Mathematical methods for Image Processing

François Malgouyres

Institut de Mathématiques de Toulouse, France

invitation by
Jidesh P., NITK Surathkal

funding
Global Initiative on Academic Network

Oct. 23–27

Plan

We look for

$$w^* \in \mathrm{Argmin}_{w \in W} \;\; E(w)$$

The function E is usually assumed

- Continuously differentiable : Its gradient is $\nabla E(w) \in W$.
- With a Lipschitz Gradient of parameter $L > 0$:

$$\forall w, w' \in W, \qquad \|\nabla E(w') - \nabla E(w)\| \leq L\|w' - w\|$$

- Proper and convex (can be relaxed)
- Optional hypothesis guaranteeing the convergence of the iterates :
    - Strongly convex (also called elliptic) of parameter $\alpha > 0$

$$\forall w, w' \in W, \qquad \left\langle \nabla E(w') - \nabla E(w), w' - w \right\rangle \geq \alpha \|w' - w\|^2$$

equivalently

$$\forall w, w' \in W, \qquad E(w') \geq E(w) + \left\langle \nabla E(w), w' - w \right\rangle + \frac{\alpha}{2}\|w' - w\|^2$$

Algorithm 1 Gradient Algorithm

Entry: Entry needed for computing E and ∇E

Output: Approximation of a minimizer : w^*

Initialize w

While Not converged **Do**

Compute $d = \nabla E(w)$

Compute a step-size $t \geq 0$

Update : $w \leftarrow w - t\, d$

End while

- **Require**: to calculate and implement a function to compute $\nabla E(w)$ and $E(w - td)$

Algorithm 1 Gradient Algorithm

Entry: Entry needed for computing E and ∇E

Output: Approximation of a minimizer : w^*

Initialize w

While Not converged **Do**

Compute $d = \nabla E(w)$

Compute a step-size $t \geq 0$

Update : $w \leftarrow w - t\, d$

End while

- **Convergence criterion**: Usually no need to be extremely accurate

Algorithm 1 Gradient Algorithm

Entry: Entry needed for computing E and ∇E
Output: Approximation of a minimizer : w^*

Initialize w
While Not converged **Do**
  Compute $d = \nabla E(w)$
  Compute a step-size $t \geq 0$
  Update : $w \leftarrow w - t\, d$
End while

- **Initialization**:
  - Does not affect the quality of the limit point.
  - Affects computational time.

Algorithm 1 Gradient Algorithm

Entry: Entry needed for computing E and ∇E

Output: Approximation of a minimizer : w^*

Initialize w

While Not converged **Do**

Compute $d = \nabla E(w)$

Compute a step-size $t \geq 0$

Update : $w \leftarrow w - t\, d$

End while

- **Step-size**: Many step-size rule exists (constant step-size, steepest descent, Armijo criterion etc)

Theorem (Convergence of the Gradient algorithm)

Let $E : W \longrightarrow \mathbb{R}$ be

- *strongly convex with constant $\alpha > 0$*

$$\forall w, w' \in W, \qquad \langle \nabla E(w') - \nabla E(w), w' - w \rangle \geq \alpha \|w' - w\|^2$$

- *differentiable, with a Lipschitz gradient of constant $L > 0$*

$$\forall w, w' \in W, \qquad \|\nabla E(w') - \nabla E(w)\| \leq L \|w' - w\|$$

Assume, there exists a and b such that the step-size t always satisfies

$$0 < a \leq t \leq b < \frac{2\alpha}{L^2}.$$

Then, the gradient algorithm converges. Its limit-point $w^* = \mathrm{Argmin}_{w \in W} E(w)$ and the sequence w_k is such that

$$\|w^k - w^*\|_2 \leq \beta^k \|w^0 - w^*\|_2,$$

for some $\beta < 1$, where w^k is the k iterate.
For instance, if we take $t = \frac{\alpha}{L^2}$, we have $\beta = \sqrt{1 - \frac{\alpha^2}{L^2}}$.

Comments :

- Convergence of the iterate $w^k \underset{k\to+\infty}{\longrightarrow} w^*$ is stronger than
  - $E(w^k) - E(w^*) \underset{k\to+\infty}{\longrightarrow} 0$
  - $\|\nabla E(w^k)\| \underset{k\to+\infty}{\longrightarrow} 0$
- "Linear convergence rate" : $\ldots \leq \beta^k \|w^0 - w^*\|_2$ is better than many others
  - convergence in $1/k^2$: $\ldots \leq \frac{C}{k^2}$, for some constante $C > 0$.
  - convergence in $1/k$: $\ldots \leq \frac{C}{k}$, for some constante $C > 0$.
- The enemy is the conditioning of E :
  - if $\frac{\alpha}{L} \sim 1 \implies \beta \sim 0$: extremely fast convergence
  - if $\frac{\alpha}{L} \sim 0 \implies \beta \sim 1$: can be very slow

Proof :

$$\begin{aligned} E \text{ strongly convex} \quad &\Longrightarrow \quad E \text{ strictly convex and coercive} \\ &\Longrightarrow \quad E \text{ has a unique global minimizer } w^* \end{aligned}$$

Moreover, $\nabla E(w^*) = 0$

Proof :

