Probabilistic Alternating Direction Method of Multipliers

Abstract

1. Introduction

2. Related Work

2.1. Exact distributed algorithms

Exact computation, requires constant communications among machines.

2.2. Communication efficient distributed algorithms

Approximate computation, only requires one communication among machines.

3. Preliminaries

3.1. Problem formulation

In the following we consider a general optimization problem of the form

$$\text{minimize} \sum_{n\in\Omega} l(y_n|\theta) + h(\theta). \tag{1}$$

Here $l(y_n|\theta)$ is the loss function, $h(\theta)$ is the regularization function, and the observed data is denoted as $\mathcal{Y} = \{y_n, n \in \Omega\}$. If $\mathcal{Y}$ is partitioned into B blocks, with each data block denoted as $\boldsymbol{y}_b = \{y_n, n \in \Omega_b\}$ and Ω_b represents the index set of b^{th} data block, we can re-write (1) into the following equivalent optimization problem:

$$\begin{aligned} &\text{minimize} \sum_{b=1}^{B} \sum_{n\in\Omega_b} l(y_n, \theta_b) + h(\theta), \\ &\text{subject to } \theta_b - \theta = 0,\ b = 1, \ldots, B. \end{aligned} \tag{2}$$

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

3.2. Distributed learning via ADMM

The ADMM formulation for the problem (2) can be derived directly from the following augmented Lagrangian

$$L_\rho(\{\theta_b, \lambda_b\}_{b=1}^B, \theta) = \sum_{b=1}^{B} L_\rho(\theta_b, \lambda_b, \theta) + h(\theta). \tag{3}$$

Here $L_\rho(\theta_b, \lambda_b, \theta)$ is the local augmented Lagrangian for the b^{th} data block defined as follows

$$L_\rho(\theta_b, \lambda_b, \theta) = \sum_{n\in\Omega_b} l(y_n, \theta_b) + \lambda_b \cdot (\theta_b - \theta) + \frac{\rho}{2}||\theta_b - \theta||_2^2. \tag{4}$$

Where $\cdot$ is the dot product operator and $\rho > 0$ is a tuning parameter. Based on (3) and (4), ADMM consists of the iterations

$$\theta_b^{k+1} = \arg\min_{\theta_b} L_\rho(\theta_b, \lambda_b^k, \theta^k) \tag{5}$$

$$\theta^{k+1} = \arg\min_{\theta} \sum_{b=1}^{B} \frac{\rho}{2}||\theta_b^{k+1} - \theta + \frac{\lambda_b^k}{\rho}||_2^2 + h(\theta) \tag{6}$$

$$\lambda_b^{k+1} = \lambda_b^k + \rho(\theta_b^{k+1} - \theta) \tag{7}$$

Note that (5) and (7) can be carried out independently for each $b = 1, \ldots, B$.

3.3. Distributed learning from a Bayesian perspective

Denote the global parameter as θ, each data block is augmented with a latent (unobserved) variable θ_b such that $\theta_b \sim p(\theta_b|\theta, \gamma_b)$, *e.g.,* the local parameter θ_b is modelled as a random variable drawn from a distribution that is dependent on global parameter θ and some hyperparameter γ_b. In the Bayesian setting, we are interested in the posterior configuration $q(\theta_b)$, instead of a point estimate θ_b as in the ADMM formulation. As a result, the optimization problem can be formulated as

$$\text{minimize} \sum_{b=1}^{B} F(q(\theta_b), \theta) + h(\theta) \tag{8}$$

Here $F(q(\theta_b), \gamma_b, \theta)$ is the local objective function for the b^{th} data block

$$F(q(\theta_b), \gamma_b, \theta) = \mathbb{D}[q(\theta_b)||p(\theta_b|\theta, \gamma_b)] + \sum_{n\in\Omega_b} \mathbb{E}_q[l(y_n, \theta_b)], \tag{9}$$

