Title: Proximal gradient methods for learning

URL: https://en.wikipedia.org/wiki/Proximal_gradient_methods_for_learning

PageID: 41200806

Categories: Category:Convex optimization, Category:First order methods, Category:Machine learning

-----

Proximal gradient (forward backward splitting) methods for learning is an area of research in optimization and statistical learning theory which studies algorithms for a general class of convex regularization problems where the regularization penalty may not be differentiable . One such example is ■ 1 {\displaystyle \ell _{1}} regularization (also known as Lasso) of the form

Proximal gradient methods offer a general framework for solving regularization problems from statistical learning theory with penalties that are tailored to a specific problem application. [1] [2] Such customized penalties can help to induce certain structure in problem solutions, such as sparsity (in the case of lasso) or group structure (in the case of group lasso).

Relevant background

Proximal gradient methods are applicable in a wide variety of scenarios for solving convex optimization problems of the form

where F {\displaystyle F} is convex and differentiable with Lipschitz continuous gradient , R {\displaystyle R} is a convex , lower semicontinuous function which is possibly nondifferentiable, and H {\displaystyle {\mathcal {H}}} is some set, typically a Hilbert space . The usual criterion of x {\displaystyle x} minimizes F (x) + R (x) {\displaystyle F(x)+R(x)} if and only if ∇ (F + R) (x) = 0 {\displaystyle \nabla (F+R)(x)=0} in the convex, differentiable setting is now replaced by

where ∂ ϕ {\displaystyle \partial \varphi } denotes the subdifferential of a real-valued, convex function ϕ {\displaystyle \varphi } .

Given a convex function ϕ : H → R {\displaystyle \varphi :{\mathcal {H}}\to \mathbb {R} } an important operator to consider is its proximal operator prox ϕ : H → H {\displaystyle \operatorname {prox} _{\varphi }:{\mathcal {H}}\to {\mathcal {H}}} defined by

which is well-defined because of the strict convexity of the ■ 2 {\displaystyle \ell _{2}} norm. The proximal operator can be seen as a generalization of a projection . [1] [3] [4] We see that the proximity operator is important because x ∗ {\displaystyle x^{*}} is a minimizer to the problem min x ∈ H F (x) + R (x) {\displaystyle \min _{x\in {\mathcal {H}}}F(x)+R(x)} if and only if

Moreau decomposition

One important technique related to proximal gradient methods is the Moreau decomposition, which decomposes the identity operator as the sum of two proximity operators. [1] Namely, let ϕ : X → R {\displaystyle \varphi :{\mathcal {X}}\to \mathbb {R} } be a lower semicontinuous , convex function on a vector space X {\displaystyle {\mathcal {X}}} . We define its Fenchel conjugate ϕ ∗ : X → R {\displaystyle \varphi ^{*}:{\mathcal {X}}\to \mathbb {R} } to be the function

The general form of Moreau's decomposition states that for any x ∈ X {\displaystyle x\in {\mathcal {X}}} and any γ > 0 {\displaystyle \gamma >0} that

which for γ = 1 {\displaystyle \gamma =1} implies that x = prox ϕ ■ (x) + prox ϕ ∗ ■ (x) {\displaystyle x=\operatorname {prox} _{\varphi }(x)+\operatorname {prox} _{\varphi ^{*}}(x)} . [1] [3] The Moreau decomposition can be seen to be a generalization of the usual orthogonal decomposition of a vector space , analogous with the fact that proximity operators are generalizations of projections. [1]

In certain situations it may be easier to compute the proximity operator for the conjugate ϕ ∗ {\displaystyle \varphi ^{*}} instead of the function ϕ {\displaystyle \varphi } , and therefore the Moreau decomposition can be applied. This is the case for group lasso .