Therefore

$$\begin{aligned}
\|w^{k+1} - w^*\|_2^2 &= \|(w^k - t\nabla E(w^k)) - w^*\|_2^2 \\
&= \|w^k - w^* - t(\nabla E(w^k) - \nabla E(w^*))\|_2^2 \\
&= \|w^k - w^*\|_2^2 - 2t\left\langle w^k - w^*, \nabla E(w^k) - \nabla E(w^*)\right\rangle \\
&\quad + t^2\|\nabla E(w^k) - \nabla E(w^*)\|_2^2 \\
&\leq \left(1 - 2\alpha t + L^2 t^2\right)\|w^k - w^*\|_2^2
\end{aligned}$$

We remind

- strongly convex with constant $\alpha > 0$:

$$\forall w, w' \in W, \qquad \left\langle \nabla E(w') - \nabla E(w), w' - w\right\rangle \geq \alpha \|w' - w\|^2$$

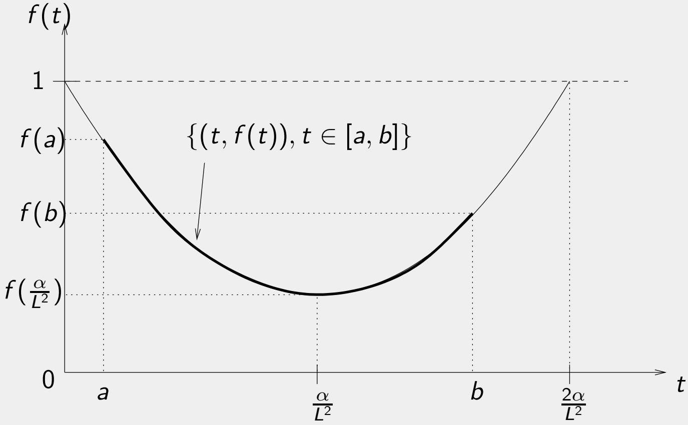
- differentiable, with a Lipschitz gradient of constant $L > 0$:

$$\forall w, w' \in W, \qquad \|\nabla E(w') - \nabla E(w)\| \leq L\|w' - w\|$$

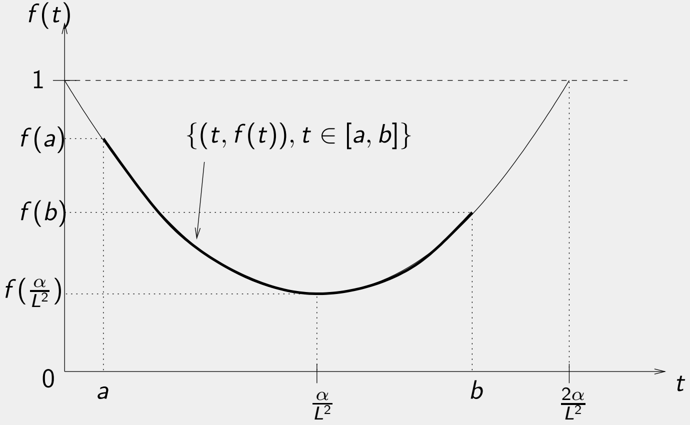
Proof :

$$\|w^{k+1} - w^*\|_2^2 \leq \left(1 - 2\alpha t + L^2 t^2\right)\|w^k - w^*\|_2^2$$

Let $f(t) = 1 - 2\alpha t + L^2 t^2$. We look for t such that $f(t) \leq \beta < 1$.



Proof :



If

$$0 < a \leq t \leq b < \frac{2\alpha}{L^2},$$

then

$$f(t) \leq \max(f(a), f(b)).$$

Proof :

Therefore for $0 < a \leq t \leq b < \frac{2\alpha}{L^2}$

$$\begin{aligned} \|w^{k+1} - w^*\|_2 &\leq \sqrt{f(t)}\|w^k - w^*\|_2 \\ &\leq \beta\|w^k - w^*\|_2, \end{aligned}$$

where $\beta = \sqrt{\max(f(a), f(b))}$.
By induction,we obtain

$$\|w^{k+1} - w^*\|_2 \leq \beta^{k+1}\|w^0 - w^*\|_2.$$

The last state comes from $f(\frac{\alpha}{L^2}) = 1 - \frac{\alpha^2}{L^2}$. □

Theorem (Other convergence result)

Let $E : W \longrightarrow \mathbb{R}$ *be*

- *lower-semicontinuous, convex and coercive*
- *differentiable, with a Lipschitz gradient of constant* $L > 0$

$$\forall w, w' \in W, \qquad \|\nabla E(w') - \nabla E(w)\| \leq L\|w' - w\|$$

if $t < \frac{1}{L}$ *then* $(E(w^k))_{k\in\mathbb{N}}$ *converges. Moreover*

$$0 \leq E(w^k) - E(w^*) \leq \frac{L}{2k}\|w^0 - w^*\|_2.$$

The proof comes later.

To go further

- **Relax the hypotheses on** E: non-differentiable (next lecture), non-convex (See statements based on Kurdyka-Lojasiewicz criterion),
- **Change the algorithm**:
  - Heavy-ball algorithm
  - Accelerated gradient algorithm (Nesterov): Convergence in $o(\frac{1}{k^2})$ (Dossal - Attouch), ease of implementation.
  - Quasi-Newton algorithm (BFGS): Good empirical convergence but requires to approximate the inverse of the Hessian matrix.
- **Adapt to problem structure:**
  - W, data or both are huge : By block algorithms, stochastic gradient methods, online algorithms. (See F. Bach, E. Mouline work.)