GI01/M055 – Supervised Learning
Proximal Methods

Massimiliano Pontil
(based on notes by Luca Baldassarre)

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 interesting in the 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: $r(w) = \lambda\|w\|_*$

Convex Analysis I

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

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{L}{2}||w - v||^2.$$

Convex Analysis II

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, $f(w) \geq f(v) + \langle \nabla f(v), w - v \rangle$. □

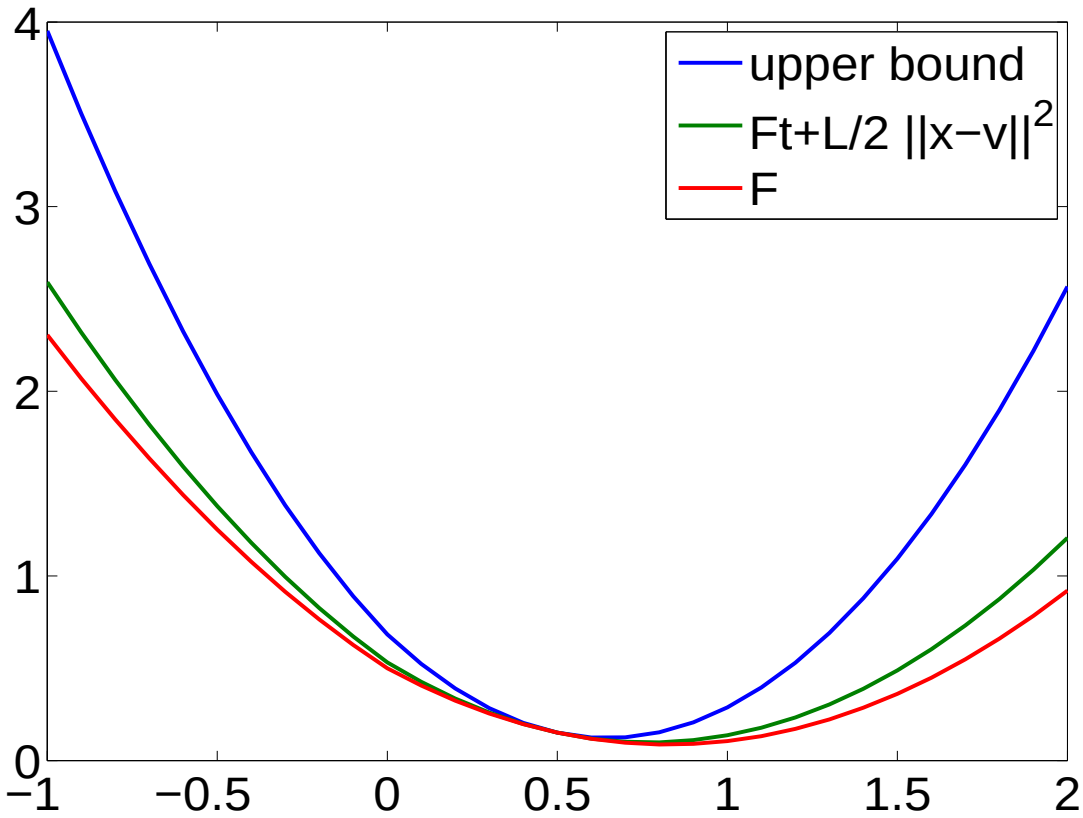
Equivalent version

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

Sandwich example

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

$$\tilde{F}(w; v) = \frac{1}{2}(av - 1)^2 + a(av - 1)(w - v) + \frac{1}{2}|w|$$



Subdifferential

If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, its **subdifferential** at w is defined as

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

- ∂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

$$\hat{w} \in \operatorname*{argmin}_{w \in \mathbb{R}^d} f(w) \iff 0 \in \partial f(\hat{w})$$

Proximal operator

The **proximal operator** of a convex function $r : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{prox}_r(v) = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{2}\|w - v\|^2 + r(w).$$

