Primal-Dual Algorithms for Subquadratic Norms

Raman Sankaran
Dept. of Computer Science & Automation
Indian Institute of Science, Bangalore
ramans@csa.iisc.ernet.in

Francis Bach
INRIA - Sierra Project-team
École Normale Supérieure, Paris, France
francis.bach@ens.fr

Chiranjib Bhattacharyya
Dept. of Computer Science & Automation
Indian Institute of Science, Bangalore
chiru@csa.iisc.ernet.in

Abstract

Reweighted least-squares formulations of Subquadratic norms appear naturally in machine learning and signal processing. The traditional algorithms associated with these formulations either do not have convergence guarantees or have large per-step computation costs. In this paper, we derive primal-dual algorithms with convergence rate $O(1/T)$ using the relevant proximal operators of the subquadratic norm. This makes it possible to solve problems with subquadratic norms resulting from wedge penalty or the Lovász extension of combinatorial penalties, for which there are no efficient algorithms available.

1 Introduction

We consider optimization problems in w and η of the following form, which are also known as reweighted least-squares formulations [1].

$$\min_{w\in\mathbb{R}^d,\eta\in\mathbb{R}^d_+} F(Xw) + \tfrac{\lambda}{2}\textstyle\sum_{j=1}^d \frac{w_j^2}{\eta_j} + \frac{\lambda}{2}\Gamma(\eta). \tag{1}$$

$X \in \mathbb{R}^{n\times d}$ is the data matrix, and λ the regularization parameter. These are special instances of the general framework $\min_w F(Xw) + \lambda\Omega(w)$ with $\Omega(w) = \min_{\eta\geq 0}\sum_{i=1}^d \frac{w_i^2}{\eta_i} + \Gamma(\eta)$ which are known as subquadratic norms [2]. Subquadratic norms of these forms arise frequently in the context of structured sparsity[2, 3, 4] with various applications in bioinformatics, image, text and audio processing [2, 5]. For such norms, existing algorithms to solve 1 essentially require to know the proximal operator for $F\circ\tilde{X} : w \mapsto F(\tilde{X}w)$ in each iteration where $\tilde{X} = X\ \mathrm{Diag}^{\frac{1}{2}}(\eta)$[1] (with the value of η at that iteration). This is costly in many situations (except for the square loss and piecewise-affine losses like the hinge loss). The simplest algorithm is the alternating optimization (which has convergence issues in general because the objective functions is not smooth at 0). The only generic first-order algorithm is subgradient descent, which is very slow. The lack of efficient solutions for this primal problem leads us to reformulate the problem 1 as a saddle point problem, for which primal-dual algorithms[6] can be used which have provable guarantees in the convergence rates. We make the following contributions in this paper.

- We propose two primal dual algorithms with convergence rate $O(1/T)$ to solve problem (1).
- In section 4.1, we propose a primal-dual algorithm with a convergence rate $O(1/T)$ to solve problem 1, when the proximal operator for $F\circ X$ is known. In general, evaluating $\mathrm{prox}_{F\circ X}$ may

[1] For a vector a, Diag(a) refers to the diagonal matrix with a as diagonal entries.

be expensive because of the dependence on X. However for the square loss, this computation turns out to be very simple.

– In section 4.2, we propose another primal-dual algorithm with convergence rate $O(1/T)$. This algorithm is applicable for all losses and norms such that prox_F and prox_Γ is available. Comparing with the previous algorithm, for the commonly used loss functions, prox_F is much simpler to obtain than $\mathrm{prox}_{F\circ X}$ since it does not depend on X. This algorithm uses the data matrix X only in terms of matrix-vector multiplications thus being a first order algorithm. Hence this becomes the first ever generic algorithm to solve problem (1) for all loss functions and norms. However if $\mathrm{prox}_{F\circ X}$ is easy to compute, it turns out that the previous algorithm will be more efficient because of the independence of the algorithm from the data matrix.

– We apply the primal dual algorithm for the cases of Γ being the wedge penalty, and the Lovász extension of a combinatorial penalty. There has not been an efficient algorithms with provable guarantees in convergence rates for these problems, which are instances of the problem (1). Our experiments show clear improvement over the state-of-the-art solutions for these problems.

