Conditional gradient algorithms for machine learning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We consider the problem of optimizing learning objectives with a regularization penalty in high-dimensional settings. For several important learning problems, state-of-the-art optimization approaches such as proximal gradient algorithms are difficult to apply and do not scale up to large datasets. We propose new conditional-type algorithms, with theoretical guarantees, for penalized learning problems. Promising experimental results are presented on two large-scale real-world datasets.

1 Introduction

We consider statistical learning problems that can be framed into convex optimization problems of the form

$$\min_{x \in K} \quad f(x) + \kappa \|x\| \tag{1.1}$$

where f is a convex function with Lipschitz continuous gradient and K is a closed convex cone of a Euclidean space. This encompasses supervised classification, regression, ranking, etc. with varying choices for the regularization penalty including regular ℓ_p-norms, as well as more structured regularization penalties such as the group-lasso penalty or the trace-norm [15]. A wealth of convex optimization algorithms was proposed to tackle such problems; see [15] for a recent overview. Among them, the celebrated Nesterov optimal gradient method for smooth and composite minimization [9, 10], and their stochastic approximation counterparts [8], are now state-of-the-art in machine learning.

These algorithms enjoy the best possible theoretical complexity estimate. For instance, Nesterov's algorithm reaches a solution with accuracy ϵ in $O(D\sqrt{L/\epsilon})$, where L is the Lipschitz constant of the gradient of the smooth component of the objective, and D is the problem-domain parameter in the norm $\|\cdot\|$. However, these algorithms work by solving at each iteration a sub-problem which involves the minimization of the sum of a linear form and a distance-generating function on the problem domain. Most often, and this explains the popularity of this family of algorithms in the last decade, exactly solving the sub-problem is computationally cheap [1].

However, in high-dimensional problems such as multi-task learning with a large number of tasks and features, solving this sub-problem is computationally challenging when the trace-norm regularization is used for instance [7, 12]. These limitations recently motivated alternative approaches, which do not require to solve hard sub-problems at each iteration, and triggered a renewed interest in conditional gradient algorithms. The *conditional gradient algorithm* [5, 3] (a.k.a. Frank-Wolfe algorithms) works by minimizing a linear form on the problem domain at each iteration, a simpler subproblem than the one involved in proximal gradient algorithms. Instances of the conditional algorithm were considered recently in the machine learning community [13, 14, 4], with new variants proposed in [6]. However, the conditional gradient algorith allows to tackle *constrained formulations* of learning problems, whereas *penalized formulations* are more widespread in applications.

The contributions of this work may be summarized as follows:

- We discuss a new conditional gradient-type algorithm for penalized formulations of learning problems, with theoretical accuracy guarantees
- We present some experimental results on real-world datasets which show that the proposed algorithms are also numerically valid – they outperform some state-of-the-art approaches in large-scale settings

The proposed algorithms apply to a large spectrum of learning problems.

2 Problem statement

Throughout the paper, we shall assume that $K \subset E$ is a closed convex cone of Euclidean space E; we do not loose anything by assuming that K linearly spans E. We assume that $\|\cdot\|$ is a norm on E, and f is a convex function with Lipschitz continuous gradient, that is $\|f'(x) - f'(y)\|_* \leq L_f\|x - y\| \ \forall x, y \in K$, where $\|\cdot\|_*$ denotes the dual norm dual of $\|\cdot\|$. We consider two kinds of problems, detailed below.

Penalized learning formulation We consider penalized learning problems of general form

$$\mathrm{Opt} = \min_x \left\{ f(x) + \kappa\|x\| : \ x \in K \right\} \ . \tag{2.2}$$

which we rewrite as

$$\mathrm{Opt} = \min_{x,r} \left\{ F([x;r]) = \kappa r + f(x) : x \in K, \|x\| \leq r \right\}. \tag{2.3}$$

We shall refer to (2.3) as the problem of *composite optimization* (CO). Note that, the x-component of an ϵ-solution (x_ϵ, r_ϵ) to (2.3) is an ϵ-solution to (2.2).

Minimization oracle We assume that for any $x \in E$ a first order "oracle" is available for f, namely, that a vector $w(x)$ is available such that $\|w(x) - f'(x)\|_* \leq \upsilon$, and the bound υ on the error in gradient is known. We assume that an exact observation of $f(x)$ is available. [1] We are going to use the conditional gradient algorithm (CG) as our working horse. As it was already mentioned in the introduction, utilizing CG hinges upon an inner problem, which consists in minimizing a linear form on the problem domain. As soon as the minimization can be performed efficiently, CG becomes an attractive alternative to proximal gradient algorithms. We state this as an assumption on the existence of a *minimization oracle*, as detailed below.

