

Optimization for Data Science

Lec 05: Accelerated Proximal Gradient

Yaoliang Yu



UNIVERSITY OF
WATERLOO

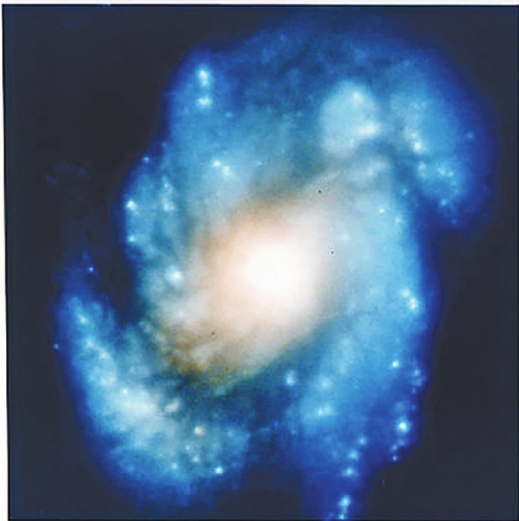
FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

Problem

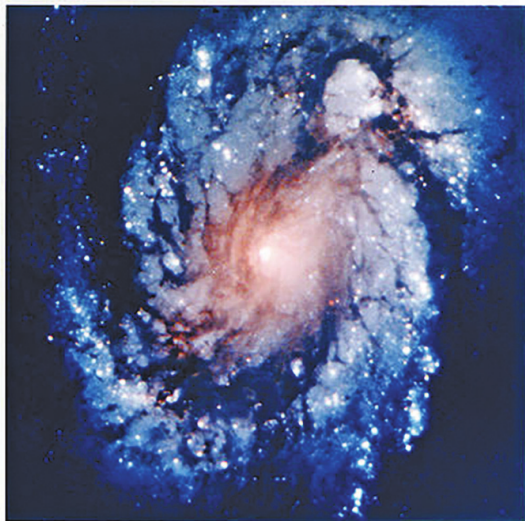
Composite smooth minimization:

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

- 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

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

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

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

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

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    $\mathbf{w}_{t+1} \leftarrow \mathbf{P}_r^{\eta_t}(\mathbf{z}_t)$  // proximal step w.r.t.  $r$ 
```

Input: $\mathbf{w}_0, \mathbf{b} \in \mathbb{R}^d$, $A \in \mathbb{S}_{++}^d \in [\sigma, L]$, $\gamma_0 = 2$, $\kappa = \frac{L}{\sigma}$