2 Examples of Subquadratic norms

We review the examples of the function Γ in (1) derived from the wedge penalty and combinatorial penalties, for which no efficient algorithm is available.

Wedge penalty. By adding ordering constraints on the vector η, we may get norms that favor certain sparsity patterns. The simplest example is the wedge penalty, so that:

$$\Gamma(\eta) = \eta^\top 1_d + I_{\eta_1 \geqslant \cdots \geqslant \eta_d \geqslant 0}(\eta). \tag{2}$$

This penalty which mandates η to be a decreasing vector in its coordinates, encourages the model w to be of the same pattern. As studied in [3], prox_Ω is difficult to compute whereas prox_Γ is easily obtained from isotonic regression in $O(d)$ time by the pool-adjacent-violator algorithm [7].

Convex relaxation of combinatorial penalties [4]. We consider $\Gamma(\eta) = f(\eta)$ where f is the Lovász extension of a non-decreasing submodular function [8, 9]. These norms extend the grouped ℓ_2-norms by favoring certain sparsity patterns over others. In order to satisfy the constraint above, we impose that $\inf_{A\subset V} f(1_A)/|A| = 1$. As shown in [9], the proximal operator may be easily obtained from the corresponding least-squares problem (by thresholding), for which many algorithms exist (see, e.g., [10] for cut-based functions).

3 Chambolle-Pock's primal dual algorithm [6]

In this section, we set up the primal dual framework by following the presentation of Chambolle and Pock [6], and consider a generic problem of the form

$$\min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} v^\top K u + G(u) - H^*(v), \tag{3}$$

with a proper strongly convex function $H^* : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ with strong convexity parameter $\gamma \geqslant 0$ and a proper convex function $G : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$, for which we can compute the proximal operators $\mathrm{prox}^\sigma_{H^*}$ and prox^τ_G for any positive σ and τ. The primal dual algorithm (CP) detailed as Algorithm 2 in [6] solves the above defined problem with convergence rate $O(1/T^2)$ if either G or F^* is strongly convex. If both are not strongly convex, the rate reduces to $O(1/T)$. It is to be noted CP only needs black box operators needed for prox_{H^*} and prox_G, and accesses K only through matrix-vector products, and hence strictly first order in nature. For the special case $K = I$, CP is shown to be equivalent to ADMM. This algorithm requires initializing the primal and dual stepsizes σ and τ respectively. We initialize them as

$$\sigma = \frac{1}{\|K\|}\frac{\|v^* - v_0\|_2}{\|u^* - u_0\|_2}, \tau = \frac{1}{\|K\|}\frac{\|u^* - u_0\|_2}{\|v^* - v_0\|_2}. \tag{4}$$

The derivation for these choices are omitted for the sake of space. Since it is not possible to calculate σ and τ exactly in practice, we will compute rough estimates for them using the information we have.

4 Primal-Dual Formulation for Subquadratic Norms

We can reformulate (1) as

$$\min_{w\in\mathbb{R}^d,\ \eta\in\mathbb{R}^d,\ t\in\mathbb{R}^d} F(Xw) + \tfrac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t + \textstyle\sum_{j=1}^d I_K(w_j, \eta_j, t_j), \tag{5}$$

where $K = \{(a, b, c) \in \mathbb{R}^3,\ a^2 \leqslant 2bc, b \geqslant 0, c \geqslant 0\}$ is the rotated second order-cone. which can naturally be split in two simple terms $F(Xw) + \frac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t$ and $\sum_{j=1}^d I_K(w_j, \eta_j, t_j)$. The cone K is self-dual and the proximal operator for $\sum_{j=1}^d I_K(w_j, \eta_j, t_j)$ involves computing the orthogonal projection into K. The projection of (a, b, c) into K may be obtained by computing the positive part of the 2×2-matrix $\begin{pmatrix} b & a/\sqrt{2} \\ a/\sqrt{2} & c \end{pmatrix}$. Depending whether the term $F(Xw) = F \circ X(w)$ is left as is or not, we obtain a second-order algorithm ADMM (that accesses X through linear systems) or a first-order algorithm Chambolle-Pock (that accesses X by matrix-vector products).