where $\mathbb{D}[\cdot||\cdot]$ represents the Kullback-Leibler divergence, $\mathbb{E}_q[\cdot]$ is the expectation take with respect to distribution $q(\cdot)$. Minimizing objective function (10) with respect to $q(\theta_b)$ gives the following optimal solution for $q(\theta_b)$

$$q(\theta_b) = \frac{1}{Z(\gamma_b, \theta)} p(\theta_b|\theta, \gamma_b) e^{\sum_{n\in\Omega_b} l(y_n, \theta_b)}, \tag{10}$$

where $Z(\gamma_b, \theta) = \int_{\theta_b} p(\theta_b|\theta, \gamma_b) e^{\sum_{n\in\Omega_b} l(y_n,\theta_b)} d\theta_b$ is the normalization constant. Note that when $l(y_n, \theta_b)$ can be interpreted as a negative log-likelihood function (*e.g.,* squared ℓ_2 and ℓ_1 loss functions as unnormalized negative Gaussian and Laplace log-likelihood function respectively, logistic loss function as unnormalized negative logistic log-likelihood function respectively), minimizing (10) can be shown to be equivalent to solving the Bayes' rule based which the optimal solution (10) can be directly derived. Given $q(\theta_b)$, the global parameter θ can be updated as follows

$$\theta = \arg\min_{\theta} \sum_{b=1}^{B} \mathbb{E}_q[-\log p(\theta_b|\theta, \gamma_b)] + h(\theta).$$

Hyper-parameter γ_b can be updated using empirical Bayes method

$$\gamma_b = \arg\min_{\gamma_b} \mathbb{E}_q[-\log p(\theta_b|\theta, \gamma_b)]$$

In the following we will focus on the latent Gaussian models, where $p(\theta_b|\theta, \gamma_b) = \mathcal{N}(\theta, \gamma_b^{-1})$, *e.g.,* the latent variable θ_b is drawn from a Gaussian distribution with mean θ and precision parameter γ_b.

3.4. Weighted average via empirical Bayes

3.5. Small sample bias in big data

We propose the following distributed learning problem

$$\begin{aligned} \text{minimize} \quad & \sum_{b=1}^{B} F(q(\theta_b), \theta, \gamma_b) + h(\theta) \\ \text{subject to} \quad & \mathbb{E}_q[\theta_b] = \theta,\ b = 1, \ldots, B \end{aligned} \tag{11}$$

, $f(x_n, \theta^b)$ is the cost function, and θ^b is a random variable with prior distribution $p(\theta^b|\theta)$, which depends on global parameter θ, in the following we focus on the Gaussian case $p(\theta^b|\theta) = \mathcal{N}(\theta^b|\theta, \gamma^b)$, with mean θ and precision parameter γ^b.

Given the global parameter θ, we have the following local optimization problem for each block b

$$\begin{aligned} \min_{q(\theta^b)} \ & \mathbb{D}[q(\theta^b)||p(\theta^b|\theta)] + \sum_{n\in\Omega^b} \mathbb{E}_q[f(x_n, \theta^b)] \\ & \text{subject to } \mathbb{E}_q[\theta^b] = \theta \end{aligned} \tag{12}$$

With the Gaussian assumption on $p(\theta^b|\theta)$, the optimal solution of $q(\theta^b)$ in (12) has the following form

$$q(\theta^b) = \frac{1}{Z(\lambda^b, \theta, \gamma^b)} e^{-\sum_{n\in\Omega^b} f(x_n,\theta^b) - \lambda^b(\theta^b-\theta) - \frac{\gamma^b}{2}(\theta^b-\theta)^2} \tag{13}$$

where $Z(\lambda^b, \theta, \gamma^b) = \int p(\theta^b|\theta) e^{-\sum_{n\in\Omega^b} f(x_n,\theta^b) - \lambda^b(\theta^b-\theta) - \frac{\gamma^b}{2}(\theta^b-\theta)^2} d\theta^b$ is the normalization constant and λ^b is the Lagrange multipliers. When $Z(\lambda^b, \theta)$ is analytically intractable, *i.e.,* for logistic regression and low-rank matrix factorization, we propose to use the dual ascent algorithm instead as in the ADMM framework, where λ^b is updated using gradient descent

$$\lambda^{b\,t+1} = \lambda^{b\,t} + \gamma^b(\mathbb{E}_q[\theta^b] - \theta) \tag{14}$$