```
1  $\mathbf{g}_0 \leftarrow A\mathbf{w}_0 - \mathbf{b}$   
2  $\mathbf{w}_1 \leftarrow \mathbf{w}_0 - \eta_0 \mathbf{g}_0$  //  $\eta_t \equiv \frac{2}{L+\sigma}$   
3 for  $t = 1, 2, \dots$  do  
4    $\mathbf{g}_t \leftarrow A\mathbf{w}_t - \mathbf{b}$  // gradient  
5    $\gamma_t \leftarrow \frac{4(\kappa+1)^2}{4(\kappa+1)^2 - (\kappa-1)^2 \gamma_{t-1}}$  //  $\gamma_t$  is the momentum size  
6    $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \gamma_t \cdot \eta_t \mathbf{g}_t + (\gamma_t - 1) (\mathbf{w}_t - \mathbf{w}_{t-1})$  //  $\eta_t \equiv \frac{2}{L+\sigma}$ 
```

Heavy Ball

$$\mathbf{w}_{t+1} = \underbrace{\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)}_{\text{gradient step}} + \underbrace{\beta_t (\mathbf{w}_t - \mathbf{w}_{t-1})}_{\text{momentum}} = \underbrace{(1 + \beta_t) \mathbf{w}_t - \beta_t \mathbf{w}_{t-1}}_{\text{extrapolation}} - \eta_t \nabla f(\mathbf{w}_t)$$

- Typically $\mathbf{w}_1 = \mathbf{w}_0$ (so that at $t = 1$ we start with the usual gradient step)
- The underlying continuous analogue:

$$\begin{aligned} \mathbf{0} &= [(\mathbf{w}_{t+1} - \mathbf{w}_t) - (\mathbf{w}_t - \mathbf{w}_{t-1})] + (1 - \beta_t)(\mathbf{w}_t - \mathbf{w}_{t-1}) + \eta_t \nabla f(\mathbf{w}_t) \\ &\approx \ddot{\mathbf{w}}(t) + (1 - \beta_t) \dot{\mathbf{w}}(t) + \eta_t \nabla f(\mathbf{w}(t)) \end{aligned}$$

- $\mathbf{w}(t)$ as the position of a heavy ball
- $\dot{\mathbf{w}}(t)$ is the velocity; $\ddot{\mathbf{w}}(t)$ is the acceleration
- f acts as the potential energy
- $\beta_t > 0$: extrapolation vs. $\beta_t < 0$:

Nesterov's Momentum

- Simultaneous gradient update and extrapolation:

$$\mathbf{w}_{t+1} = \underbrace{\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)}_{\text{gradient step}} + \underbrace{\beta_t (\mathbf{w}_t - \mathbf{w}_{t-1})}_{\text{momentum}}$$

- Sequential gradient update and extrapolation:

$$\mathbf{z}_{t+1} = \mathbf{w}_t + \beta_t (\mathbf{w}_t - \mathbf{w}_{t-1})$$

$$\mathbf{w}_{t+1} = \mathbf{z}_{t+1} - \eta_t \nabla f(\mathbf{z}_{t+1})$$

- Continuous analogue:

$$\ddot{\mathbf{w}}(t) + \frac{a}{t} \dot{\mathbf{w}}(t) + \nabla f(\mathbf{w}(t)) = \mathbf{0}$$

Theorem: Optimal rate for Nesterov's momentum

Let $r = 0$ and ℓ be L -smooth convex. Then, with the momentum size choice

$$\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}, \quad \text{where} \quad \gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2},$$

Nesterov algorithm satisfies:

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

where the constant step size $\eta \in (0, 1/L]$ and $\mathbf{w}_* \in \operatorname{argmin} f$ with $f_* = f(\mathbf{w}_*)$.

Back to the Composite Problem

$$\min_{\mathbf{w}} \ell(\mathbf{w}) + r(\mathbf{w})$$

Algorithm 1: Accelerated Proximal Gradient, a.k.a. FISTA

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

```
1 for  $t = 1, 2, \dots$  do
2     choose step size  $\eta_t \leq \eta_{t-1}$  // step size can only decrease
3      $\mathbf{u}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$  // gradient step w.r.t.  $\ell$ 
4      $\mathbf{w}_t = \mathbf{P}_r^{\eta_t}(\mathbf{u}_t) = \operatorname{argmin}_{\mathbf{u}} \frac{1}{2\eta_t} \|\mathbf{u}_t - \mathbf{u}\|_2^2 + r(\mathbf{u})$  // proximal step w.r.t.  $r$ 
5      $\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2}$ 
6      $\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}$  // momentum size
7      $\mathbf{z}_{t+1} = \mathbf{w}_t + \beta_t(\mathbf{w}_t - \mathbf{w}_{t-1})$  // extrapolation
```

A. Beck and M. Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". *SIAM Journal on Imaging Sciences*, vol. 2, no. 1 (2009), pp. 183–202, Y. E. Nesterov. "Gradient Methods for Minimizing Composite Functions". *Mathematical Programming, Series B*, vol. 140 (2013), pp. 125–161.

Discussions

- When $r \equiv 0$, FISTA reduces to the original algorithm of Nesterov.
- When $\gamma_1 = 1$, $\mathbf{w}_0$ does not really play any role: the first step is simply a proximal gradient step
- The smooth function ℓ needs to be defined over the entire space $\mathbb{R}^d$
- The proximal sequence $\mathbf{w}_t$ remains in $\text{dom } r$ by construction
- The momentum choice $\beta_t = \frac{\gamma_t - 1}{\gamma_{t+1}}$ is universal: given any sequence β_t ,

$$\forall t \in [j, i] \text{ s.t. } \beta_t \neq 0, \quad \gamma_{t+1} = \frac{\gamma_t - 1}{\beta_t} = \frac{\gamma_j - 1 - \sum_{m=j}^{t-1} \prod_{k=j}^m \beta_k}{\prod_{k=j}^t \beta_k}.$$

In particular, the choice

$$\gamma_t = \frac{t + a - 2}{a - 1}, \quad \text{or equivalently} \quad \beta_t = \frac{t - 1}{t + a - 1}, \quad a \geq 3,$$

works equally well.

Theorem: Optimal rate for Nesterov's momentum

Suppose $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ is $L^{[1]}$ -smooth and convex, $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ is closed and convex, and $\eta_t \equiv \eta \leq 1/L^{[1]}$. Then, the proximal sequence $\{\mathbf{w}_t\}$ generated by FISTA satisfies: for all $\mathbf{w}$ and $t \geq 1$,

$$f(\mathbf{w}_t) \leq f(\mathbf{w}) + \frac{\|\mathbf{w} - \mathbf{z}_1\|_2^2}{2\eta_t\gamma_t^2} \leq f(\mathbf{w}) + \frac{2\|\mathbf{w} - \mathbf{z}_1\|_2^2}{\eta_t(t+1)^2}.$$

$$\begin{aligned} \frac{1}{2} + \gamma_t &\leq \frac{1 + 2\gamma_t}{2} \leq \gamma_{t+1} \leq \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2} \leq \frac{1 + 1 + 2\gamma_t}{2} = 1 + \gamma_t \implies \\ \frac{t-1}{2} + \gamma_1 &\leq \gamma_t \leq t - 1 + \gamma_1 \end{aligned}$$

- FISTA is not monotonic: it could happen that $f(\mathbf{w}_{t+1}) > f(\mathbf{w}_t)$!

Refinements

If we choose $\mathbf{w} \in \operatorname{argmin} f$, then we can make the following refinements:

- The extrapolation constants need only satisfy

$$\gamma_{t-1}^2 \geq \gamma_t^2 - \gamma_t.$$

In particular, the choice for $\gamma_t = \frac{t+a-2}{a-1}$, $a \geq 3$ works and enjoys the same guarantee (with slightly worse constants).

- We can use Amijo's rule to adaptively choose η_t . However, the condition $\eta_t \leq \eta_{t-1}$ needs to be respected, meaning that each Amijo step should start with the step size from the previous iteration.

Algorithm 2: Monotonic FISTA

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