4.1 ADMM formulation of problem (1)

By Fenchel duality, Eq. (5) may be cast as the saddle-point problem

$$\min_{w \in \mathbb{R}^d,\ \eta \in \mathbb{R}^d,\ t \in \mathbb{R}^d} \max_{\kappa \in K^d} \Big(F(Xw) + \frac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t \Big) - \sum_{j=1}^d I_K(\kappa_j) + \sum_{j=1}^d \kappa_j^\top (w_j, \eta_j, t_j). \quad (6)$$

We can equate this formulation to (3) by denoting $u = [(w_j, \eta_j, t_j)]_{j=1}^d$, $v = \kappa = [(\delta_j, \beta_j, \nu_j)]_{j=1}^d$, $K = -I$, $G(u) = F(Xw) + \frac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t$ and $H^*(v) = \sum_{j=1}^d I_K(\kappa_j)$. Computing prox_G involves computing the proximal operator for $F \circ X$. Since $K = -I$, the CP algorithm is invariant to the data matrix X. We refer to the application of CP to this instance as "ADMM-η". From Eq. (4), it is easy to see that we can derive estimates of the stepsizes for for the case of Square loss and ℓ_1 norm combination as as $\sigma = \|X\|^2/\sqrt{3}n$ and $\tau = \sqrt{3}n/\|X\|^2$. In our experiments, these stepsize choices work well for the other norms and losses.

4.2 Chambolle-Pock

The previous reformulation (6) may not be applicable when it is difficult to compute $\mathrm{prox}_{F \circ X}$. Hence we reformulate (6) using Fenchel's duality for F as

$$\min_{w \in \mathbb{R}^d,\ \eta \in \mathbb{R}^d,\ t \in \mathbb{R}^d} \max_{\alpha \in \mathbb{R}^n,\ \kappa \in K^d} \alpha^\top Xw - F^*(\alpha) + \frac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t + r \sum_{j=1}^d \kappa_j^\top (w_j, \eta_j, t_j) - r \sum_{j=1}^d I_K(\kappa_j) \quad (7)$$

This is also an instance of the problem in Eq. (3) with the following choices: $u = [(w_j, \eta_j, t_j)]_{j=1}^d$, $v = (\alpha, \kappa)$, $K = [X, rI]$, $G(u) = \frac{\lambda}{2}\Gamma(\eta) + \lambda 1_d^\top t$ and $H^*(v) = F^*(\alpha) + r \sum_{j=1}^d I_K(\kappa_j)$.

Comparing this with ADMM-η, we only require prox_F which is much easier to obtain than $\mathrm{prox}_{F \circ X}$. We refer to this algorithm as "CP-η". This emerges as the most general algorithm with least assumptions on the functions involved. In this formulation, $\|K\|^2 = \|X^\top X + r^2 I\|^2 \leqslant \|X\|^2 + r^2$. We have introduced the constant r to balance the scale of α, against κ, with $r = \|X\|$. Similar to the previous case, we can derive stepsize estimates as $\sigma = 1/n$ and $\tau = n/(2\|X\|^2)$.

5 Experiments

The algorithms for the problem (1) which we use for comparison purposes are (a) Alternating minimization between w and η denoted as "Alt-η", (b)"FISTA-η", which represents the (1) as a function of η similar to SimpleMKL [11], and (c) generic subgradient descent algorithm ("Subgrad"). We compared the results for the settings of low and high correlations among the columns of X[2]. We have fixed $n = d = 5000$, $\lambda = 0.01$ and set the stopping criterion to be the certified duality gap less than the convergence threshold of 10^{-3}.

5.1 Wedge penalty

When the norm Ω is the wedge penalty, the usual algorithms which access prox_Ω are not applicable. We consider the Square loss and the Hinge loss as examples of smooth and non-smooth loss functions. The plots of the objective value versus iterations is given for the Hinge loss in Figure 1.

Square loss. Here the ADMM-η algorithm requires solving a linear system, which is constant across iterations and hence compares against CP-η and subgradient descent as a first order algorithm. For ADMM-η running times do not vary across the two different settings of correlation (which is true

[2]The details of experiments are omitted here for the sake of space.