Lasso regularization

Consider the regularized empirical risk minimization problem with square loss and with the ■ 1 {\displaystyle \ell _{1}} norm as the regularization penalty:

where x i ∈ R d and y i ∈ R . {\displaystyle x_{i}\in \mathbb {R} ^{d}{\text{ and }}y_{i}\in \mathbb {R} .} The ■ 1 {\displaystyle \ell _{1}} regularization problem is sometimes referred to as lasso (least absolute shrinkage and selection operator). [5] Such ■ 1 {\displaystyle \ell _{1}} regularization problems are interesting because they induce sparse solutions, that is, solutions w {\displaystyle w} to the minimization problem have relatively few nonzero components. Lasso can be seen to be a convex relaxation of the non-convex problem

where ■ w ■ 0 {\displaystyle \|w\|_{0}} denotes the ■ 0 {\displaystyle \ell _{0}} "norm", which is the number of nonzero entries of the vector w {\displaystyle w} . Sparse solutions are of particular interest in learning theory for interpretability of results: a sparse solution can identify a small number of important factors. [5]

Solving for L 1 proximity operator

For simplicity we restrict our attention to the problem where λ = 1 {\displaystyle \lambda =1} . To solve the problem

we consider our objective function in two parts: a convex, differentiable term F (w) = 1 n ∑ i = 1 n (y i − ■ w , x i ■) 2 {\displaystyle F(w)={\frac {1}{n}}\sum _{i=1}^{n}(y_{i}-\langle w,x_{i}\rangle)^{2}} and a convex function R (w) = ■ w ■ 1 {\displaystyle R(w)=\|w\|_{1}} . Note that R {\displaystyle R} is not strictly convex.

Let us compute the proximity operator for R (w) {\displaystyle R(w)} . First we find an alternative characterization of the proximity operator prox R ■ (x) {\displaystyle \operatorname {prox} _{R}(x)} as follows:

u = prox R ■ (x) ■ 0 ∈ ∂ (R (u) + 1 2 ■ u − x ■ 2 2) ■ 0 ∈ ∂ R (u) + u − x ■ x − u ∈ ∂ R (u) . {\displaystyle {\begin{aligned}u=\operatorname {prox} _{R}(x)\iff &0\in \partial \left(R(u)+{\frac {1}{2}}\|u-x\|_{2}^{2}\right)\\\iff &0\in \partial R(u)+u-x\\\iff &x-u;\in \partial R(u).\end{aligned}}}

For R (w) = ■ w ■ 1 {\displaystyle R(w)=\|w\|_{1}} it is easy to compute ∂ R (w) {\displaystyle \partial R(w)} : the i {\displaystyle i} th entry of ∂ R (w) {\displaystyle \partial R(w)} is precisely

Using the recharacterization of the proximity operator given above, for the choice of R (w) = ■ w ■ 1 {\displaystyle R(w)=\|w\|_{1}} and γ > 0 {\displaystyle \gamma >0} we have that prox γ R ■ (x) {\displaystyle \operatorname {prox} _{\gamma R}(x)} is defined entrywise by

which is known as the soft thresholding operator S γ (x) = prox γ ■ · ■ 1 ■ (x) {\displaystyle S_{\gamma }(x)=\operatorname {prox} _{\gamma \|\cdot \|_{1}}(x)} . [1] [6]

Fixed point iterative schemes

To finally solve the lasso problem we consider the fixed point equation shown earlier:

Given that we have computed the form of the proximity operator explicitly, then we can define a standard fixed point iteration procedure. Namely, fix some initial w 0 ∈ R d {\displaystyle w^{0}\in \mathbb {R} ^{d}} , and for k = 1 , 2 , … {\displaystyle k=1,2,\ldots } define

Note here the effective trade-off between the empirical error term F (w) {\displaystyle F(w)} and the regularization penalty R (w) {\displaystyle R(w)} . This fixed point method has decoupled the effect of the two different convex functions which comprise the objective function into a gradient descent step (w k − γ ∇ F (w k) {\displaystyle w^{k}-\gamma \nabla F\left(w^{k}\right)}) and a soft thresholding step (via S γ {\displaystyle S_{\gamma }}).