Convexity of r ensures that the minimizer always exists and is unique.
Examples:

- LASSO: $r(w) = \lambda\|w\|_1$, $\mathrm{prox}_r(v) = H_\lambda(v)$, which is the component-wise soft-thresholding operator $(H_\lambda(v))_i = \mathrm{sign}(v_i)(|v_i| - \lambda)_+$
- ℓ_2 NORM: $r(w) = \lambda\|w\|_2$, $\mathrm{prox}_r(v) = \frac{v}{\|v\|_2}(\|v\|_2 - \lambda)_+$.
- GROUP LASSO: $r(w) = \sum_{\ell=1}^K \|w_{|J_\ell}\|_2$, $(\mathrm{prox}_r(v))_{|J_\ell} = \frac{v_{|J_\ell}}{\|v_{|J_\ell}\|}(\|v_{|J_\ell}\|_2 - \lambda)_+$.

$O(1/T)$ algorithm

Algorithm

$w_0 \leftarrow 0$
for $t = 0, 1, \ldots, T$ **do**
$\quad w_{t+1} = \text{argmin}_w \frac{L}{2}\|w - w_t\|^2 + \tilde{F}(w; w_t)$
end for

Recall,

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

The term $f(w_t)$ does not depend on w and can be discarded.
By completing the square, we also obtain

$$\frac{L}{2}\|w - w_t\|^2 + \langle \nabla f(w_t), w - w_t \rangle + \left(\frac{1}{L}\nabla f(w_t)\right)^2 = \frac{L}{2}\left\|w - \left(w_t - \frac{1}{L}\nabla f(w_t)\right)\right\|^2$$

where the extra term $\left(\frac{1}{L}\nabla f(w_t)\right)^2$ does not depend on w.

$O(1/T)$ algorithm (cont'd)

Algorithm 1

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

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

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

Theorem (Convergence rate)

Let $w^* \in \text{argmin}_w F(w)$, then, at iteration T, Algorithm 1 yields a solution w_T that satisfies

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

$O(1/T)$ algorithm - Examples

LASSO

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

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

GROUP LASSO

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

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

$$(w_{t+1})_{|J_\ell} = \frac{v_{|J_\ell}}{\|v_{|J_\ell}\|}\left(\|v_{|J_\ell}\|_2 - \frac{\lambda}{L}\right)_+$$

$O(1/T)$ 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.

$$\begin{aligned}
F(w_{t+1}) &\leq \tilde{F}(w_{t+1}; w_t) + \frac{L}{2}\|w_{t+1} - w_t\|^2 && \text{(Sandwich-left)} \\
&\leq \tilde{F}(w^*; w_t) + \frac{L}{2}\|w^* - w_t\|^2 - \frac{L}{2}\|w^* - w_{t+1}\|^2 && \text{(3-point with } w = w^*) \\
&\leq F(w^*) + \frac{L}{2}\|w^* - w_t\|^2 - \frac{L}{2}\|w^* - w_{t+1}\|^2 && \text{(Sandwich-right)}
\end{aligned}$$

$O(1/T)$ algorithm - convergence rate proof cont'd

Let us now define $\varepsilon_t := F(w_t) - F(w^*)$, so that

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

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

$$\begin{aligned} F(w_{t+1}) &\leq \tilde{F}(w_{t+1}; w_t) + \frac{L}{2}\|w_{t+1} - w_t\|^2 && \text{(Sandwich-left)} \\ &\leq \tilde{F}(w_t; w_t) + \frac{L}{2}\|w_t - w_t\|^2 = F(w_t) && \text{(Def. of } w_{t+1}\text{)} \quad \square \end{aligned}$$

Since ε_t is monotone non-increasing,

$$T\varepsilon_T \leq \sum_{t=0}^{T-1} \varepsilon_{t+1} \leq \frac{L}{2}\|w^* - w_0\|^2 - \frac{L}{2}\|w^* - w_T\|^2 \leq \frac{L}{2}\|w^* - w_0\|^2$$

$$\varepsilon_T = F(w_T) - F(w^*) \leq \frac{L\|w^* - w_0\|^2}{2T} \quad \square$$

