Convex Optimisation

Matthew Tam
(University of Melbourne)
matthew.tam@unimelb.edu.au

ACE Network
S1 2021

---

* Assignment 3 online:
matthewktam.github.io/cvxopt

Due: May 13th @ 5pm

Last time

- proximal gradient algorithm

$$\min_{x\in\mathbb{R}^n} f(x)+g(x)$$

convex, diff. with L-Lipschitz grad.

proper, lsc, convex

Get

$$\min_{x\in C} f(x) \quad \text{by setting } g=\iota_C$$

$g=0$

$$\min_{x\in\mathbb{R}^n} f(x)$$

Choose $x_0\in\mathbb{R}^n$ and $\lambda\in(0,\frac{1}{L}]$. Set

$$x_{k+1} := \operatorname{prox}_{\lambda g}(x_k-\lambda\nabla f(x_k)) \quad \forall k\in\mathbb{N}.$$

$g=0$

$$x_{k+1} = x_k - \lambda\nabla f(x_k) \quad \forall k\in\mathbb{N}.$$

We showed that:

$$(f+g)(x_k) - (f+g)(x^*) \le \boxed{O\left(\frac{1}{k}\right)} \quad \forall k \in \mathbb{N}$$

soln

$$O\left(\frac{1}{k^2}\right)$$

3.3 Nestrov acceleration

In this section, we consider the unconstrained minimisation problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with L-Lipschitz continuous gradient.

This is the special case of the problem from Section 3.2 where $g = 0$. Consequently, Theorem 3.2.2 explains how we generate a sequence (x_k) such that

$$f(x_k) - f(x^*) \leq O\left(\frac{1}{k}\right).$$

The natural question that arises is whether this result can be improved.

That is, can the RHS be replaced

with something that decays faster than $O(\frac{1}{k})$, while still using the same function information (i.e. gradient of f).

Given initial points $x_0, y_0 \in \mathbb{R}^n$, Nesterov acceleration iterates according to

$$\begin{cases} y_{k+1} = x_k - \lambda \nabla f(x_k) \\ x_{k+1} = (1-\gamma_k) y_{k+1} + \gamma_k y_k \end{cases} \tag{3.12}$$

where $\lambda \in (0, \frac{1}{L}]$ and (γ_k) is the sequence of positive real numbers given by

$$\alpha_0 = 0, \quad \alpha_k = \frac{1 + \sqrt{1 + 4\alpha_{k-1}^2}}{2}, \quad \gamma_k = \frac{1-\alpha_k}{\alpha_{k+1}} \tag{3.13}$$

Note that 2nd line of (3.12) can be written as

$$x_{k+1} = y_{k+1} + \gamma_k (y_k - y_{k+1})$$

Remark 3.3.1

Note that (α_k) satisfies

$$1 + 4\alpha_{k-1}^2 = (2\alpha_k - 1)^2 = 4\alpha_k^2 - 4\alpha_k + 1$$

$$\iff \alpha_{k-1}^2 = \alpha_k^2 - \alpha_k . \qquad (3.14)$$

and that

$$x_{k+1} = y_{k+1} + \overset{\frac{1-\alpha_k}{\alpha_{k+1}}}{\overset{\|}{\gamma_k}} (y_k - y_{k+1})$$

$$\iff \alpha_{k+1} x_{k+1} = \alpha_{k+1} y_{k+1} + (1 - \alpha_k)(y_k - y_{k+1})$$

$$\iff \alpha_{k+1} x_{k+1} - (\alpha_{k+1} - 1) y_{k+1} = \alpha_k y_{k+1} - (\alpha_k - 1) y_k \qquad (3.15)$$

Lemma 3.3.2

The sequence (α_k) given by (3.13) satisfies $\alpha_k \geq \frac{k+1}{2}$ for all $k \geq 1$.

Proof

The proof is by induction on k. First note $\alpha_0 = 0$, so

$$\alpha_1 = \frac{1+\sqrt{1+4\alpha_0^2}}{2} = \frac{1+1}{2},$$

Thus the result holds for $k=1$.