Convergence of this fixed point scheme is well-studied in the literature [1] [6] and is guaranteed under appropriate choice of step size γ {\displaystyle \gamma } and loss function (such as the square loss taken here). Accelerated methods were introduced by Nesterov in 1983 which improve the rate of convergence under certain regularity assumptions on F {\displaystyle F} . [7] Such methods have been studied extensively in previous years. [8] For more general learning problems where the proximity operator cannot be computed explicitly for some regularization term R {\displaystyle R} , such fixed point schemes can still be carried out using approximations to both the gradient and the proximity operator. [4] [9]

Practical considerations

There have been numerous developments within the past decade in convex optimization techniques which have influenced the application of proximal gradient methods in statistical learning theory. Here we survey a few important topics which can greatly improve practical algorithmic performance of these methods. [2] [10]

Adaptive step size

In the fixed point iteration scheme

one can allow variable step size γ k {\displaystyle \gamma _{k}} instead of a constant γ {\displaystyle \gamma } . Numerous adaptive step size schemes have been proposed throughout the literature. [1] [4] [11] [12] Applications of these schemes [2] [13] suggest that these can offer substantial improvement in number of iterations required for fixed point convergence.

Elastic net (mixed norm regularization)

Elastic net regularization offers an alternative to pure ■ 1 {\displaystyle \ell _{1}} regularization. The problem of lasso (■ 1 {\displaystyle \ell _{1}}) regularization involves the penalty term R (w) = ■ w ■ 1 {\displaystyle R(w)=\|w\|_{1}} , which is not strictly convex. Hence, solutions to min w F (w) + R (w) , {\displaystyle \min _{w}F(w)+R(w),} where F {\displaystyle F} is some empirical loss function, need not be unique. This is often avoided by the inclusion of an additional strictly convex term, such as an ■ 2 {\displaystyle \ell _{2}} norm regularization penalty. For example, one can consider the problem

where x i ∈ R d and y i ∈ R . {\displaystyle x_{i}\in \mathbb {R} ^{d}{\text{ and }}y_{i}\in \mathbb {R} .} For 0 < μ ≤ 1 {\displaystyle 0<\mu \leq 1} the penalty term λ ((1 − μ) ■ w ■ 1 + μ ■ w ■ 2 2) {\displaystyle \lambda \left((1-\mu)\|w\|_{1}+\mu \|w\|_{2}^{2}\right)} is now strictly convex, and hence the minimization problem now admits a unique solution. It has been observed that for sufficiently small μ > 0 {\displaystyle \mu >0} , the additional penalty term μ ■ w ■ 2 2 {\displaystyle \mu \|w\|_{2}^{2}} acts as a preconditioner and can substantially improve convergence while not adversely affecting the sparsity of solutions. [2] [14]

Exploiting group structure

Proximal gradient methods provide a general framework which is applicable to a wide variety of problems in statistical learning theory . Certain problems in learning can often involve data which has additional structure that is known a priori . In the past several years there have been new developments which incorporate information about group structure to provide methods which are tailored to different applications. Here we survey a few such methods.

Group lasso

Group lasso is a generalization of the lasso method when features are grouped into disjoint blocks. [15] Suppose the features are grouped into blocks { w 1 , … , w G } {\displaystyle \{w_{1},\ldots ,w_{G}\}} . Here we take as a regularization penalty

which is the sum of the ■ 2 {\displaystyle \ell _{2}} norm on corresponding feature vectors for the different groups. A similar proximity operator analysis as above can be used to compute the proximity operator for this penalty. Where the lasso penalty has a proximity operator which is soft thresholding on each individual component, the proximity operator for the group lasso is soft thresholding on each group. For the group w g {\displaystyle w_{g}} we have that proximity operator of λ γ (∑ g = 1 G ■ w g ■ 2) {\displaystyle \lambda \gamma \left(\sum _{g=1}^{G}\|w_{g}\|_{2}\right)} is given by

where w g {\displaystyle w_{g}} is the g {\displaystyle g} th group.

In contrast to lasso, the derivation of the proximity operator for group lasso relies on the Moreau decomposition . Here the proximity operator of the conjugate of the group lasso penalty becomes a projection onto the ball of a dual norm . [2]

Other group structures

In contrast to the group lasso problem, where features are grouped into disjoint blocks, it may be the case that grouped features are overlapping or have a nested structure. Such generalizations of group lasso have been considered in a variety of contexts. [16] [17] [18] [19] For overlapping groups one common approach is known as latent group lasso which introduces latent variables to account for overlap. [20] [21] Nested group structures are studied in hierarchical structure prediction and with directed acyclic graphs . [18]

See also

Convex analysis

Proximal gradient method

Regularization

Statistical learning theory

References