Note that the exponent in equation (13) can be re-written in the following way

$$-\sum_{n\in\Omega^b} f(x_n, \theta^b) - \frac{\gamma^b}{2}(\theta^b - \theta + \lambda^b/\gamma^b)^2 + C(\lambda^b, \gamma^b)$$

Where $C(\lambda^b, \gamma^b)$ is independent of θ^b. If we treat $f(x_n, \theta^b)$ as the negative log-likelihood function, then intuitively θ^b's prior distribution has a Gaussian form: $\mathcal{N}(\theta^b|\theta - \lambda^b/\gamma^b, \gamma^b)$. Based on this observation, we propose to update γ^b as follows:

$$\gamma^b = \frac{|\mathcal{P}^b|}{\sum_{p\in\mathcal{P}^b} \mathbb{E}_q[(\theta_p^b - \theta_p + \lambda_p^b/\gamma^b)^2]} \tag{15}$$

Finally, given $q(\theta^b)$, λ^b and γ^b from each local block b, the global optimization problem is

$$\min_{\theta} \sum_{b=1}^{B} \left(\mathbb{D}[q(\theta^b)||p(\theta^b|\theta)] + \lambda^b(\mathbb{E}_q[\theta^b] - \theta)\right) + h(\theta) \tag{16}$$

Assume $h(\theta)$ is the following elastic net regularizer, which absorbs ℓ_1 or ℓ_2 norms regularizer as special cases

$$h(\theta) = \lambda(\alpha||\theta||_1 + (1-\alpha)||\theta||_2^2) \tag{17}$$

After simple algebra, with the Gaussian assumption on $p(\theta^b|\theta)$ the optimal solution for θ is

$$\theta = \frac{S\left(\sum_{b=1}^{B} \gamma^b \mathbb{E}_q[\theta^b], \lambda\alpha\right)}{\sum_{b=1}^{B} \gamma^b + \lambda(1-\alpha)} \tag{18}$$

where $S(\cdot)$ is the soft-thresholding operator.

4. Extenstions

4.1. Full Bayesian treatment

4.2. Network structure induced hierarchical modeling

5. Random thoughts

On evaluating $\mathbb{E}_q[\theta_b]$ We have not explored importance resampling methods because of the potential for importance resampling to collapse to a single point in high dimensions.

On updating γ_b One is tempted to argue that the posterior distributions in each shard will have roughly the same shape, and so the draws from different shards can be combined with equal weighting. However, large samples on each shard are necessary for the preceding argument to apply. In particular, each worker must have enough information to estimate all components of θ. Almost by definition, models that are complex enough to require very large data sets will have subsets of parameters that cannot be estimated on every shard.

On adding the equality constraint Oddly enough, small sample bias can play a role when analyzing large data sets (Zhang et al., 2012). When the data are divided among many machines, bias that vanishes in the full data may still be present in each shard. Bootstrap/Subsampling based methods could be used to address this problem, however, they are sensitive to the subsampling rate (as all subsampling based methods do). In this work we propose to model such sample bias explicitly, which improves the overall performance with negligible computational overhead.

Remarks Our approach is a combine/trade-off of pure optimization based approach, and pure Bayesian based approach, and is a combine/trade-off between exact distributed algorithm and communication-efficient algorithm.

6. Application Examples

6.1. Logistic regression

6.2. Matrix completion

References

Zhang, Y., Duchi, J., and Wainwright, M. Communication-efficient algorithms for statistical optimization. In *NIPS*, 2012.