Distributed ML

Krishna Pillutla

ntroductio

IPM

Literature Review Algorithm

ADMM Literature

Review
A general framework f distributed

Experiments and Results

Distributed Machine Learning: Iterative Convex Optimization Methods

Krishna Pillutla

IIT Bombay

krishna.p@iitb.ac.in www.cse.iitb.ac.in/~krishna.p

May 5, 2014

Advisor

Distributed ML

Krishna Pillutla

ntroduction

IPM

Literatui Review

. _ . . .

Litaratus

A general framework for distributed optimization Approach

Experiments

Prof. SakethaNath J saketh@cse.iitb.ac.in

Overview

Distributed ML

Krishna Pillutla

Introductio

IPM

Literatur Review

A D. AA

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

1 Introduction

- 2 IPM
 - Literature Review
 - Algorithm
- 3 ADMM
 - Literature Review
 - A general framework for distributed optimization
 - Approach
- 4 Experiments and Results

Distributed ML

Krishna Pillutla

Introduction

IDA4

Review

VDMV

Literatui Review

A general framework for distributed optimization Approach

Experiments and Results

1 Introduction

- 2 IPM
 - Literature Review
 - Algorithm

3 ADMM

- Literature Review
- A general framework for distributed optimization
- Approach
- 4 Experiments and Results

Distributed ML

Krishna Pillutla

Introduction

......

IPIVI

Review

ADMN

Literatur

A general framework distributed optimizatio

Experiments

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

Distributed ML

Krishna Pillutla

Introduction

IPM

Review

ADMN

Literatui Review

A general framework for distributed optimization Approach

Experiments and Results

To solve learning problems of the form

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

■ *I* is the loss function (convex and ∇I is Lipschitz continuous)

Distributed ML

Krishna Pillutla

Introduction

IPM

Literature Review

A D. A.A.

Literatur Review

A general framework for distributed optimization Approach

Experiments

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)

Distributed ML

Krishna Pillutla

Introduction

IPM

Literatur Review Algorithm

ΔDΜМ

ADIVIN Literati

Review
A general
framework for
distributed
optimization
Approach

Experiments and Results

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)
- w is the model

Distributed ML

Krishna Pillutla

Introduction

IPM

Review

ADMN

A general framework fo distributed optimization Approach

Experiments and Results

$$min_w \sum_{i=1}^{N} I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)
- w is the model
- Regularizer $R(w) = w^T w/2$

Distributed ML

Krishna Pillutla

Introduction

IPM

Literature Review

ADMN

Review
A general

A general framework for distributed optimization Approach

Experiments and Results

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)
- w is the model
- Regularizer $R(w) = w^T w/2$
- We wish to solve it in a distributed setting

Distributed ML

Krishna Pillutla

Introduction

IРМ

Literatur Review

ΔDΜΝ

Literature Review A general

A general framework for distributed optimization Approach

Experiments and Results

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)
- w is the model
- Regularizer $R(w) = w^T w/2$
- We wish to solve it in a distributed setting
 - data is distributed across the nodes.

Distributed ML

Krishna Pillutla

Introduction

IPM

Literatur Review

A DMA

ADMI

Review
A general
framework fo
distributed
optimization
Approach

Experiments and Results

$$min_w \sum_{i=1}^N I(y_i w^T x_i) + \gamma R(w)$$

- I is the loss function (convex and ∇I is Lipschitz continuous)
- x_i a training example and y_i is its label (+1 or -1 for binary classification)
- w is the model
- Regularizer $R(w) = w^T w/2$
- We wish to solve it in a distributed setting
 - data is distributed across the nodes.
 - locally optimise and communicate the models to get one common global model.

Distributed ML

Krishna Pillutla

Introductio

IPM Literature Review

A D MAN

Literatui Review

A general framework for distributed optimization Approach

- 1 Introduction
- 2 IPM
 - Literature Review
 - Algorithm
- 3 ADMM
 - Literature Review
 - A general framework for distributed optimization
 - Approach
- 4 Experiments and Results

Distributed

Literature Review

■ Parameter Mixing(PM): Independently solve optimisation on each node and take a convex combination of these to represent the global model [Man], [MMM⁺09].

