Proximal methods

S. Villa

21st October 2013

0.1 Review of the basics

Often machine learning problems require the solution of minimization problems. For instance, the ERM algorithm requires to solve a problem of the form

$$\min_{c \in \mathbb{R}^d} \|y - Kc\|^2,$$

for various choices of the loss function. Another typical problem is the regularized one, e.g. Tikhonov regularization where, for linear kernels one looks for

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} V(\langle w, x_i \rangle, y_i) + \lambda R(w).$$

More generally, we are interested in solving a minimization problem

$$\min_{w \in \mathbb{R}^d} F(w).$$

We review the basic concepts that allow to study the problem.

Existence of a minimizer We will consider extended real valued functions $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The **domain** of F is

$$\mathrm{dom} F = \{w \in \mathbb{R}^d \colon F(w) < +\infty\}.$$

This all F is proper if the domain is nonempty. It is useful to consider extended valued functions since they allow to include constraints in the regularization.
F is lower semicontinuous if epi$F$ is closed (example). F is coercive if $\lim_{\|w\| \to +\infty} F(w) = +\infty$.

Theorem 0.1.1. *If F is lower semicontinuous and coercive then there exists w_* such that $F(w_*) = \min F$.*

We will always assume that the functions we consider are lower semicontinuous.

0.1.1 Convexity concepts

Convexity F is convex if

$$(\forall w, w' \in \mathrm{dom} F)(\forall \lambda \in [0,1]) \quad F(\lambda w + (1-\lambda)w') \le \lambda F(w) + (1-\lambda)F(w').$$

If F is differentiable, we can write an equivalent characterization of convexity based on the gradient:

$$(\forall w, w' \in \mathbb{R}^d) \quad F(w') \ge F(w) + \langle \nabla F(w), w' - w \rangle$$

If F is twice differentiable, and $\nabla^2 F$ is the Hessian matrix, convexity is equivalent to $\nabla^2 F(w)$ positive semidefinite for all $w \in \mathbb{R}^d$.

If a function is convex and differentiable, then $\nabla F(w) = 0$ implies that w is a global minimizer.

Strict Convexity F is strictly convex if $(\forall w, w' \in \mathrm{dom} F)(\forall \lambda \in (0,1))$

$$F(\lambda w + (1-\lambda)w') < \lambda F(w) + (1-\lambda)F(w').$$

If F is differentiable, we can write an equivalent charcterization of strct convexity based on the gradient:

$$(\forall w, w' \in \mathbb{R}^d) \quad F(w') > F(w) + \langle \nabla F(w), w' - w \rangle$$

If F is twice differentiable, and $\nabla^2 F$ is the Hessian matrix, convexity is implied by $\nabla^2 F(w)$ positive definite for all $w \in \mathbb{R}^d$. The minimizer of a strictly convex function is unique (if it exists)

Strong Convexity F is μ-strongly convex if the function $f - \mu\|\cdot\|^2$ is convex, i.e. $(\forall w, w' \in \mathrm{dom}F)(\forall \lambda \in [0,1])$

$$F(\lambda w + (1-\lambda)w') \leq \lambda F(w) + (1-\lambda)F(w') - \frac{\mu}{2}\lambda(1-\lambda)\|w - w'\|^2.$$

If F is differentiable, then strong convexity is equivalent to $(\forall w, w' \in \mathbb{R}^d)$

$$F(w') \geq F(w) + \langle \nabla F(w), w' - w \rangle + \frac{\mu}{2}\|w - w'\|^2$$

If F is twice differentiable, and $\nabla^2 F$ is the Hessian matrix, strong convexity is equivalent to $\nabla^2 F(w) \geq \mu I$ for all $w \in \mathbb{R}^d$. If F is strongly convex then it is coercive. Therefore if it is lsc, it admits a unique minimizer. Moreover

$$F(w) - F(w_*) \geq \frac{\mu}{2}\|w - w_*\|^2 .$$

We will often assume Lipschitz continuity of the gradient

$$\|F(w) - F(w')\| \leq L\|w - w'\|.$$

This gives a useful quadratic upper bound of F

$$F(w') \leq F(w) + \langle \nabla F(w), w - w' \rangle + \frac{L}{2}\|w' - w\|^2 \quad (\forall w, w' \in \mathrm{dom}F) \tag{1}$$

Moreover, for every $w \in \mathrm{dom}F$ and w_* is a minimizer,

$$\frac{1}{2L}\|\nabla F(w)\|^2 \leq F(w) - F(w_*) \leq \frac{L}{2}\|w - w_*\|^2 .$$

The second inequality follows by substituting in the quadratic upper bound $w = w^*$ and $w' = w$. The first follows by substituting $w' = w - \frac{1}{L}\nabla F(w)$.

