swissQuant
Mastering Complexity

Mathematical Challenge August 2017

Distributed Optimization

References

- [1] Boyd, Stephen, et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends® in Machine Learning* 3.1 (2011): 1-122.
- [2] Nedic, Angelia, and Asuman Ozdaglar. "Cooperative distributed multi-agent." *Convex Optimization in Signal Processing and Communications* 340 (2010).
- [3] Bertsekas, Dimitri P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014..

Description

Large scale optimization problems arise often due to a large number of optimization variables or, as in machine learning tasks, due to a large number of training data to fit the model on. Distributed optimization algorithms, where computations are performed in parallel relying only on local observations and information, are appealing to reduce computational time and sometimes, in case of lack of a centralized access to information for example, unavoidable.

More formally, following [2], consider the task of optimizing a global-objective function $T\big(f_1(w),..,f_N(w)\big)$:

$$\min\Big(T\big(f_1(w),..,f_N(w)\big)\Big), w \in \ (\cap\, C_i) \cap C$$

where each convex local-objective functions $f_i(w)$ and constraints set C_i may be known at the local node i only. The goal of the agents is to “cooperatively” find an approximate solution of the global-objective function.

Lagrangian dual ascent methods are an example of how a problem can re casted in a form suited to distributed optimization in the special case where problem is “separable”: $T\big(f_1(w),..,f_N(w)\big) = \sum_i^N f_i(w_i)$, $w \in C \equiv \ \big\{\sum_i^N h_i(w_i) \le b\big\}$, i.e. when the optimization takes the form

$$\min_w \sum\nolimits_i^N f_i(w_i)$$

$$\sum\nolimits_i^N h_i(w_i) \le b, w_i \in C_i$$

In this case the dual problem has the form

$$\max_{\lambda} g(\,\lambda)$$

where

$g(\lambda) = \min_{w} L(w, \lambda)$

$L(w, \lambda) \equiv \sum_i^N L_i(w_i, \lambda)\,, L_i(w_i, \lambda) = \ f_i(w_i) + \lambda^T(h_i(w_i) - b/N)$.

Applying the dual ascent method, which tries to identify the solutions $(w^* \equiv argmin_w L(w, \lambda^*\,),\ \lambda^* \equiv argmax_\lambda g(\lambda))$ iterating the update steps

$$\begin{cases} {w_i}^{\,k+1} = argmin_{w_{i\in C_i}} L_i\big(w_i, \lambda^k\big)\ (*) \\ \quad {r^{k+1}}_i = h_i({w_i}^{\,k+1}) - b/N \end{cases} \qquad \lambda^{k+1} = \lambda^k + \rho\ \sum_i^N {r^{k+1}}_i\ \text{(**)}$$

we obtain an algorithm that (under some, not necessarily mild, conditions) solve the original problem by distributing the dual λ^k to N nodes/workers (broadcast step), running in parallel the optimizations (*), collecting the residuals ${r^{k+1}}_i$ (gather step) and updating the dual (**) at the central node

When the problem is not separable the dual ascent method approach will not lead to distributed formulation. Even for separable problems actually often to improve the convergence of the algorithms the Lagrangian has to be augmented with a non-separable regularization term [3]. For such problems, an approach called global variable consensus algorithms [1] may be used, as we will now illustrate.

Large datasets

Consider an additive problem, such as the ones that arise while fitting a single model on a large set of data possibly stored in different locations S_i (in this case $f_i(w) = Loss(y_{n\in S_i}, Model(x_{n\in S_i}; w)))$

$\min_{w} \sum_i^N f_i(w)$

$w \in \cap\, C_i$

The problem is not separable, but can be rewritten in a separable form introducing a global variable z

$\min_{w_i} \sum_i^N f_i(w_i)$

$w_i = z, \forall i,\ z \in \cap\, C_i$

The problem can then be solved using the Alternating Direction Method of Multipliers (ADMM) [1], consisting in applying the an iterative gradient ascent schema for the augmented Lagrangian $\sum_i^N f_i(w_i) + \lambda_i(w_i - z) + \rho/2\|w_i - z\|_2^2$:

