not easy to find \mathbf{w}_{t+1} with such property; but nice theoretical guarantees

Algorithm 2: Proximal gradient algorithm for composite minimization

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, smooth function $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$, $r : \mathbb{R}^d \rightarrow \mathbb{R}$

```
1 for  $t = 0, 1, \dots$  do
2    $\mathbf{z}_t \leftarrow \mathbf{w}_t - \eta_t \cdot \nabla \ell(\mathbf{w}_t)$  // gradient step w.r.t.  $\ell$ 
3    $\mathbf{w}_{t+1} \leftarrow P_r^{\eta_t}(\mathbf{z}_t)$  // proximal step w.r.t.  $r$ 
```

- $r \equiv 0$: reduces to gradient descent
- $\ell \equiv 0$: reduces to proximal point
- $r = \iota_C$: reduces to projected gradient
- Motivation from L -smoothness of ℓ :

$$\begin{aligned} \ell(\mathbf{w}) + r(\mathbf{w}) &\leq \ell(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla \ell(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 + r(\mathbf{w}) \\ &= \frac{1}{2\eta_t} \|\mathbf{w} - (\mathbf{w}_t - \eta_t \cdot \nabla \ell(\mathbf{w}_t))\|_2^2 + r(\mathbf{w}) + \ell(\mathbf{w}_t) - \frac{\eta_t}{2} \|\nabla \ell(\mathbf{w}_t)\|_2^2 \end{aligned}$$

R. E. Bruck (1977). "On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space". *Journal of Mathematical Analysis and Applications*, vol. 61, no. 1, pp. 159–164; M. Fukushima and H. Mine (1981). "A Generalized Proximal Point Algorithm for Certain Non-Convex Minimization Problems". *International Journal of Systems Science*, vol. 12, no. 8, pp. 989–1000.

A Technical Result

The Bregman divergence induced by a (differentiable) convex function f is

$$D_f(\mathbf{z}; \mathbf{w}) := f(\mathbf{z}) - f(\mathbf{w}) - \langle \mathbf{z} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle \geq 0$$

- $D_f(\mathbf{z}; \mathbf{w}) \equiv D_f(\mathbf{w}, \mathbf{z})$ iff $f = \frac{1}{2} \|\cdot\|_2^2$

Theorem: composite optimality

Let ℓ be differentiable convex and r be convex. Then,

$$\mathbf{w}_* \in \operatorname{argmin} \ell + r \iff \forall \mathbf{w}, \ell(\mathbf{w}) + r(\mathbf{w}) \geq \ell(\mathbf{w}_*) + r(\mathbf{w}_*) + D_\ell(\mathbf{w}; \mathbf{w}_*)$$

Corollary: Euclidean projection revisited

Let $\ell(\mathbf{w}) = \frac{1}{2} \|\mathbf{w} - \mathbf{w}_0\|_2^2$ and $r = \iota_C$ for some convex set C .

Theorem: convergence of proximal gradient

Let ℓ be convex and L -smooth and r be convex. Then,

$$f(\mathbf{w}_t) \leq f(\mathbf{w}) + \frac{\|\mathbf{w} - \mathbf{w}_0\|_2^2}{2t\bar{\eta}_t}, \quad \text{where } \bar{\eta}_t := \frac{1}{t} \sum_{s=0}^{t-1} \eta_s.$$

$$\begin{aligned} f(\mathbf{w}_{t+1}) &\leq \ell(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla \ell(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + r(\mathbf{w}_{t+1}) \\ &\leq \ell(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, \nabla \ell(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 + r(\mathbf{w}) - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 \\ &\leq \ell(\mathbf{w}) + r(\mathbf{w}) + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2, \end{aligned}$$

- With $\mathbf{w} = \mathbf{w}_t$ we know $f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t)$
- Multiply η_t and telescope

Discussions

- Where is L -smoothness of ℓ used?
- Where is convexity used?
- What is the condition on the step size η_t ?
 - open-loop: $\sum_t \eta_t \rightarrow \infty, \eta_t \rightarrow 0$
- With $\eta_t = \frac{1}{L}$, obtain the nice bound:

$$f(\mathbf{w}_t) - f_* \leq \frac{L \|\mathbf{w}_0 - \mathbf{w}_*\|_2^2}{2t}$$

- $O(\frac{1}{t})$ rate of convergence, **no dependence on dimension d**
- Amijo's backtracking for the step size?

Example: Elastic net

$$\min_{\mathbf{w}} \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$$

Here we have two choices:

- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$ and $r(\mathbf{w}) = \lambda \|\mathbf{w}\|_1$.
- Set $\ell = \frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2$ and $r(\mathbf{w}) = \lambda \|\mathbf{w}\|_1 + \frac{\gamma}{2} \|\mathbf{w}\|_2^2$.

What are the pros and cons?

$$P_{\lambda \|\cdot\|_1 + \frac{\gamma}{2} \|\cdot\|_2^2}^\eta(\mathbf{w}) = P_{\frac{\gamma}{2} \|\cdot\|_2^2}^\eta \left(P_{\lambda \|\cdot\|_1}^\eta(\mathbf{w}) \right)$$

