Advanced Topics in Machine Learning
Part II
5. Proximal Methods

Luca Baldassarre
with Massimiliano Pontil

Department of Computer Science
University College London
Today’s Plan

- Problem setting
- Convex analysis concepts
- Proximal operators
- $O(1/T)$ algorithm
- $O(1/T^2)$ algorithm
- Empirical comparison
Problem setting

We are interested in following optimization problem

\[
\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w).
\]

We assume that \( r \) is convex and \( f \) is convex and differentiable

Examples:

- **SQUARE LOSS**: \( f(w) = \frac{1}{2} \| Aw - y \|^2 \)
- **LASSO**: \( r(w) = \lambda \| w \|_1 \)
- **TRACE NORM**: \( r(w) = \lambda \| w \|_* \)
We assume that $f$ has Lipschitz continuous gradient:

$$\|\nabla f(w) - \nabla f(v)\| \leq L\|w - v\|.$$ 

**Lemma**

The above assumption is equivalent to

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{L}{2} \|w - v\|^2.$$
Define the linear approximation of $F$ in $v$, w.r.t. $f$

$$\tilde{F}(w; v) := f(v) + \langle \nabla f(v), w - v \rangle + r(w).$$

**Lemma (Sandwich)**

$$F(w) - \frac{L}{2}\|w - v\|^2 \leq \tilde{F}(w; v) \leq F(w).$$

**Proof.**

The left inequality follows from Lemma 1, the right inequality follows from the convexity of $f$, $f(w) \geq f(v) + \langle \nabla f(v), w - v \rangle$. 

**Equivalent version**

$$F(w) \leq \tilde{F}(w; v) + \frac{L}{2}\|w - v\|^2 \leq F(w) + \frac{L}{2}\|w - v\|^2.$$
Sandwich example

\[ F(w) = \frac{1}{2} (aw - 1)^2 + \frac{1}{2} |w| \]

\[ \tilde{F}(w, v) = \frac{1}{2} (aw - 1)^2 + a(aw - 1)(w - v) + \frac{1}{2} |w| \]
If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, its subdifferential at $w$ is defined as

$$\partial f(w) = \{u : f(v) \geq f(w) + \langle u, v - w \rangle, \ \forall v \in \mathbb{R}^d\}$$

- $\partial f$ is a set-valued function
- the elements of $\partial f(w)$ are called the subgradients of $f$ at $w$
- intuition: $u \in \partial f(w)$ if the affine function $f(w) + \langle u, v - w \rangle$ is a global underestimator of $f$

**Theorem**

$$\hat{w} \in \text{argmin}_{w \in \mathbb{R}^d} f(w) \iff 0 \in \partial f(\hat{w})$$
The **proximal operator** of a function $r : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as

$$\text{prox}_r(v) = \arg\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w - v\|^2 + r(w).$$

**Examples:**

- **LASSO:** $r(w) = \lambda \|w\|_1$, $\text{prox}_r(v) = H_\lambda(v)$, which is the component-wise soft-thresholding operator
  $$(H_\lambda(v))_i = \text{sign}(v_i)(|v_i| - \lambda)_+$$

- **$\ell_2$ NORM:** $r(w) = \lambda \|w\|_2$, $\text{prox}_r(v) = \frac{v}{\|v\|_2} (\|v\|_2 - \lambda)_+.$

- **GROUP LASSO:** $r(w) = \sum_{\ell=1}^{K} \|w_{|J_\ell}\|_2$, $$(\text{prox}_r(v))_{|J_\ell} = \frac{v_{|J_\ell}}{\|v_{|J_\ell}\|_2} (\|v_{|J_\ell}\|_2 - \lambda)_+. $$
$O(1/T)$ algorithm

Algorithm

\[
\begin{align*}
\mathbf{w}_0 & \leftarrow 0 \\
\text{for } t = 0, 1, \ldots, T & \text{ do} \\
\mathbf{w}_{t+1} & = \arg\min_{\mathbf{w}} \frac{L}{2} \| \mathbf{w} - \mathbf{w}_t \|^2 + \tilde{F}(\mathbf{w}; \mathbf{w}_t) \\
\text{end for}
\end{align*}
\]

Recall,

\[
\tilde{F}(\mathbf{w}; \mathbf{w}_t) := f(\mathbf{w}_t) + \langle \nabla f(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + r(\mathbf{w})
\]

The term $f(\mathbf{w}_t)$ does not depend on $\mathbf{w}$ and can be discarded.