0.2 Convergence of the gradient method with constant step-size

Assume F to be convex, differentiable, with L Lipschitz continuous gradient, and that a minimizer exists. The first order necessary condition is $\nabla F(w) = 0$. Therefore

$$w_* - \alpha \nabla F(w_*) = w_*$$

This suggests an algorithm based on the fixed point iteration

$$w_{k+1} = w_k - \alpha \nabla F(w_k) .$$

We want to study convergence of this algorithm. Convergence can be intended in two senses, towards the minimum or towards a minimizer. Start from the first one. Different strategis to choose stepsize. We keep α fixed and determine a priori conditions guaranteeing convergence. From the quadratic upper bound (1) we get

$$\begin{aligned} F(w_{k+1}) &\leq F(w_k) - \alpha\|\nabla F(w_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla F(w_k)\|^2 \\ &= F(w_k) - \alpha\left(1 - \frac{L}{2}\alpha\right)\|\nabla F(w_k)\|^2 \end{aligned}$$

If $0 < \alpha < 2/L$ the iteration decreases the function value. Choose $\alpha = 1/L$ (which gives the maximum decrease) and get

$$\begin{aligned}
F(w_{k+1}) &\leq F(w_k) - \frac{1}{2L}\|\nabla F(w_k)\|^2 \\
&\leq F(w_*) + \langle \nabla F(w_k), w_k - w_* \rangle - \frac{1}{2L}\|\nabla F(w_k)\|^2 \\
&= F(w_*) + \frac{L}{2}\left(\langle \nabla \frac{1}{L} F(w_k), w_k - w_* \rangle - \frac{1}{L^2}\|\nabla F(w_k)\|^2 - \|w_k - w_*\|^2 + \|w_k - w_*\|^2 \right) \\
&= F(w_*) + \frac{L}{2}(\|w_k - w_*\|^2 - \|w_k - \frac{1}{L}\nabla F(w_k) - w_*\|^2) \\
&= F(w_*) + \frac{L}{2}(\|w_k - w_*\|^2 - \|w_{k+1} - w^*\|^2)
\end{aligned}$$

Summing the above inequality for $k = 0, \ldots, K-1$ we get

$$\begin{aligned}
\sum_{k=0}^{K-1} F(w_k) - F(w_*) &\leq \sum_{k=0}^{K-1} \frac{L}{2}(\|w_k - w_*\|^2 - \|w_{k+1} - w^*\|^2) \\
\sum_{k=0}^{K-1} F(w_k) - F(w_*) &\leq \frac{L}{2}\|w_0 - w_*\|^2
\end{aligned}$$

Noting that $F(w_k)$ is decreasing, $F(w_K) - F(w_*) \leq F(w_k) - F(w^*)$ for every k, therefore we obtain

$$F(w_K) - F(w_*) \leq \frac{L}{2K}\|w_0 - w_*\|^2 .$$

This is called sublinear rate of convergence. For strongly convex functions, it is possible to prove that the operator $I - \alpha\nabla F$ is a contraction, and therefore we get linear convergence rate:

$$\|w_K - w_*\|^2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{2K} \|w_0 - w_*\|^2$$

which gives, using the bound following (1)

$$F(w_K) - F(w_*) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2K} \|w_0 - w_*\|^2$$

which is much better.

It is known that for general convex problems problems, with Lipschitz continuous gradient, the performance of any first order method is lower bounded by $1/k^2$. Nesterov in 1983 devised an algorithm reaching the lower bound. The algorithm is called **accelerated gradient descent** and is very similar to the gradient. It needs to store two iterates, instead of only one. It is of the form

$$\begin{aligned}
w_{k+1} &= u_k - \frac{1}{L}\nabla F(u_k) \\
u_{k+1} &= a_k w_k + b_k w_{k+1},
\end{aligned}$$

for some $w_0 \in \mathrm{dom}F$, and $u_1 = w_0$ and a suitable (a priori determined) sequence of parameters a_k and b_k. More precisely, choose $w_0 \in \mathrm{dom}F$, and $u_1 = w_0$. Set $t_1 = 1$. Then define

$$\begin{aligned}
w_{k+1} &= u_k - \frac{1}{L}\nabla F(u_k) \\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\
u_{k+1} &= \left(1 + \frac{t_k - 1}{t_{k+1}}\right) w_k + \frac{1 - t_k}{t_{k+1}} w_{k+1}.
\end{aligned}$$