Distributed ML

Pillutla

Introduction

IРМ

Literature Review

ADMN

Literatu

A general framework for distributed optimization Approach

- Parameter Mixing(PM): Independently solve optimisation on each node and take a convex combination of these to represent the global model [Man], [MMM+09].
- Use local gradient information in coefficients of the convex combination (to give weights to different components)
 [ACDL11]

Distributed ML

Krishna Pillutla

Introduction

IPM

Literature Review

ADMN

Literature Review A general

A general framework for distributed optimization Approach

- Parameter Mixing(PM): Independently solve optimisation on each node and take a convex combination of these to represent the global model [Man], [MMM+09].
- Use local gradient information in coefficients of the convex combination (to give weights to different components)
 [ACDL11]
- Use gradient and hessian information from quadratic Taylor approximations from other nodes (did in my Summer Internship at MSR).

Distributed ML

Krishn Pillutla

Introduction

IΡΝ

Literature Review

ADMN

Literature Review A general framework f distributed optimization

- Parameter Mixing(PM): Independently solve optimisation on each node and take a convex combination of these to represent the global model [Man], [MMM+09].
- Use local gradient information in coefficients of the convex combination (to give weights to different components)
 [ACDL11]
- Use gradient and hessian information from quadratic Taylor approximations from other nodes (did in my Summer Internship at MSR).
- What if the Taylor approximation does not hold?

Distributed ML

Krishna Pillutla

ntraduction

IPIVI

Literature Review

Aigonti

ADMM

Literatur Review

framework fo distributed optimization Approach

Experiments and Results

■ IPM: Run multiple iterations of PM- by using the model obtained at each PM step as the starting guess for the next step- until global convergence. Theoretical bounds exist for the perceptron [MHM]

Distributed ML

Pillutla

ntroduction

IPM

Literature Review

ADMN

Literature Review A general

A general framework for distributed optimization Approach

- IPM: Run multiple iterations of PM- by using the model obtained at each PM step as the starting guess for the next step- until global convergence. Theoretical bounds exist for the perceptron [MHM]
- [MKSB13] uses functional approximation for IPM (published in January 2014, after stage 1). It has theoretical guarantees but requires a global line search step.

Distributed ML

Pillutla

Introduction

IΡΝ

Literature Review

ADMN

Literature Review

A general framework for distributed optimization Approach

- IPM: Run multiple iterations of PM- by using the model obtained at each PM step as the starting guess for the next step- until global convergence. Theoretical bounds exist for the perceptron [MHM]
- [MKSB13] uses functional approximation for IPM (published in January 2014, after stage 1). It has theoretical guarantees but requires a global line search step.
- Node *i* minimises $f_i(w) + C_i(w)$ where $C_i(w)$ is a quadratic satisfying some mild requirements.

Distributed ML

Krishn: Pillutla

Introductio

IPM

Literature Review

ADMN

Literature Review A general framework for distributed optimization Approach

- IPM: Run multiple iterations of PM- by using the model obtained at each PM step as the starting guess for the next step- until global convergence. Theoretical bounds exist for the perceptron [MHM]
- [MKSB13] uses functional approximation for IPM (published in January 2014, after stage 1). It has theoretical guarantees but requires a global line search step.
- Node *i* minimises $f_i(w) + C_i(w)$ where $C_i(w)$ is a quadratic satisfying some mild requirements.
- Can we achieve theoretical guarantees and practical results without this line search step?

Distributed ML

Krishn Pillutl

lankan alau akta.

......

IPM

Literatui

Algorithm

<u> ДОМИ</u>

ADMI

Review

A general framework fo distributed optimization Approach

Experiments and Results

 Each nodes approximates the objective function at other nodes by a linear or quadratic approximation about the globally accepted starting point.

Distributed ML

Krishn Pillutla

Introduction

IPM

Review

Algorithm

ADMN

Review
A general framework

Experiments

- Each nodes approximates the objective function at other nodes by a linear or quadratic approximation about the globally accepted starting point.
- This way, each node has a global picture.