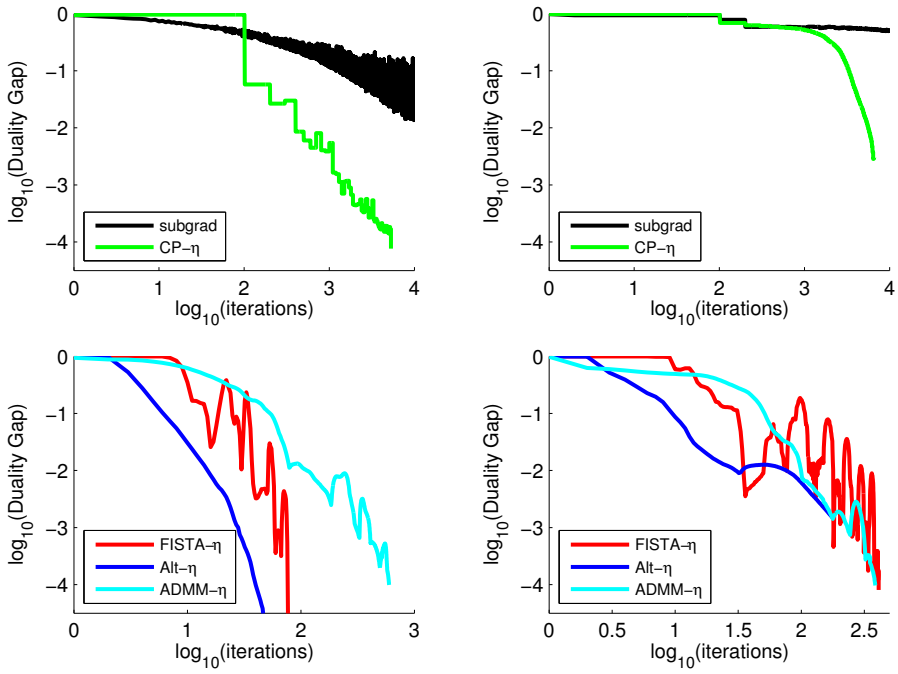

Figure 1: Primal dual gap convergence for the hinge loss and wedge penalty combination. Top: first-order algorithms. Bottom: second-order algorithms. Left: low correlation. Right: high correlation.

for the other loss / norms also). Second-order algorithms Alt-η and FISTA-η are similar to the Lasso case, will have very high per-step cost, as shown in the table below which shows the running time in seconds. Overall, our new primal-dual algorithms outperform existing ones.

Loss	Corr.	Alt-η	FISTA-η	ADMM-η	subgrad	CP-η
Square	Low	232	1997	165	-	121
Square	High	2510	4187	277	-	830
Hinge	Low	245	2893	2765	-	269
Hinge	High	2721	9632	2043	-	2954

Hinge loss. When we move to the hinge loss, CP-η is the only available first order algorithm, which gives $O(1/T)$ compared to $O(1/\sqrt{T})$ given by subgradient descent. The plots are given in Figure 1 and the corresponding with running times (in seconds) are given in the above table.

5.2 Combinatorial penalty

We choose the submodular function $S(A) = \max(A)$. The Lovász extension $f(\eta) = \sum_{i=1}^{d} \|\eta(i : d)\|_\infty$. $\Gamma(\eta) = f(\eta)$ leads to a subquadratic norm, which can be proven to be the same norm which we get using the Wedge penalty. And as shown in [12], the proximal operator of Γ can be derived using composition of the proximal operator for the ℓ_∞ norm. We show below the corresponding running times of the algorithms.

Loss	Corr.	Alt-η	FISTA-η	ADMM-η	subgrad	CP-η
Square	Low	183	2512	210	-	152
Square	High	1980	4221	255	-	798
Hinge	Low	202	3103	2976	-	220
Hinge	High	1962	9925	2563	-	3032

6 Conclusion

In this paper, we showed how primal-dual algorithms could be extended to reweighted least-squares formulations, with simple algorithms improving over the state of the art.

References

[1] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.

[3] C. Micchelli, J. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.

[4] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Technical Report 00694765, HAL, 2012.

[5] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, May 2011.

[7] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

[8] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

[9] F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6-2-3:145–373, 2011.

[10] M. Babenko, J. Derryberry, A. Goldberg, R. Tarjan, and Y. Zhou. Experimental evaluation of parametric max-flow algorithms. In *Proceedings of the International Conference on Experimental algorithms (WEA)*, 2007.

[11] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[12] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.*, 12:2297–2334, July 2011.