We obtain

$$F(w_k) - F(w_*) \leq \frac{L\|w_0 - w_*\|^2}{2k^2}$$

0.3 Regularized optimization

We often want to minimize

$$\min_{w \in \mathbb{R}^d} F(w) + R(w),$$

where either F is smooth (e.g. square loss) and R is convex and nonsmooth, either R is smooth and F is not (SVM). We would like to write a similar condition to $\nabla = 0$ to characterize a minimizer. We use the subdifferential. Let R be a convex, lsc proper function. $\eta \in \mathbb{R}^d$ is a subgradient of R at w if

$$R(w') \geq R(w) + \langle \eta, w' - w \rangle.$$

The subdifferential $\partial R(w)$ is the set of all subgradients. It is easy to see that

$$R(w_*) = \min R \iff 0 \in \partial R(w_*).$$

If R is differentiable, the subdifferential is a singleton and coincides with the gradient.
Example 1) Indicator function of a convex set C (constrained regularization). Let $w \notin C$. Then $\partial i_C = \varnothing$. If $w \in C$, then $\eta \in \partial i_C(w)$ if and only if, for all $v \in C$

$$i_C(v) - i_C(w) \geq \langle \eta, w - v \rangle \iff 0 \geq \langle \eta, w - v \rangle.$$

This is the normal cone to C.
2) Subdifferential of $R(w) = \|w\|_1$.

$$\sum_{i=1}^n |v_i| - \sum_{i=1}^n |w_i| \geq \langle \eta, v - w \rangle.$$

If, η is such that for all $i = 1, \ldots, d$

$$|v_i| - |w_i| \geq \eta_i(v_i - w_i),$$

then $\eta \in \partial R(w)$. Vice versa, taking $v_j = w_j$ for all $j \neq i$ we get that $\eta \in \partial R(w)$ implies that $|v_i| - |w_i| \geq \eta_i(v_i - w_i)$, and thus $\eta_i \in \partial |\cdot|(w_i)$. We therefore proved that

$$\partial R(w) = \left(\partial |\cdot|(w_1), \ldots, \partial |\cdot|(w_d)\right).$$

Proximity operator Let R be lsc, convex, proper. Then

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

is well-defined and is unique. Imposing the first order necessary conditions, we get

$$u = \mathrm{prox}_R(v) \iff 0 \in \partial R(u) + (u - v) \iff v - u \in \partial R(u) \iff u = (I + \partial R)^{-1}(v)$$

Examples If $R = 0$ then $\mathrm{prox}(v) = v$. If $R = i_C$ then $\mathrm{prox}_R(v) = P_C(v)$. Proximity operator of the l_1 norm. Let $v \in \mathbb{R}^d$ and $u = \mathrm{prox}_R(v)$. Then $v - u \in \partial \|\cdot\|_1(u)$. SInce the subdifferential can be computed componentwise, the same holds for the prox. In particular, $u = (I + \partial R)^{-1}(v)$ By the previous example, this is equivalent to $u = (I + \partial R)^{-1}(v)$. To compute this quantity first note that

$$((I + \partial R)(v))_i = \begin{cases} v_i + 1 & \text{if } v_i > 1 \\ [-1, 1] & \text{if } v_i = 0 \\ v_i - 1 & \text{if } v_i < -1 \end{cases}$$

Inverting the previous relationship we get

$$(\text{prox}_{\|\cdot\|_1}(u))_i = \begin{cases} u_i - 1 & \text{if } u_i > 1 \\ 0 & \text{if } u_i \in [-1,1] \\ u_i + 1 & \text{if } u_i < -1 \end{cases}$$

0.4 Basic proximal algorithm (forward-backward splitting)

Assume that F is convex and differentiable with Lipschitz continuous gradient. As for gradient descent, the idea is to start from a fixed point equation characterizing the minimizer. If we write the first order conditions, we get

$$\begin{aligned} & 0 \in \nabla F(w_*) + \partial R(w_*) \\ \iff & -\alpha \nabla F(w_*) \in \alpha \partial R(w_*) \\ \iff & w_* - \alpha \nabla F(w_*) - w_* \in \partial \alpha R(w_*) \\ \iff & w_* = \text{prox}_{\alpha R}(w_* - \alpha \nabla F(w_*)). \end{aligned}$$