Distributed ML

Krishna Pillutla

straductio

.....

IPM

Poviou

Algorithm

Aigoritiii

Literature

Review

framework f distributed optimization

Experiments and Results

■ Each node has an objective function $f_i(w)$.

Distributed ML

Krishna Pillutla

ntroduction

..... o d d c c.i.c

IPM

Literatui

Algorithm

Algorith

Literature

A general framework for distributed optimization

- Each node has an objective function $f_i(w)$.
- If *L* is the Lipschitz constant of ∇f , define

$$\hat{f}^{w_0}(w) := f(w_0) + \nabla f(w_0)^T (w - w_0) + \frac{L}{2} \|w - w_0\|^2$$

Distributed ML

Krishna Pillutla

Introductio

IPM

Review

Algorithm

ADMI

Review
A general

A general framework for distributed optimization Approach

Experiments

- **Each** node has an objective function $f_i(w)$.
- If *L* is the Lipschitz constant of ∇f , define

$$\hat{f}^{w_0}(w) := f(w_0) + \nabla f(w_0)^T (w - w_0) + \frac{L}{2} \|w - w_0\|^2$$

■ If $w^{(k)}$ is the global model from the previous iteration, node i solves,

$$\min_{w} f_i(w) + \sum_{i \neq i} \hat{f}_j^{w^{(k)}}(w)$$

Algorithm

Distributed

Algorithm

Algorithm 1 Our Algorithm for IPM

- 1. Initialise $w^{(0)}$
- 2: **for** t = 1, 2, ... (outer iterations) **do**

3:
$$\tilde{f}_i^{(t)}(w) = f_i(w) + \hat{f}_{-i}^{w^{(k)}}(w)$$

- $w_i^{(t)} = \operatorname{argmin}(\tilde{f}_{i,t}(w))$ by some method
- $w^{(t+1)} = \text{ParameterMixing}(w_i^{(t)})$ 5:
- Obtain $f(w^{(t+1)})$ and $\nabla f(w^{(t+1)})$ by communication 6:
 - end for
- 8: return $w^{(t+1)}$

ParameterMixing
$$(w_i^{(t)}) = \sum_{i=1}^{m} \alpha_i w_i^{(t)}$$

Algorithm from [MKSB13]

Distributed

Algorithm

Algorithm 2 IPM Algorithm proposed in [MKSB13] applied to our setting

- 1: Initialise $w^{(0)}$
- 2: **for** t = 1, 2, ... (outer iterations) **do**

3:
$$\tilde{f}_{i}^{(t)}(w) = f_{i}(w) + \hat{f}_{-i}^{w^{(k)}}(w)$$

- $w_i^{(t)} = \operatorname{argmin}_w(\tilde{f}_{i,t}(w))$ by some method 4:
- 5: $d^{(t)} = \text{ParameterMixing}(w_i^{(t)}) w^{(t)}$
- $w^{(t+1)} = w^{(t)} + \tau d^{(t)}$ where τ is a step length satisfying Armijo-Wolfe conditions.
- Obtain $f(w^{(t+1)})$ and $\nabla f(w^{(t+1)})$ by communication
- 8: end for
- 9: return $w^{(t+1)}$

Algorithm from [MKSB13]

Distributed

Algorithm

■ Linear Convergence is guaranteed for Algorithm 2 [MKSB13].

Distributed ML

Krishna Pillutla

Introductio

IPM

Literatur

Algorithm

A DIMI

ADIVII

Literati

A general framework for distributed optimization

Experiment

Theorem (Convergence)

If f is convex and differentiable, ∇f is Lipschitz continuous with constant L, a strict decrease in global objective function f can be guaranteed in every outer iteration (parameter mixing step) under suitable conditions. In particular, for gradient descent, the condition is that the step size h satisfies

$$0 < h \le 1/L$$

Because f is convex, IPM converges to the unique global minimizer w*.

Distributed ML

Pillutla

atroduction

.....

