Stochastic Dual Coordinate Descent

Guillaume Desforges & Michaël Karpe & Matthieu Roux

June 14th 2018

1 Introduction

In machine learning, the process of fitting an algorithm often requires to solve some optimization problem, more or less complex. Most of the time, it consists in minimizing an empirical risk (the case we have studied in class), but there exists other ways of learning. The difficulty resides in the fact that those optimization problems are often very complex. Their solutions are often neither exact, nor a solution that we can obtain with a closed form. In addition, in real use cases the computation of the solution is often hard.

Because the problems we work with have a great amount of data and dimensions, it can be interesting to have a look on the dual problem. In fact, looking at the dual problem can lead to easier problems to solve, with interesting solving methods different from the ones used to solve the primal problems. Moreover, if we work in a a case where we have strong duality it will lead to the optimal solution of the primal problem. In this paper, we will study this approach to obtain another optimization problem, with a different structure from the primal one. Therefore, the goal of the project will be to study and implement the method of stochastic coordinate ascent (descent is equivalent but we will stick to the litterature) to solve the dual problem.

First we introduce the general forms of the machine learning problems we want to apply SDCA on. Then we develop on how we get to the dual form, and how it can be solved. Then we try to compare the theoretical advantages of solving dual problem instead of the primal one, and make sure we have guarantees over the strong duality. Then we study computational performances of the method by trying to apply SDCA on concret problems. Finally we conclude on SDCA strengths and weaknesses.

2 The dual problem of the logistic regression

In this paper, we study the SDCA algorithm on several problems. The logistic regression is an interesting problem for an introduction. The multiclass logistic regression, with the cross entropy loss, is a generalization of the binary logistic regression that requires more work, so for now we stick to the binary problem which is interesting enough to present our work.

We use the following usual notations :

$X \in \mathbf{X} = \mathbb{R}^p$ the random variable for the description space;

$Y \in \mathbf{Y} = \{0, 1\}$ the random variable for the label.

We recall that the model is the following :

$$\frac{\mathbb{P}(y = 1|X = x)}{\mathbb{P}(y = -1|X = x)} = w^T x, \quad w \in \mathbb{R}^p \tag{1}$$

We want to find w so that it maximizes the likelihood, or log-likelihood, with a term of regularization:

$$\min_w C \sum_i \log(1 + e^{-y_i w^T x_i)} + \frac{1}{2} w^T w \tag{2}$$

In order to get the dual problem, we rewrite it with an artificial constraint $z_i = e^{-y_i w^T x_i}$, and we have the following lagrangian :

$$\mathcal{L}(w, z, \alpha) = \sum_i (C \log(1 + z_i) + \alpha_i z_i) - \sum_i \alpha_i e^{-y_i w^T x_i} + \frac{1}{2} w^T w \tag{3}$$

We will note w^* and z^* the variables solution of the optimization problem

$$\min_{w,z} \mathcal{L}(w, z, \alpha) = \mathcal{L}(w^*, z^*, \alpha) = \psi(\alpha) \tag{4}$$

Where, $w^* = \sum_i \alpha_i y_i x_i$.

It leads to the following dual problem :

$$\begin{array}{ll} & \max_\alpha \sum_{i \in I} (-\alpha_i \log(\alpha_i) - (C - \alpha_i) \log(C - \alpha_i)) - \frac{1}{2} \alpha^T Q \alpha \\ s.t. & I = \{i,\ 0 < \alpha_i < C\} \\ & 0 \leq \alpha_i \leq C \end{array} \tag{5}$$

We used the notation $Q = (Q_{ij})_{i,j}$ where $Q_{ij} = y_i x_i^T x_j y_j$

Now that we have the dual problem, we need to solve a maximize it. To do so, we will use in the coordinate ascent method, which consist in optimizing the objective function coordinate by coordinate (or with groups of coordinates).

3 Methods

(a) specific problem studied
(b) related work
(c) model, main idea
(d) specific methodology

(e) algorithms

Let $x_1, \ldots, x_n \in \mathbb{R}^d$, $\phi_1, \ldots, \phi_n$ scalar convex functions, $\lambda > 0$ regularization parameter. Let focus on the following optimization problem:

$$\min_{w \in \mathbb{R}^d} P(w) \quad \text{where } P(w) = \left[\frac{1}{n} \sum_{i=1}^{n} \phi_i(w^\top x_i) + \frac{\lambda}{2} \|w\|^2 \right] \tag{6}$$

with solution $w^* = \arg\min_{w \in \mathbb{R}^d} P(w)$. Moreover, we say that a solution w is ϵ_P-sub-optimal if $P(w) - P(w^*) \leq \epsilon_P$. We analyze here the required runtime to find an ϵ_P-sub-optimal solution using SDCA. (détailler + haut différence entre SGD et SDCA)

Let $\phi_i^* : \mathbb{R} \to \mathbb{R}$ be the convex conjugate of ϕ_i : $\phi_i^*(u) = \max_z(zu - \phi_i(z))$. The dual problem of (6) is defined as follows:

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha) \quad \text{where } D(\alpha) = \left[\frac{1}{n} \sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right] \tag{7}$$

with solution $\alpha^* = \arg\max_{a \in \mathbb{R}^n} D(\alpha)$. If we define $w(\alpha) = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i$, then we have $w(\alpha^*) = w^*$, $P(w^*) = D(\alpha^*)$, $\forall (w, \alpha), P(w) = D(\alpha)$ due to classic optimization results, and the duality gap $P(w(\alpha)) - P(w^*)$.