(A) [Minimization oracle] Given $\eta \in E_*$, we can find an optimal solution $x[\eta]$ to the optimization problem

$$\min_x \left\{ \langle \eta, x \rangle : \|x\| \leq 1, \ \ x \in K \right\}.$$

We do not sacrifice anything by assuming that for every η, $x[\eta]$ is either zero, or a vector of the $\|\cdot\|$-norm equal to 1. To ensure this, it is unnecessary to compute $\|x[\eta]\|$. Indeed, it suffices to compute $\langle \eta, x[\eta] \rangle$; if this product is 0, we can reset $x[\eta] = 0$, otherwise $\|x[\eta]\|$ is just equal to 1.

3 Conditional-gradient-like algorithm for penalized learning problems

We assume that there is an *a priori* known upper bound D^+ on $\|x_*\|$, x_* being an optimal solution to (2.2).

(B) [Upper bound] There exists $D < \infty$ such that $\kappa r + f(x) \leq f(0)$ together with $\|x\| \leq r$ imply that $r \leq D$. Further, we assume that a finite upper bound D^+ on D is available.

[1] A close inspection of the proofs of the results below reveals that the discussed algorithms are robust with respect to errors in observation of $f(x)$. In other words, a bound ν on the error in $f(x)$ results in an extra term $O(1)\nu$ in the corresponding accuracy bounds. However, to streamline the presentation we do not discuss this modification here.

Further, there is another upper bound D on $\|x_*\|$, “induced by the problem data” (but not available prior to solving (2.2). An important property of the proposed algorithm is that its accuracy guarantees, provided in Proposition 3.1 below, are expressed in terms of D, but not in terms of D^+ which may be very loose.

We now describe our algorithm for solving (2.3)

$$\mathrm{Opt} = \min_{x,r} \{F([x;r]) = \kappa r + f(x) : x \in K, \|x\| \le r\}\,.$$

Let $E^+ = E \times \mathbf{R}$, and $K^+ = \{[x;r] :\ x \in K,\ r \ge \|x\|\}$. In the remainder of the paper, we shall use the notation $z := [x;r]$, and for $z = [x;r]$ we put $x(z) := x$, $r(z) := r$. We also denote $K^+[\rho]$ the set

$$K^+[\rho] = \left\{[x;\rho] \in K^+ :\ \|x\| \le \rho\right\}.$$

Given $z = [x;r] \in K^+$, let us look at the linear form

$$\zeta = [\xi;\rho] \to \langle f'(x),\, \xi\rangle + \kappa\rho = \langle F'(z),\, \zeta\rangle.$$

For every $\rho \ge 0$, the minimum of this form on $K^+[\rho]$ is attained at the point $[\rho x[f'(x)];\, \rho]$. As ρ varies, these points form a ray. Let

$$\Delta(z) = \{\rho[x[f'(x)];1] :\ 0 \le \rho \le D^+\}$$

be the segment of this ray. By Assumption (A), given $z = [x;r]$, we can identify $\Delta(z)$ via a single call to the first order oracle for f and a single call for the minimization oracle $x[\cdot]$ for $(K, \|\cdot\|)$. Now let the point $z^*(z) = [x^*;r^*]$ be a minimizer of $F(\cdot)$ over the convex hull of $\Delta(z)$ and z, which is itself the convex hull of the origin and the points z and $D^+[x[f'(x)];1]$:

$$\begin{array}{rcl} [x^*;r^*] &=& \lambda_* D^+[x[f'(x)];1] + \mu_*[x;r], \quad \text{where} \\ (\lambda_*, \mu_*) &\in& \mathrm{Argmin}_{\lambda,\mu} \left\{ F(\lambda D^+[x[f'(x)];1] + \mu[x;r]) :\ \lambda + \mu \le 1,\ \lambda \ge 0, \mu \ge 0 \right\}. \end{array} \tag{3.4}$$

The iteration of the Conditional Gradient algorithm for composite optimization (COCG) algorithm is defined according to

$$z_{t+1} \in \mathop{\mathrm{Argmin}}_{z} \{F(z) :\ z \in \mathrm{Conv}\{0, z_t,\ D^+[x[f'(x(z_t))];1]\},\ \ t = 1,2,...,\tag{3.5}$$

where we put $z_1 = 0$.

Let $z_* = [x_*;r_*]$ be an optimal solution to (2.3), with $F_* = F(z_*)$.