IDM 4

Review

Algorithm

Algorithi

ADMM

Literatur Review

A general framework for distributed optimization Approach

Experiments and Results

Local Rate of Convergence:

lacksquare f if convex and differentiable, ∇f is Lipschitz continuous with constant L

Distributed ML

Pillutla

ntroduction

IIItioductio

IPM

Review

Algorithm

ADMM.

Review

A general framework for distributed optimization Approach

Experiments and Results

- f if convex and differentiable, ∇f is Lipschitz continuous with constant L
- Gradient Descent (fixed number of iterations) is used for inner optimization

Distributed ML

Krishna Pillutla

ntroductio

Literati

Algorithm

-A D.MA.

ADMN

Review
A general

A general framework for distributed optimization Approach

Experiments and Results

- f if convex and differentiable, ∇f is Lipschitz continuous with constant L
- Gradient Descent (fixed number of iterations) is used for inner optimization
- $w^{(0)}$ is sufficiently close to the global optimum w^*

Distributed ML

Krishna Pillutla

Introductio

IPN

Review Algorithm

A DAMA

ADMN

Review
A general framework f

Approach

Experiments

Experiments and Results

- f if convex and differentiable, ∇f is Lipschitz continuous with constant L
- Gradient Descent (fixed number of iterations) is used for inner optimization
- $w^{(0)}$ is sufficiently close to the global optimum w^*
- Convergence of IPM is $\mathcal{O}(1/k)$, for k outer iterations

Distributed

Algorithm

- f if convex and differentiable, ∇f is Lipschitz continuous with constant L
- Gradient Descent (fixed number of iterations) is used for inner optimization
- $w^{(0)}$ is sufficiently close to the global optimum w^*
- Convergence of IPM is $\mathcal{O}(1/k)$, for k outer iterations
- If f is strongly convex, convergence is linear.

Distributed ML

Krishna Pillutla

Introductio

miroductic

IPIVI

Review

Algorithm

ADMI

ADIVII

A general framework distributed

Experiments and Results

Theorem (Local Rate of Convergence)

If f if convex and differentiable, ∇f is Lipschitz continuous with constant L, the inner optimisation is solved with gradient descent i.e. $w_i^{(k,j+1)} = w_i^{(k,j)} - h \nabla \tilde{f}_{i,k}(w_i^{(k,j)})$ with a fixed number of steps c and fixed step size $h = 1/L(c^2 + 2c - 2)$ and the initial guess $w^{(0)}$ is sufficiently close to the global optimum w^* then IPM converges as

$$f(w^{(k)}) - f(w^*) \le \frac{2L\|w^{(0)} - w^*\|^2}{k+4}\beta^2$$

where $\beta > 0$ is a constant i.e., convergence is $\mathcal{O}(1/k)$, for k outer iterations. If f is strongly convex with constant μ , convergence is linear as $\|w^{(k)} - w^*\| \le (\frac{L-\mu}{L+\mu})^k \|w^{(0)} - w^*\|$

Distributed ML

Krishna Pillutla

ntroduction

Literatur

Algorithm

. .

A [] A A A A

Review

A general framework fo distributed optimization Approach

Experiments

Global Rate of Convergence

Distributed ML

Krishna Pillutla

Introductio

meroducero

Literatu

Review

Algorithm

ADMI

ADMI

Review
A general framework for distributed

Experiments

We tried to but were unable to prove this theorem:

Theorem (Global Rate of Convergence)

If f if convex and differentiable, ∇f is Lipschitz continuous with constant L, the inner optimisation is solved with gradient descent i.e. $w_i^{(k,j+1)} = w_i^{(k,j)} - h \nabla \tilde{f}_{i,k}(w_i^{(k,j)})$ with a fixed number of steps c and fixed step size $h = 1/L(c^2 + 2c - 2)$, then IPM converges as $f(w^{(k)}) - f(w^*) \leq \frac{2L\|w^{(0)} - w^*\|^2}{k+4}\beta^2$ where $\beta > 0$ is a constant i.e., convergence is $\mathcal{O}(1/k)$, for k outer iterations. If f is strongly convex with constant μ , convergence is linear as $\|w^{(k)} - w^*\| \leq (\frac{L-\mu}{L+\mu})^k \|w^{(0)} - w^*\|$

Distributed ML

Krishn Pillutl

ntroduction

....