$O(1/T^2)$ algorithm

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:

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

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

$$\begin{aligned} F(w_{t+1}) &\leq \tilde{F}(w_{t+1}; v_t) + \frac{L}{2}\|w_{t+1} - v_t\|^2 && \text{(Sandwich-left)} \\ &\leq \tilde{F}(v^*; v_t) + \frac{L}{2}\|v^* - v_t\|^2 - \frac{L}{2}\|v^* - w_{t+1}\|^2 && \text{(3-point with } w = v^*) \\ &\leq F(v^*) + \frac{L}{2}\|v^* - v_t\|^2 - \frac{L}{2}\|v^* - w_{t+1}\|^2 && \text{(Sandwich-right)} \end{aligned}$$

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

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

$O(1/T^2)$ algorithm - cont'd

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

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

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

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

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.

$O(1/T^2)$ algorithm - cont'd

In summary, we have

$$
\begin{aligned}
v^* &= (1-\theta_t)w_t + \theta_t w^* \\
u_{t+1} &= \frac{w_{t+1} - (1-\theta_t)w_t}{\theta_t} \\
v_t &= (1-\theta_t)w_t + \theta_t u_t.
\end{aligned}
$$

Accelerated Algorithm

$w_0, u_0 \leftarrow 0$
for $t = 0, 1, \ldots, T$ **do**
  $v_t \leftarrow (1-\theta_t)w_t + \theta_t u_t$
  $w_{t+1} \leftarrow \operatorname{argmin}_w \tilde{F}(w; v_t) + \frac{L}{2}\|w - v_t\|^2 = \operatorname{prox}_{\frac{r}{L}}\left(v_t - \frac{1}{L}\nabla f(v_t)\right)$
  $u_{t+1} \leftarrow \frac{w_{t+1} - (1-\theta_t)w_t}{\theta_t}$
end for

$O(1/T^2)$ algorithm - convergence rate

Let $w^* \in \operatorname{argmin}_w F(w)$, then, at iteration T, *Algorithm 2* yields w_T that satisfies

$$F(w_T) - F(w^*) \leq \frac{L}{2}\theta_T^2 \|w^*\|^2 \tag{2}$$

Let us consider the sequence $\{\theta_t\}$

$$\theta_0 = 1$$
$$\frac{1-\theta_{t+1}}{\theta_{t+1}^2} = \frac{1}{\theta_t^2}. \tag{Theta-Def}$$

The sequence $\{\theta_t\}$ satisfies

$$\theta_t \leq 2/(t+2), \tag{3}$$

Proof by induction: use arithmetic-geometric inequality on (Theta-Def).