We consider the fixed point iteration

$$w_{k+1} = \text{prox}_{\alpha_k R}(w_k - \alpha_k \nabla F(w_k))\,.$$

Another interpretation:

$$\begin{aligned} w_{k+1} &= \text{argmin}\{\alpha_k R(w) + \frac{1}{2}\|w - (w_k - \alpha_k \nabla F(w_k))\|^2\} \\ &= \text{argmin}\{R(w) + \frac{1}{2\alpha_k}\|w\|^2 + \langle w - w_k, \nabla F(w_k)\rangle + F(w_k)\} \end{aligned}$$

Special cases: $R = 0$ (gradient method), $R = i_C$ (projected gradient method). The proof of convergence for the sequence of objective values with $\alpha_k = 1/L$ is similar to the proof of convergence for the differentiable case. The rate of convergence is the same as in the differentiable case (this would not be the case if a subdifferential method was used, compare...)

$$F(w_k) - F(w_*) \leq \frac{L\|w_0 - w_*\|^2}{2k}$$

Convergence proof Set $\alpha_k = 1/L$ and define the "gradient mapping" as

$$G_{1/L}(w) = L(w - \frac{1}{L}\text{prox}_{R/L}(w - \frac{1}{L}\nabla F(w)))$$

Then

$$w_{k+1} = w_k - \frac{1}{L}G_{1/L}(w_k).$$

Note that $G_{1/L}$ is not a gradient or a subgradient of $F + R$ but is called gradient mapping. By wiriting the first order condition for the prox operator, we get:

$$G_{1/L}(w) \in \nabla F(w) + \partial R(w - \frac{1}{L}G_{1/L}(w))$$

Recalling the upper bound (1), we obtain

$$F(w - \frac{1}{L}G_{1/L}(w)) \leq F(w) - \frac{1}{L}\langle \nabla F(w), G_{1/L}(w)\rangle + \frac{1}{2L}\|G_{1/L}(w)\|^2 \tag{2}$$

If inequality (2) holds, then for every $v \in \mathbb{R}^d$:

$$F(w - \frac{1}{L}G_{1/L}(w)) \leq F(v) + \langle G_{1/L}(w), w - v\rangle + \frac{1}{2L}\|G_{1/L}(w)\|^2$$

....

Accelerated versions As for the gradient.
The problem is that the forward-backward algorithm is effective only when prox is easy to compute. Note indeed that we replaced our original problem with a sequence of new minimization problems. They are strongly convex (therefore easier), but in general not solvable in closed form.

0.5 Fenchel conjugate and Moreau decomposition

Fenchel conjugate The Fenchel conjugate is a function $R^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined as

$$R^*(\eta) = \sup_{w \in \mathbb{R}^d} \{\langle \eta, w\rangle - R(w)\}.$$

R^* is a convex function (even if R is not), since it is the pointwise supremum of convex (linear) functions.
Example

1. Conjugate of an affine function. It is the support function. If $R(w) = \langle a, w\rangle + b$, then $R^*(w) = -b\iota_{\{a\}}$
2. Conjugate of an Indicator function.
3. Conjugate of the norm $R(w) = \|w\|$. In this case define $\|\eta\|_* = \sup_{w:\|w\|\leq 1}\langle \eta, w\rangle$. Then $R^* = I_{B_*(1)}$. (for the l_1 norm, it is the l_∞ norm) Let $\eta \notin B_*(1)$. Then $\|\eta\|_* = \sup_{w:\|w\|\leq 1}\langle \eta, w\rangle > 1$. Therefore there exists $\bar{w}$, $\|\bar{w}\| \leq 1$ such that $\langle \eta, \bar{w}\rangle > 1$. Thus,

   $$R^*(\eta) = \sup_{w \in \mathbb{R}^d} \langle \eta, w\rangle - \|w\| \geq \langle \eta, \bar{w}\rangle - \|\bar{w}\| > 0.$$

   Now, taking $w = t\bar{w}$, we derive $R^*(\eta) = +\infty$. On the other hand, if $\eta \in B_*(1)$, $\sup_{w:\|w\|\leq 1}\langle \eta, w\rangle \leq 1$ and thus

   $$R^*(\eta) = \sup_{w \in \mathbb{R}^d} \langle \eta, w\rangle - \|w\| \leq 0.$$

   Taking $w = 0$ we obtain $R^*(\eta) = 0$.