IF IV

Review

Algorithm

4 D 1 4 1 4

Literatur

A general framework for distributed optimization Approach

Experiments and Results

We had to settle for a weaker bound:

• f if convex and differentiable, ∇f is Lipschitz continuous with constant I

Distributed ML

Pillutl

ntroductio

Literatui Review

Algorithm

ADMN

Review A general

A general framework for distributed optimization Approach

Experiments

We had to settle for a weaker bound:

- f if convex and differentiable, ∇f is Lipschitz continuous with constant L
- Gradient Descent (fixed number of iterations or with line search) is used for inner optimization

Distributed

Algorithm

We had to settle for a weaker bound:

- f if convex and differentiable, ∇f is Lipschitz continuous with constant I
- Gradient Descent (fixed number of iterations or with line) search) is used for inner optimization
- Convergence of IPM is $\mathcal{O}(1/\sqrt{k})$, for k outer iterations

Distributed ML

Krishna Pillutla

troduction

miroductic

IPM

Literatu

Algorithm

A D A 41

ADMI

Literature Review A general framework for distributed optimization Approach

Experiments and Results

Theorem (Global Rate of Convergence)

If f if convex and differentiable, ∇f is Lipschitz continuous with constant L, the inner optimization is solved with gradient descent i.e. $w_i^{(k,j+1)} = w_i^{(k,j)} - h\nabla \tilde{f}_{i,k}(w_i^{(k,j)})$ with a fixed number of steps c, and a fixed step size of h = 1/L, we have,

$$\|\nabla f(w^{(k)}\| \le \sqrt{\frac{2L(f(w^{(0)} - f(w^*))}{k+1}}$$

In other words, convergence is $O(1/\sqrt{k})$, for k outer iterations.

Distributed ML

Krishna Pillutla

Introduction

Literature Review Algorithm

ADMM

Review

A general framework for distributed optimization Approach

Experiments and Results

- 1 Introduction
- 2 IPM
 - Literature Review
 - Algorithm
- 3 ADMM
 - Literature Review
 - A general framework for distributed optimization
 - Approach
- 4 Experiments and Results

Distributed ML Krishna

Pillutl

ntroduct

IPM

Literature Review

ADMN

Literature Review A general framework fo

distributed optimization Approach

Experiments and Results

■ A separable problem of the form $min_w f(w) + g(w)$ is recast as $min_{x,y} f(x) + g(y)$ subject to x = y.

Distributed ML

Krishn Pillutla

Introductio

IPN

Literatu Review Algorith

ADMM Literature

Review
A general framework for distributed

Experiment

■ A separable problem of the form $min_w f(w) + g(w)$ is recast as $min_{x,y} f(x) + g(y)$ subject to x = y.

■ Based on theory of Lagrange Duality.

Distributed ML

Rrishn Pillutla

Introductio

IPI

Literatur Review Algorithi

ADMM

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

- A separable problem of the form $min_w f(w) + g(w)$ is recast as $min_{x,y} f(x) + g(y)$ subject to x = y.
- Based on theory of Lagrange Duality.
- Because of the separable nature, x is updated keeping y fixed and then, y is updated keeping x fixed 1.

46 / 104

Distributed ML

Pillutl

Introductio

IPI

Literatu Review Algorith

ADMN

Literature Review

A general framework for distributed optimization Approach

Experiments

- A separable problem of the form $min_w f(w) + g(w)$ is recast as $min_{x,y} f(x) + g(y)$ subject to x = y.
- Based on theory of Lagrange Duality.
- Because of the separable nature, x is updated keeping y fixed and then, y is updated keeping x fixed ¹.
- Dual variables are updated and the process is repeated.