By completing the square, we also obtain

\[
\frac{L}{2} \| \mathbf{w} - \mathbf{w}_t \|^2 + \langle \nabla f(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + \left( \frac{1}{L} \nabla f(\mathbf{w}_t) \right)^2 = \frac{L}{2} \| \mathbf{w} - (\mathbf{w}_t - \frac{1}{L} \nabla f(\mathbf{w}_t)) \|^2,
\]

where the extra term $\left( \frac{1}{L} \nabla f(\mathbf{w}_t) \right)^2$ does not depend on $\mathbf{w}$.
$O(1/T)$ algorithm (cont’d)

Algorithm 1

\[
\begin{align*}
    &w_0 \leftarrow 0 \\
    &\text{for } t = 0, 1, \ldots, T \text{ do} \\
    &\quad w_{t+1} = \text{prox}_{\frac{r}{L}} \left( w_t - \frac{1}{L} \nabla f(w_t) \right) \\
    &\text{end for}
\end{align*}
\]

Remark. If $r = 0$, we recover Gradient Descent,

\[
w_{t+1} = w_t - \frac{1}{L} \nabla f(w_t)
\]

Theorem (Convergence rate)

Let $w^* \in \arg\min_w F(w)$, then, at iteration $T$, Algorithm 1 yields a solution $w_T$ that satisfies

\[
F(w_T) - F(w^*) \leq \frac{L \|w^* - w_0\|^2}{2T} \tag{1}
\]
**LASSO**

- \( f(w) = \frac{1}{2} \|Aw - y\|^2 \)
- \( r(w) = \lambda \|w\|_1 \)

\[
    w_{t+1} = H_{\frac{\lambda}{L}} \left( w_t - \frac{1}{L} A^\top (A w_t - y) \right)
\]

**GROUP LASSO**

- \( f(w) = \frac{1}{2} \|Aw - y\|^2 \)
- \( r(w) = \lambda \sum_{\ell=1}^K \|w_{|J_\ell}\|_2 \)

\[
    v = w_t - \frac{1}{L} A^\top (A w_t - y)
\]

\[
    (w_{t+1})_{|J_\ell} = \frac{v_{|J_\ell}}{\|v_{|J_\ell}\|} \left( \|v_{|J_\ell}\|_2 - \frac{\lambda}{L} \right)_+
\]
\( O(1/ T) \) algorithm - convergence rate proof

**Sandwich.** \( F(w) - \frac{L}{2} \|w - v\|^2 \leq \tilde{F}(w; v) \leq F(w) \)

**Lemma: 3-point property.**
If \( \hat{w} = \text{argmin}_{w \in \mathbb{R}^d} \frac{1}{2} \|w - w_0\|^2 + \phi(w) \), then, for any \( w \in \mathbb{R}^d \)

\[
\phi(\hat{w}) + \frac{1}{2} \|\hat{w} - w_0\|^2 \leq \phi(w) + \frac{1}{2} \|w - w_0\|^2 - \frac{1}{2} \|w - \hat{w}\|^2.
\]