By definition, $R^*(\eta) + R(w) \geq \langle \eta, w\rangle$ for $\eta, w \in \mathbb{R}^d$ (Fenchel Young inequality). Moreover,

$$R(w) + R^*(\eta) = \langle \eta, w\rangle \iff \eta \in \partial R(w) \iff w \in \partial R^*(\eta)$$

Suppose that $R^*(\eta) = \langle \eta, w\rangle - R(w)$ iff $\langle \eta, w'\rangle - R(w') \leq \langle \eta, w\rangle - R(w)$ for every w' iff $\eta \in \partial R(w)$. From $R(w) + R^*(\eta) = \langle w, \eta\rangle$ we get

$$\begin{aligned} R^*(\eta') &= \sup_u \langle \eta', u\rangle - R(u) \\ &\geq \langle \eta', w\rangle - R(w) \\ &= \langle \eta' - \eta, w\rangle + \langle \eta, w\rangle - R(w) \\ &= \langle \eta' - \eta, w\rangle - R^*(\eta). \end{aligned}$$

If R is lsc and convex, then $R^{**} = R$ (which gives the other equivalence).

Moreau decomposition

$$w = \mathrm{prox}_R(w) + \mathrm{prox}_{R^*}(w)$$

It follows from the properties stated above of the subdifferential and of the conjugate:

$$\begin{aligned} u = \mathrm{prox}_R(w) &\iff w - u \in \partial R(u) \\ &\iff u \in \partial R^*(w-u) \\ &\iff w - (w-u) \in \partial R^*(w-u) \\ &\iff w - u = \mathrm{prox}_{R^*}(w). \end{aligned}$$

Note that this is a generalization of the classical decomposition on orthogonal components. So if V is a linear subspace and $V^\perp$ is the orthogonal subspace, we know $w = P_V(w) + P_{V^\perp}(w)$. This is a special case of the Moreau decomposition obtained by choosing $R = i_V$ (and noting that $R^* = i_{V^\perp}$).

Properties of the proximity operators – examples Separable sum: If $R(w) = R_1(w_1) + R_2(w_2)$, then $\mathrm{prox}_R(w) = (\mathrm{prox}_{R_1}(w_1), \mathrm{prox}_{R_2}(w_2))$. Scaling:

$$\mathrm{prox}_{R+\frac{\mu}{2}\|\cdot\|^2}(v) = \mathrm{prox}_{\frac{1}{1+\mu}R}\left(\frac{v}{1+\mu}\right)$$

"Generalized" Moreau decomposition: for every $\lambda > 0$:

$$w = \mathrm{prox}_{\lambda R}(w) + \lambda \mathrm{prox}_{R^*/\lambda}(w/\lambda)$$

Sometimes, Moreau decomposition is useful to compute proximity operators. Let $R(w) = \lambda\|w\|$. We have seen that $R^* = i_{B_*(\lambda)}$. Therefore, from the Moreau decomposition, we get

$$\mathrm{prox}_R(w) = w - P_{B_*(\lambda)}(w).$$

In particular, if $R = \|\cdot\|_1$, noting that $\|\cdot\|_* = \|\cdot\|_\infty$, we obtain again the formula for the soft-thresholding seen before.

Elastic-net.

Let $G = \{G_1, \ldots, G_t\}$ be a partition of the indices $\{1, \ldots, d\}$. The following norm is called group lasso penalty:

$$R(w) = \sum_{i=1}^{t} \|w\|_{G_i},$$

where $\|w\|_{G_i}^2 = \sum_{j \in G_i} w_j^2$. The dual norm is

$$\max_{j=1,\ldots,t} \|w\|_{G_j},$$

and therefore

$$\mathrm{prox}_R(w) = w - P_{B_*(\lambda)}(w),$$

where $B_*(\lambda) = \{w \in \mathbb{R}^d \colon \|w\|_{G_j} \le \lambda, \forall j = 1, \ldots, t\}$. The projection on this set can be expressed componentwise as

$$(P_{B_*(\lambda)}(w))_{G_j} = \begin{cases} w_{G_j} & \text{if } \|w\|_{G_j} \le \lambda \\ \lambda \dfrac{w_{G_j}}{\|w\|_{G_j}}, & \text{otherwise} \end{cases}$$

0.6 references

Combettes, Pesquet Proximal splitting methods in signal processing, 2009
Combettes and Wajs, SIGNAL RECOVERY BY PROXIMAL FORWARD-BACKWARD SPLITTING, Multiscale Model Simul, 2005
Nesterov, A basic course in otpimization
Beck- Teboulle, A fast iterative soft-thresholding algorithm for linear inverse problems, SIAM J Imaging Sciences 2009