The SDCA algorithm is described below. T_0 can be chosen between 1 to T, and is generally chosen equal to $T/2$. However, in pratice, these parameters are not required as the duality gap is used to terminate the algorithm.

Algorithm 1 Procedure SCDA with averaging option

procedure SCDA($\alpha^{(0)}, \phi, T_0$)
  $w^{(0)} \leftarrow w(\alpha^{(0)})$
  for $t = 1, \ldots, T$ **do**
    Randomly pick i
    $\Delta\alpha_i \leftarrow \arg\max -\phi_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2} \left\| w^{(t-1)} + (\lambda n)^{-1} \Delta\alpha_i x_i \right\|^2$ (*)
    $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$
    $w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1} \Delta\alpha_i x_i$
  return $\overline{w} = \frac{1}{T - T_0} \sum_{i=T_0+1}^{T} w^{(t-1)}$

4 Experiments

(a) Description of the dataset(s) considered / general problem associated with the data
(b) Description of the protocol of the experiments (setting of the hyperparameters/cross-validation procedure/evaluation methodology)
(c) Factual description of the type of results reported (explanation pertaining to the Figures, tables, etc)
(d) Interpretation and discussion of the results (comparison with the baselines, advantages of each algorithm, etc)

In this study, SDCA is computed either for L-Lipschitz loss functions or for $(1/\gamma)$-smooth loss functions. We recall that a function $\phi_i : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz if $\forall a, b \in \mathbb{R}$, $|\phi_i(a) - \phi_i(b)| \leq L\,|a - b|$, and that a function $\phi_i : \mathbb{R} \to \mathbb{R}$ is $(1/\gamma)$-smooth if it is differentiable and its derivative is $(1/\gamma)$-Lipschitz. Moreover, if ϕ_i is $(1/\gamma)$-smooth, then ϕ_i^* is γ-strongly convex. The different loss functions used are described in the table below.

Squared loss: $\phi_i(a) = (a - y_i)^2$ $\phi_i^*(-a) = -ay_i + a^2/4$ $\Delta\alpha_i = \dfrac{y_i - x_i^\top w^{(t-1)} - 0.5\alpha_i^{(t-1)}}{0.5 + \|x_i\|^2/(\lambda n))}$	**Absolute deviation loss:** $\phi_i(a) = \|a - y_i\|$ $\phi_i^*(-a) = -ay_i,\ a \in [-1, 1]$ $\Delta\alpha_i = \max\left(1, \min\left(1, \dfrac{y_i - x_i^\top w^{(t-1)}}{\|x_i\|^2/(\lambda n)} + \alpha_i^{(t-1)}\right)\right) - \alpha_i^{(t-1)}$
Log loss: $\phi_i(a) = \log(1 + \exp(-y_i a))$ $\phi_i^*(-a) = -ay_i \log(ay_i) + (1 - ay_i)\log(1 - ay_i)$ $\Delta\alpha_i = \dfrac{(1 + \exp(x_i^\top w^{(t-1)} y_i))^{-1} y_i - \alpha_i^{(t-1)}}{\max(1, 0.25 + \|x_i\|^2/(\lambda n))}$	**(γ-smoothed) Hinge loss:** $\phi_i(a) = \max\{0, 1 - y_i a\}$ $\phi_i^*(-a) = -ay_i + \gamma a^2/2,\ ay_i \in [0, 1]$ $\Delta\alpha_i = y_i \max\left(0, \min\left(1, \dfrac{1 - x_i^\top w^{(t-1)} y_i - \gamma\alpha_i^{(t-1)} y_i}{\|x_i\|^2/(\lambda n) + \gamma} + \alpha_i^{(t-1)} y_i\right)\right) - \alpha_i^{(t-1)}$

Table 1: Used loss functions, convex conjugates and closed form of solutions of problem (*).

For the sake of simplicity, we consider the following assumptions: $\forall i, \|x_i\| \leq 1$, $\forall (i, a), \phi_i(a) \geq 0$ and $\forall i, \phi_i(0) \leq 1$. Under these assumptions, we have the following theorem:

Theorem Consider Procedure SDCA with $\alpha^{(0)} = 0$. Assume that $\forall i, \phi_i$ is L-Lipschitz (resp. $(1/\gamma)$-smooth). To obtain an expected duality gap of $\mathbb{E}[P(\overline{w}) - D(\overline{\alpha})] \leq \epsilon_P$, it suffices to have a total number of iterations of

$$T \geq n + \max\left(0, \left\lceil n \log\left(\frac{\lambda n}{2L^2}\right)\right\rceil\right) + \frac{20L^2}{\lambda\epsilon_P} \quad \left(\text{resp. } T > \left(n + \frac{1}{\lambda\gamma}\right)\log\left[\frac{1}{(T - T_0)\epsilon_P}\left(n + \frac{1}{\lambda\gamma}\right)\right]\right) \tag{8}$$

5 Conclusion

(a) summary
(b) main conclusions and take home messages
(c) Remaining questions/ future directions (only if relevant)