```
1 for  $t = 1, 2, \dots$  do
2     choose step size  $\eta_t \leq \eta_{t-1}$  // step size can only decrease
3      $\mathbf{u}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$  // gradient step w.r.t.  $\ell$ 
4      $\tilde{\mathbf{w}}_t = \text{P}_r^{\eta_t}(\mathbf{u}_t) = \operatorname{argmin}_{\mathbf{u}} \frac{1}{2\eta_t} \|\mathbf{u}_t - \mathbf{u}\|_2^2 + r(\mathbf{u})$  // proximal step w.r.t.  $r$ 
5     choose  $\mathbf{w}_t$  such that  $f(\mathbf{w}_t) \leq f(\tilde{\mathbf{w}}_t)$  // local improvment
6      $\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2}$ 
7      $\mathbf{z}_{t+1} = \mathbf{w}_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{\gamma_t}{\gamma_{t+1}}(\tilde{\mathbf{w}}_t - \mathbf{w}_t)$  // extrapolation
```

- Can also restart the algorithm: roll back to the previous $\mathbf{w}_{t-1}$
 - does the algorithm simply repeat and get stuck?
 - what to do with γ_t ?

Algorithm 3: Optimized gradient descent

Input: $\mathbf{w}_0 = \mathbf{z}_1, \gamma_1 = 1, \eta_0$

```
1 for  $t = 1, 2, \dots, T$  do
2     choose step size  $\eta_t \leq \eta_{t-1}$  // step size can only decrease
3      $\mathbf{w}_t = \mathbf{z}_t - \eta_t \nabla \ell(\mathbf{z}_t)$  // gradient step w.r.t.  $\ell$ 
4      $\gamma_{t+1} = \frac{1 + \sqrt{1 + 4(1 + \llbracket t = T \rrbracket)\gamma_t^2}}{2}$ 
5      $\mathbf{z}_{t+1} = \mathbf{w}_t + \frac{\gamma_t - 1}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{w}_{t-1}) + \frac{\gamma_t}{\gamma_{t+1}}(\mathbf{w}_t - \mathbf{z}_t)$  // extrapolation
```

$$f(\mathbf{z}_{T+1}) - f_\star \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{2\eta\gamma_{T+1}^2} \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{\eta(T+1)(T+1+\sqrt{2})}, \quad \eta_t \equiv \eta \leq 1/L^{[1]}$$

$$f(\mathbf{w}_t) - f_\star \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{4\eta\gamma_t^2} \leq \frac{\|\mathbf{z}_1 - \mathbf{w}_\star\|_2^2}{\eta(t+1)^2}$$