Distributed ML

Krishn Pillutla

Introductio

IPI

Literatu Review Algorith

ADMN

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

- A separable problem of the form $min_w f(w) + g(w)$ is recast as $min_{x,y} f(x) + g(y)$ subject to x = y.
- Based on theory of Lagrange Duality.
- Because of the separable nature, x is updated keeping y fixed and then, y is updated keeping x fixed ¹.
- Dual variables are updated and the process is repeated.
- Can trivially be parallelized ([BPC+11]).

Distributed ML

Krishna Pillutla

Introductio

IPN

Literatur Review Algorithr

ADMN

Literature

Review

framework for distributed optimization Approach

Experiments and Results

The same idea can be generalized to $min_w \sum_{i=1}^m f_i(w)$ ([BT97]). We rewrite the problem as

$$\min_{w_i,\dots,w_m} \qquad \sum_{i=1}^m f_i(w_i)$$
 subject to $w_i=w_{i+1}; i=1,\dots,m-1.$

where $f_i(.)$ is the objective function at node i.

Distributed ML

Krishn Pillutla

ntroductio

minoduction

Literatu Review

ADIVIIV

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

■ λ_j represents the Lagrangian dual variable corresponding to the j^{th} constraint

Distributed ML

Krishna Pillutla

Introduction

Literatur Review Algorithr

ADMN

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

- λ_j represents the Lagrangian dual variable corresponding to the j^{th} constraint
- c is the augmented lagrangian parameter

Distributed

Literature Review

- λ_i represents the Lagrangian dual variable corresponding to the j^{th} constraint
- c is the augmented lagrangian parameter
- Apply equation (4.75) of [BT97] to get:

Distributed ML

Krishna Pillutla

Introduction

....

Literature Review

ADMM

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

- λ_j represents the Lagrangian dual variable corresponding to the i^{th} constraint
- c is the augmented lagrangian parameter
- Apply equation (4.75) of [BT97] to get:

$$w_{i}^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_{i}(w) + c \|w\|^{2} + w^{T} (\lambda_{i}^{(t)} - \lambda_{i-1}^{(t)} - c(w_{i}^{(t)} + \frac{w_{i-1}^{(t)} + w_{i+1}^{(t)}}{2})) \}$$

$$(1)$$

Distributed ML

Krishna Pillutla

Introduction

....

Literatur Review

л Били

Literature Review

A general framework for distributed optimization Approach

Experiments and Results

 λ_j represents the Lagrangian dual variable corresponding to the i^{th} constraint

- c is the augmented lagrangian parameter
- Apply equation (4.75) of [BT97] to get:

$$w_{i}^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_{i}(w) + c \|w\|^{2} + w^{T} (\lambda_{i}^{(t)} - \lambda_{i-1}^{(t)} - c(w_{i}^{(t)} + \frac{w_{i-1}^{(t)} + w_{i+1}^{(t)}}{2})) \}$$

$$(1)$$

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \frac{c}{2} (w_i^{(t+1)} - w_{i+1}^{(t+1)})$$
 (2)

ADMM: continued: Implementation

Distributed ML

Krishna Pillutla

ntroductio

ΙΡΜ

Literatur Review

DMM

Literature Review

A general framework f distributed

Experiments

lacktriangleq w updates happen parallely

ADMM: continued: Implementation

Distributed ML

Krishna

Introductio

IPM

Literature Review

45141

Literature

Review

A general framework for distributed optimization Approach

Experiments and Results

- w updates happen parallely
- Node i communicates model with neighbours i-1 and i+1.

Distributed

A general framework for distributed optimization

Introduction

- 2 IPM
 - Literature Review
 - Algorithm
- 3 ADMM
 - Literature Review
 - A general framework for distributed optimization
 - Approach
- **Experiments and Results**

Distributed ML

Krishn Pillutl

atroductic

minoducti

Literatur Review

.

Literatur

A general framework for distributed optimization

Experiments

■ Add a linear or a quadratic correction C(w) to the objective function at each node.

Distributed ML

Krishna Pillutla

Introductio

IPM

Review

A D.M.M.