Suppose the result holds for α_k. Then

$$\alpha_{k+1} = \frac{1+\sqrt{1+4\alpha_k^2}}{2}$$

$$\geq \frac{1+\sqrt{1+\cancel{4}\left(\frac{(k+1)^2}{\cancel{4}}\right)}}{2}$$

$$\geq \frac{1 + \sqrt{0 + (k+1)^2}}{2}$$

$$= \frac{1 + (k+1)}{2}.$$

So the result holds for α_{k+1} which completes the proof. $\boxtimes$

Lemma 3.3.3

Let $f: \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable with L-Lipschitz gradient, and let $\lambda \in (0, \frac{1}{L}]$. For any $x, y \in \mathbb{R}^n$, we have

$$f(x - \lambda \nabla f(x)) - f(y)$$
$$\leq -\frac{\lambda}{2} \|\nabla f(x)\|^2 + \langle \nabla f(x), x - y \rangle.$$

<u>Proof</u>

Using the descent lemma (Prop. 2.2.5), we deduce

$$f(x - \lambda \nabla f(x)) - f(y)$$

$$\leq \langle \nabla f(x), (x - \lambda \nabla f(x)) - y \rangle + \frac{L}{2} \| (\cancel{x} - \lambda \nabla f(x)) - \cancel{x} \|^2$$

$$= \langle \nabla f(x), x - y \rangle - \lambda \|\nabla f(x)\|^2 + \frac{L\lambda^2}{2} \|\nabla f(x)\|^2$$

$$\leq \langle \nabla f(x), x - y \rangle - \lambda \|\nabla f(x)\|^2 + \frac{\lambda}{2} \|\nabla f(x)\|^2$$

$$= \langle \nabla f(x), x - y \rangle - \frac{\lambda}{2} \|\nabla f(x)\|^2.$$

$\blacksquare$

Theorem 3.3.4

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable with L-Lipschitz continuous gradient, and let $x^* \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$. Given initial points $x_0, y_0 \in \mathbb{R}^n$ and $\lambda \in (0, \frac{1}{L}]$, the sequence (y_k) given by (3.12) satisfies

$$f(y_{k+1}) - f(x^*) \leq \frac{2\|x_1 - x^*\|^2}{\lambda (k+1)^2} \quad \forall k \in \mathbb{N}.$$

Proof

Assume WLOG that $f(x^*) = 0$. Setting $x = x_k$ and $y = y_k$ in Lemma 3.3.3 gives

$y_{k+1} = x_k - \lambda \nabla f(x_k)$

$$f(y_{k+1}) - f(y_k) \leq -\frac{\lambda}{2}\|\nabla f(x_k)\|^2 + \langle \nabla f(x_k), x_k - y_k \rangle$$
$$= -\frac{1}{2\lambda}\|y_{k+1} - x_k\|^2 - \frac{1}{\lambda}\langle y_{k+1} - x_k, x_k - y_k \rangle.$$

(3.16).

Setting $x = x_k$ and $y = x^*$ in Lemma 3.3.3 gives

$$f(y_{k+1}) = f(y_{k+1}) - f(x^*)$$
$$\leq -\frac{1}{2\lambda}\|y_{k+1} - x_k\|^2 - \frac{1}{\lambda}\langle y_{k+1} - x_k, x_k - x^* \rangle. \tag{3.17}$$

Using (3.16) and (3.17), we deduce that

$$\alpha_k f(y_{k+1}) - (\alpha_k - 1) f(y_k)$$
$$= (\alpha_k - 1)\left[f(y_{k+1}) - f(y_k)\right] + f(y_{k+1})$$
$$\leq (\alpha_k - 1)\left[-\frac{1}{2\lambda}\|\cancel{y_{k+1}} - x_k\|^2 - \frac{1}{\lambda}\langle y_{k+1} - x_k, x_k - y_k \rangle\right]$$
$$+ \left[-\frac{1}{\cancel{2\lambda}}\|y_{k+1} - x_k\|^2 - \frac{1}{\lambda}\langle y_{k+1} - x_k, x_k - x^* \rangle\right]$$

$$= -\frac{1}{2\lambda}\alpha_k \|y_{k+1}-x_k\|^2$$
$$-\frac{1}{\lambda}\langle y_{k+1}-x_k, (\alpha_k-1)(x_k-y_k)+(x_k-x^*)\rangle$$

$$= -\frac{1}{2\lambda}\alpha_k \|y_{k+1}-x_k\|^2 - \frac{1}{\lambda}\langle y_{k+1}-x_k, \alpha_k x_k - (\alpha_k-1)y_k - x^*\rangle.$$