$w_i^{k+1} = argmin_{w_i}\left(f_i(w_i) + \lambda_i^k\big(w_i - \overline{w}^k\big) + \rho/2\big\|w_i - \overline{w}^k\big\|_2^2\right)$

$\lambda_i^{k+1} = \lambda_i^k + \rho(w_i^{k+1} - \overline{w}^{k+1})$

(Here and in the following $\bar{x}$ denotes the average of x)

Note that this approach can be extended to handle local variables sharing, local weighted averages and more general regularization functions. This is especially useful when the flow of information take place on a network with communication constraints [2].

Large number of features

Another problematic situation arises when the optimization is on a large number of variables in this case the problem has often the form (we assume a separable regularization term):

$$\min_{w_i} Loss\left(y, \sum_{i=1}^{B} w_i b_i(x)\right) + \sum_{i=1}^{B} r_i(w_i)$$

which can be reformulated as

$$\min_{w_i} Loss\left(y, \sum_{i=1}^{B} z_i\right) + \sum_{i=1}^{B} r_i(w_i)$$

$$w_i b_i(x) = z_i$$

Using ADMM we obtain the iterative and distributable method

$$w_i^{k+1} = argmin_{w_i} r_i(w_i) + \lambda_i^k\left(w_i b_i(x) - z_i^k\right) + \frac{\rho}{2}\left\|w_i b_i(x) - z_i^k\right\|_2^2$$

$$z^{k+1} = argmin_z\ Loss(y, \sum_{i=1}^{B} z_i)\ + \sum_{i=1}^{B}\left(\lambda_i^k\left(w^{k+1}{}_i b_i(x) - z_i\right) + \frac{\rho}{2}\left\|w_i^{k+1} b_i(x) - z_i^k\right\|_2^2\right)$$

$$\lambda_i^{k+1} = \ \lambda_i^k + \rho\left(w^{k+1}{}_i b_i(x) - z_i^{k+1}\right)$$

It can be shown (see [1]) that the second optimization problem boils down to a non-parametric regularized fitting problem:

$$\bar{z}^{k+1} = argmin_{\bar{z}}\ Loss(y, B\ \bar{z}) + \frac{\rho}{2} B\left\|\bar{z} - \bar{a}^{k+1}\right\|_2^2$$

$$z_i^{k+1} = \bar{z}^{k+1} + \left(a_i^{k+1} - \bar{a}^{k+1}\right),\ a_i^{k+1} = w_i^{k+1} b_i(x) + \lambda_i^k/\rho$$

This furthermore implies that there is “consensus” on the dual variable $\lambda_i^{k+1} = \overline{a}^{k+1} -\ z^{k+1}$. Indeed this problem is similar to the sharing problem $\min_{x_i} \sum_{i=1}^{N} f_i(x_i) + g(\ \bar{x})$, which in its ADMM formulation is the dual of the consensus problem and vice versa.

Finally, note that the ADMM approach can be also used for some special cases where the constraints set S is not convex. For this general problem

$$\min_x f(x), x \in S$$

the (scaled) ADMM steps have the form

$$x^{k+1} = argmin_x\ f(x) + \frac{\rho}{2\|x - z^k + u^k\|_2^2}$$

$$z^{k+1} = \Pi_S\left(x^{k+1} + u^k\right)$$

$$u^{k+1} = u^k + x^{k+1} - z^{k+1}$$

Where Π_S is the projection on the constraints set, for example in case of a cardinality constraints, it consist in retaining the k largest elements of the vector

Questions:

- Q1: Show that the ADMM sharing problem is dual to the regularized consensus problem
- Q2: Consider a Lasso-regression problem with a large number of variables and observations. We can imagine distributing the calibration optimization problem to multiple workers either splitting across the training examples or the features. How do the two distributed problem formulation differ? In which situations do you think one formulation is to be favored w.r.t the other?

We look forward to your opinions and insights.

Best Regards,

swissQuant Group Leadership Team