**Proof of the convergence rate.**

\[
F(w_{t+1}) \leq \tilde{F}(w_{t+1}; w_t) + \frac{L}{2} \|w_{t+1} - w_t\|^2 \quad \text{(Sandwich-left)}
\]
\[
\leq \tilde{F}(w^*; w_t) + \frac{L}{2} \|w^* - w_t\|^2 - \frac{L}{2} \|w^* - w_{t+1}\|^2 \quad \text{(3-point with } w = w^*)
\]
\[
\leq F(w^*) + \frac{L}{2} \|w^* - w_t\|^2 - \frac{L}{2} \|w^* - w_{t+1}\|^2 \quad \text{(Sandwich-right)}
\]
Let us now define $\epsilon_t := F(w_t) - F(w^*)$, so that

$$\epsilon_{t+1} \leq \frac{L}{2} \| w^* - w_t \|^2 - \frac{L}{2} \| w^* - w_{t+1} \|^2.$$ 

The sequence $\epsilon_t$, for $t = 0, \ldots, T$ is monotone non-increasing.

Let us consider the telescopic sum

$$\sum_{t=1}^{T-1} \epsilon_{t+1} \leq \frac{L}{2} \| w^* - w_0 \|^2 - \frac{L}{2} \| w^* - w_T \|^2 \leq \frac{L}{2} \| w^* - w_0 \|^2.$$

Since $\epsilon_t$ is monotone non-increasing,

$$T \epsilon_T \leq \sum_{t=1}^{T-1} \epsilon_{t+1}$$

which leads to

$$F(w_T) - F(w^*) = \epsilon_T \leq \frac{L \| w^* - w_0 \|^2}{2T}.$$
We want to accelerate Algorithm 1, by introducing some factors tending to zero.

We define $w_{t+1}$ by taking the linear approximation at an auxiliary point $v_t$:

$$w_{t+1} := \arg\min_w \tilde{F}(w; v_t) + \frac{L}{2} \|w - v_t\|^2.$$ 

We perform the same analysis as above, letting $v^*$ be a reference vector that will be chosen later

$$F(w_{t+1}) \leq \tilde{F}(w_{t+1}; v_t) + \frac{L}{2} \|w_{t+1} - v_t\|^2 \quad \text{(Sandwich-left)}$$

$$\leq \tilde{F}(v^*; v_t) + \frac{L}{2} \|v^* - v_t\|^2 - \frac{L}{2} \|v^* - w_{t+1}\|^2 \quad \text{(3-point with } w = v^*)$$

$$\leq F(v^*) + \frac{L}{2} \|v^* - v_t\|^2 - \frac{L}{2} \|v^* - w_{t+1}\|^2 \quad \text{(Sandwich-right)}$$

By introducing yet another sequence $\{u_t\}$, we would like to obtain

$$F(w_{t+1}) \leq F(v^*) + \frac{L\theta_t^2}{2} \|w^* - u_t\|^2 - \frac{L\theta_t^2}{2} \|w^* - u_{t+1}\|^2. \quad \text{(WANT)}$$
\[ F(w_{t+1}) \leq F(v^*) + \frac{L}{2} \| v^* - v_t \|^2 - \frac{L}{2} \| v^* - w_{t+1} \|^2 \]

\[ F(w_{t+1}) \leq F(v^*) + \frac{L\theta_t^2}{2} \| w^* - u_t \|^2 - \frac{L\theta_t^2}{2} \| w^* - u_{t+1} \|^2. \]  

(WANT)

In order for (WANT) to hold, we need

\[ v^* - v_t = \theta_t (w^* - u_t) \]
\[ v^* - w_{t+1} = \theta_t (w^* - u_{t+1}). \]

To satisfy the second relation we can choose

\[ v^* = \alpha_t + \theta_t w^* \]
\[ u_{t+1} = \frac{\alpha_t - w_{t+1}}{\theta_t} \]

In order to to exploit the convexity of \( F \), we can choose

\[ \alpha_t = (1 - \theta_t) w_t \]

so that \( v^* \) becomes a convex combination of \( w^* \) and the previous point \( w_t \).
In summary, we have

\[ v^* = (1 - \theta_t)w_t + \theta_t w^* \]

\[ u_{t+1} = \frac{w_{t+1} - (1 - \theta_t)w_t}{\theta_t} \]

\[ v_t = (1 - \theta_t)w_t + \theta_t u_t. \]