Multiply both sides by α_k, using the identity $\alpha_{k-1}^2 = \alpha_k^2 - \alpha_k$ from (3.14) and (3.15) gives

$$\boxed{2\langle u,v\rangle = \|u+v\|^2 - \|u\|^2 - \|v\|^2}$$

$$\alpha_k^2 f(y_{k+1}) - \alpha_{k-1}^2 f(y_k)$$

$$\le -\frac{1}{2\lambda}\alpha_k^2 \|y_{k+1}-y_k\|^2$$
$$-\frac{1}{\lambda}\alpha_k\langle y_{k+1}-x_k, \alpha_k x_k - (\alpha_k-1)y_k - x^*\rangle$$

$$= -\frac{1}{2\lambda}\Big(\|\underline{\alpha_k(y_{k+1}-y_k)}\|^2$$
$$+ 2\langle \underline{\alpha_k(y_{k+1}-x_k)}, \alpha_k x_k - (\alpha_k-1)y_k - x^*\rangle\Big)$$

$$= -\frac{1}{2\lambda}\Big(\| \alpha_k (y_{k+1} - x_k) + \alpha_k x_k - (\alpha_k - 1) y_k - x^* \|^2$$

$$- \| \alpha_k x_k - (\alpha_k - 1) y_k - x^* \|^2 \Big)$$

$$= -\frac{1}{2\lambda}\Big(\| \alpha_k y_{k+1} - (\alpha_k - 1) y_k - x^* \|^2$$

$$- \| \alpha_k x_k - (\alpha_k - 1) y_k - x^* \|^2 \Big)$$

$$y_n$$

$$= \frac{1}{2\lambda}\Big(\| \alpha_k x_k - (\alpha_k - 1) y_k - x^* \|^2$$

$$- \| \alpha_{k+1} x_{k+1} - (\alpha_{k+1} - 1) y_{k+1} - x^* \|^2 \Big)$$

(3.15)

we can telescope this inequality to obtain

$$\alpha_k^2 f(y_{k+1})$$

$$= \alpha_k^2 f(y_{k+1}) - \alpha_0^2 f(y_1)$$

$$= \sum_{i=1}^{k} \left[\alpha_i^2 f(y_{i+1}) - \alpha_{i-1}^2 f(y_i) \right]$$

$$\leq \frac{1}{2\lambda}\sum_{i=1}^{k}\left[\begin{aligned}&\|\alpha_i x_i-(\alpha_i-1)y_i-x^*\|^2\\&-\|\alpha_{i+1}x_{i+1}-(\alpha_{i+1}-1)y_{i+1}-x^*\|^2\end{aligned}\right]$$

$$=\frac{1}{2\lambda}\left[\begin{aligned}&\|\overset{=1}{\alpha_1}x_1-\overset{=0}{(\alpha_1-1)}y_1-x^*\|^2\\&-\|\alpha_{k+1}x_{k+1}-(\alpha_{k+1}-1)y_{k+1}-x^*\|^2\end{aligned}\right]$$

$$\leq \frac{1}{2\lambda}\|x_1-x^*\|^2$$

By Lemma 3.3.2, we have $\alpha_k \geq \frac{k+1}{2}$.

Thus, we have

$$f(y_{k+1})-f(x^*) \leq \frac{1}{2\lambda\alpha_k^2}\|x_1-x^*\|^2$$

$$\leq \frac{4}{2\lambda(k+1)^2}\|x_1-x^*\|^2$$

$$=\frac{2}{\lambda(k+1)^2}\|x_1-x^*\|^2.$$

$\boxed{\mathcal{Q}}$

$$\alpha_k^2 f(y_{k+1}) - \alpha_0^2 f(y_1)$$

$$= \left[\alpha_k^2 f(y_{k+1}) - \cancel{\alpha_{k-1}^2 f}(y_k)\right]$$

$$+ \left[\cancel{\alpha_{k-1}^2 f}(y_k) - \cancel{\alpha_{k-2}^2} f(y_{k-2})\right]$$

$$\vdots$$

$$+ \left[\alpha_1^2 \cancel{f}(y_2) - \alpha_0^2 f(y_1)\right]$$

$$= \sum_{i=1}^{k} \left[\alpha_i^2 f(y_{i+1}) - \alpha_{i-1}^2 f(y_i)\right]$$

Remark 3.3.5 (FISTA)

In the literature, the fast iterative soft-thresholding alg. (FISTA) is a version of Nesterov acceleration that solve (3.20) (i.e. $\min_x f(x)+g(x)$). It takes the form:

$$\begin{cases} y_{n+1} = \operatorname{prox}_{\lambda g}(x_n - \lambda \nabla f(x_n)) \\ x_{n+1} = (1-\gamma_k) y_{n+1} + \gamma_k y_n \end{cases}$$

where (γ_n) is given by (3.17) ⊠