We then have the following convergence rate

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

$O(1/T^2)$ algorithm - convergence rate proof

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

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

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

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

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

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

Taking the sum from $t = 1$ to $t = T$ gives

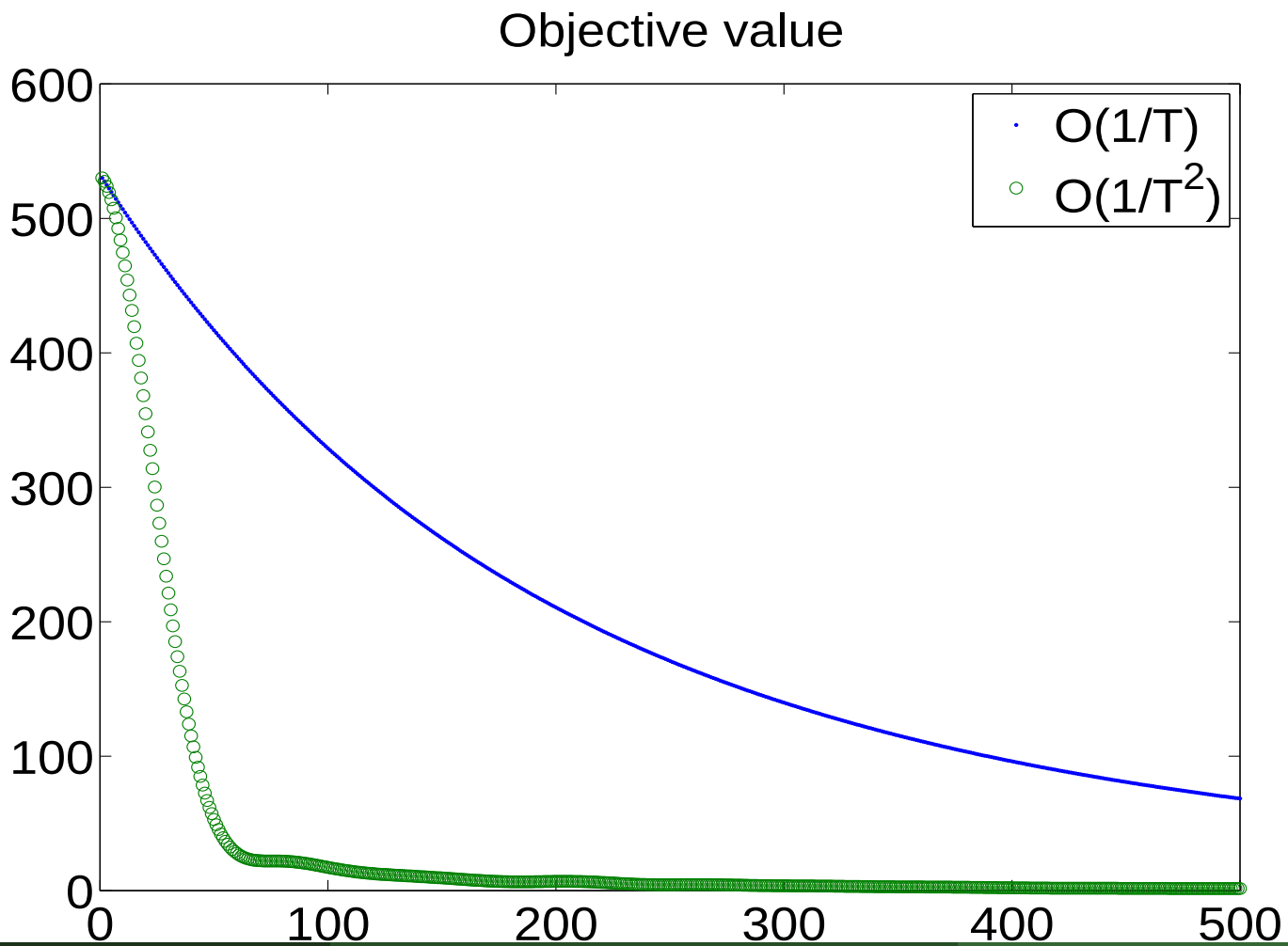
$$\frac{1-\theta_{T+1}}{\theta_{T+1}^2}\varepsilon_{T+1} \leq \Phi_0 - \Phi_{T+1} + \frac{1-\theta_0}{\theta_0^2}\varepsilon_0$$

$$\frac{1-\theta_{T+1}}{\theta_{T+1}^2}\varepsilon_{T+1} = \frac{1}{\theta_T^2}\varepsilon_{T+1} \leq \Phi_0 = \frac{L}{2}\|w^* - u_0\|^2 \quad \square$$

Simple numerical comparison

Solve LASSO with $d = 100$ variables and

- Regression vector $\tilde{w}$ has 20 nonzero components with random ± 1
- $n = 40$ examples, $x_{ij} \backsim \mathcal{N}(0, 1)$ and $y = Xw + \varepsilon$, $\varepsilon_i \backsim \mathcal{N}(0, 0.01)$.



Summary

- Proximal algorithms are fast first-order methods.
- 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