**Accelerated Algorithm**

\[ w_0, u_0 \leftarrow 0 \]

**for** \( t = 0, 1, \ldots, T \) **do**

\[ v_t \leftarrow (1 - \theta_t)w_t + \theta_t u_t \]

\[ w_{t+1} \leftarrow \text{argmin}_w \tilde{F}(w; v_t) + \frac{\ell}{2} \|w - v_t\|^2 = \text{prox}_{\frac{\ell}{2}} \left( v_t - \frac{1}{L} \nabla f(v_t) \right) \]

\[ u_{t+1} \leftarrow \frac{w_{t+1} - (1 - \theta_t)w_t}{\theta_t} \]

**end for**
If we consider the sequence $\{\theta_t\}$

$$
\theta_0 = 1 \\
\frac{1 - \theta_{t+1}}{\theta_{t+1}^2} = \frac{1}{\theta_t^2}
$$

(Theta-Def)

which satisfies

$$
\theta_t \leq \frac{2}{(t + 2)},
$$

(2)

we have the following convergence rate

$$
F(w_{T+1}) - F(w^*) \leq \frac{L}{2} \theta_T^2 \|w^*\|^2 \leq \frac{2L}{T^2} \|w^*\|^2
$$

where $w^*$ is a minimizer of $F$. 
\( O(1/T^2) \) algorithm - convergence rate proof

\[
F(w_{t+1}) \leq F(v^*) + \frac{L\theta_t^2}{2} \| w^* - u_t \|^2 - \frac{L\theta_t^2}{2} \| w^* - u_{t+1} \|^2. \quad \text{(WANT)}
\]

\[
F(w_{t+1}) \leq (1 - \theta_t)F(w_t) + \theta_t F(w^*) + \frac{L\theta_t^2}{2} \| w^* - u_t \|^2 - \frac{L\theta_t^2}{2} \| w^* - u_{t+1} \|^2.
\]

Define \( \varepsilon_t := F(w_t) - F(w^*) \) and \( \Phi_t := \frac{1}{2} \| w^* - u_t \|^2 \),

\[
\varepsilon_{t+1} \leq (1 - \theta_t)\varepsilon_t + \theta_t^2 (\Phi_t - \Phi_{t+1})
\]

\[
\frac{1}{\theta_t^2} \varepsilon_{t+1} - \frac{1 - \theta_t}{\theta_t^2} \varepsilon_t \leq \Phi_t - \Phi_{t+1}
\]

\[
\frac{1 - \theta_{t+1}}{\theta_{t+1}^2} \varepsilon_{t+1} - \frac{1 - \theta_t}{\theta_t^2} \varepsilon_t \leq \Phi_t - \Phi_{t+1} \quad \text{Using (Theta-Def)}
\]

Taking the sum from \( t = 1 \) to \( t = T \) gives

\[
\frac{1 - \theta_{T+1}}{\theta_{T+1}^2} \varepsilon_{T+1} \leq \Phi_0 - \Phi_{T+1} + \frac{1 - \theta_0}{\theta_0^2} \varepsilon_0
\]

\[
\frac{1 - \theta_{T+1}}{\theta_{T+1}^2} \varepsilon_{T+1} = \frac{1}{\theta_T^2} \varepsilon_{T+1} \leq \Phi_0 = \frac{L}{2} \| w^* - u_0 \|^2 \]
Simple numerical comparison

Solve LASSO with $d = 100$ variables and
- Regression vector $\tilde{w}$ has 20 nonzero components with random $\pm 1$
- $n = 40$ examples, $x_{ij} \sim \mathcal{N}(0, 1)$ and $y = Xw + \varepsilon$, $\varepsilon_i \sim \mathcal{N}(0, 0.01)$.
Proximal algorithms are fast first-order method.
The accelerated algorithm has $O(1/T^2)$ convergence rate.
All you need is:
  - The gradient of the smooth part
  - The proximal operator of the non-smooth part