Proposition 3.1. *The algorithm based on the above recursion is a descent algorithm. In addition, we have*

$$F(z_t) - F_* \le 8L_f D^2 (t+1)^{-1},\ k = 2,3,...$$

Remarks The recursion can be modified to obtain a bundle version of COCG. Let for some $M \in \mathbf{N}_+$,

$$\mathcal{C}_t = \left\{ \begin{array}{ll} \mathrm{Conv}\{0;\ D^+[x[f'(0)];1],\ ...,\ D^+[x[f'(x(z_t))];1]\}, & \text{for } t \le M, \\ \mathrm{Conv}\{0; z_{t-M+1}, ...,\ z_t;\ D^+[x[f'(x(z_{t-M+1}))];1], ...,\ D^+[x[f'(x(z_t))];1]\}, & \text{for } t > M. \end{array} \right. \tag{3.6}$$

For $t = 1,2,...$ set

$$z_{t+1} \in \mathop{\mathrm{Argmin}}_{z} \{F(z) :\ z \in \mathcal{C}_t\}.\tag{3.7}$$

Results of Proposition 3.1 still hold for $(z_t)_{t=1,2,...}$ defined in (3.6), (3.7).

The basic implementation of the COCG requires solving a two-dimensional auxiliary problem (3.4) at each iteration. To solve the problem one can utilize, for instance, the ellipsoid method, the first order information about (3.4) being supplied by the first-order oracle for f.

4 Experiments

We conducted experiments on two real-world datasets corresponding resp. to our two working examples, i.e., matrix completion with noise and multi-class classification with logistic loss: i) MovieLens10M, ii) ImageNet dataset. On both datasets, we compared the proposed algorithm to the accelerated version of the proximal gradient algorithm, denoted by the acronym Prox-Grad on the figures.

	Algo.	Time	Training error	Test error
MovieLens10M	Prox-Grad	42.37s	0.64	0.86
	COCG	19.15s	0.63	0.86
ImageNet	Prox-Grad	129.43hrs	0.51	0.63
	COCG	47.94hrs	0.51	0.60

Table 1: Comparison of the COCG algorithm with the proximal gradient algorithm on the MovieLens10M and the ILSVRC2010 ImageNet dataset. Performance is measured in RMSE on the MovieLens10M dataset and misclassification error on the ILSVRC2010 ImageNet dataset. Note that the time is measured in *seconds* on the MovieLens10M dataset while it is measured in *hours* on the ILSVRC2010 ImageNet dataset.

MovieLens10M We utilise the COCG algorithm, described in Section 3, on the MovieLens10M dataset, corresponding to 10^7 ratings of 69878 users on 10677 movies. Same as in [6], we use half of the sample for training. Note that, because of the inherent sparsity of the problem, the matrix-vector computations can be handled very efficiently. Yet, experiments on this dataset are interesting, as they show that our COCG approach is competitive with the state-of-the-art proximal gradient algorithm. A lower value of the objective is reached within a slightly shorter time than the proximal gradient algorithm.

Imagenet We test the COCG algorithm on an image categorization application. We consider the Pascal ILSVRC2010 ImageNet dataset and focus on the "Vertebrate-craniate" subset, yielding 2943 classes with and average of 1000 examples each. We compute $65,000$-dimensional visual descriptors for each example using Fisher vector representation [11], a state-of-the-art feature vector representation in image categorization. Note that, in contrast to the previous experiment, this representation yields *dense* feature vectors. Hence making the matrix-vector computations efficient is a challenging task. We can leverage here one of the feature of our COCG algorithm, that is, the robustness of gradient computation which allows making inexact objective and gradient evaluations on mini-batches of examples. In order to save enough RAM space to perform matrix-vector computations at each iteration, we set the size of the mini-batches to $b = 5000$. Note that the "raw" batch proximal gradient algorithm is inappropriate for such a large dataset, where no structure can be leveraged to speed-up the computations. The straightforward "stochastic gradient" version of the proximal algorithm is out of scope, as it would involve an SVD per example. Thus we use as a contender the mini-batch version of the proximal gradient algorithm [2], and we use the same reasoning to set the mini-batch size of the proximal gradient algorithm. Results in Table 1 show that our COCG algorithm has an edge over the mini-batch proximal gradient algorithm and manages to achieve a similar value of the objective significantly faster.

5 Conclusion

We discuss new conditional gradient-type algorithms for penalized formulations of learning problems. These algorithms feature good theoretical guarantees and promising experimental results.

References

[1] F. R. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 2012.

[2] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *NIPS*, 2011.

[3] V. Demyanov and A. Rubinov. *Approximate Methods in Optimization Problems*. American Elsevier, 1970.

[4] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS*, 2012.

[5] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[6] M. Jaggi and M. Sulovský. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.

[7] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

[8] G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 2012.

[9] Y. Nesterov. *Introductory lectures on convex optimization. A basic course.* Kluwer, 2004.

[10] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, CORE Discussion Paper, 2007.

[11] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2006.

[12] T. K. Pong, S. J. Paul Tseng, and J. Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM J. Optimization*, 20(6):3465–3489, 2010.

[13] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. In *ICML*, 2011.

[14] C. Shen, J. Kim, L. Wang, and A. van den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *JMLR*, 2012.

[15] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, 2010.