TDIVIIV

Review

A general framework for distributed optimization Approach

Experiments

- Add a linear or a quadratic correction C(w) to the objective function at each node.
- IPM adds $\hat{f}_{-i}^{w^{(k)}}(w)$

Distributed ML

Krishna

Introductio

IPN

Literatur Review Algorithm

4 D.M.

ADIVIIV

Literatui Review

A general framework for distributed optimization Approach

Experiments

- Add a linear or a quadratic correction C(w) to the objective function at each node.
- IPM adds $\hat{f}_{-i}^{w^{(k)}}(w)$
- The correction is iteratively improved along with the solution ([MKSB13], [BPC+11], [HMS08]).

Distributed ML

Krishna Pillutla

ntroductic

Literature Review

Review Algorithm

Literature

A general framework for distributed optimization Approach

Experiments

■ IPM adds $\hat{f}_{-i}^{w^{(k)}}(w)$

Distributed ML

Krishna Pillutla

ntroductio

IPM

Literature Review

MMQA

Literature Review

A general framework for distributed optimization Approach

Experiments

■ IPM adds $\hat{f}_{-i}^{w^{(k)}}(w)$

■ ADMM, based on Lagrangian Duality adds

$$C_i^{(t)}(w) = c||w||^2 + w^T d_i^{(t)}$$

where
$$d_i^{(t)} = (\lambda_i^{(t)} - \lambda_{i-1}^{(t)} - c(w_i^{(t)} + \frac{w_{i-1}^{(t)} + w_{i+1}^{(t)}}{2}))$$

Distributed ML

Krishna Pillutla

Introductio

IΡΝ

Review
Algorithm

ADMN

Literatur Review

A general framework for distributed optimization Approach

Experiments and Results

■ IPM adds $\hat{f}_{-i}^{w^{(k)}}(w)$

ADMM, based on Lagrangian Duality adds

$$C_i^{(t)}(w) = c||w||^2 + w^T d_i^{(t)}$$

where
$$d_i^{(t)} = (\lambda_i^{(t)} - \lambda_{i-1}^{(t)} - c(w_i^{(t)} + \frac{w_{i-1}^{(t)} + w_{i+1}^{(t)}}{2}))$$

■ When Fenchel duality is used, we have a linear term ([HMS08], equation (2)) used to tie together solutions from various nodes.

Distributed ML

Krishn Pillutl

minoductio

IPN

Review

. _ . . .

ADIVIIV

Review A general

A general framework for distributed optimization

Experiments

■ Observation: Each of the above three methods involves a communication of $\mathcal{O}(n)$ where n is the number of features.

Distributed ML

Krishn Pillutl

ntroductio

.....

IPIV

Review Algorithm

ADMI

Literatur Review

A general framework for distributed optimization

Experiments

- Observation: Each of the above three methods involves a communication of $\mathcal{O}(n)$ where n is the number of features.
- Can we reduce this further?

Distributed ML

Krishna Pillutla

Introduction

Literati Review

Algorith

ADMM

Literature Review

A general framework for distributed optimization Approach

Experiment and Results

Algorithm to capture all variants described above

Algorithm 3 General Algorithm

- 1: Initialise $w^{(0)}$ and other quantities required arbitrarily
- 2: for t = 1, 2, .. (outer iterations) do
- 3: Compute $C_i^{(t)}(w)$, the correction
- 4: $w_i^{(t)} = \underset{w}{\operatorname{argmin}} (f_i(w) + C_i^{(t)}(w))$ by some method
- 5: Communication: communicate the required vectors (such as dual vectors, or gradients)
- 6: end for
- 7: **return** $w^{(t)} = \text{ParameterMixing}(w_i^t)$

Distributed ML

Krishna Pillutla

ntroductio

......

Literature Review Algorithm

Literature Review A general

Approach

Experiments

■ To reduce communication costs further.

Distributed ML

Krishna Pillutla

ntroduction

. . . .

Review

4 D L 4 L

ADMIN

Review

optimizati Approach

Experiments

■ To reduce communication costs further.

Original Problem:

$$\min_{w_i,\dots,w_m}$$
 $\sum_{i=1}^m f_i(w_i)$ subject to $w_i=w_{i+1}; i=1,\dots,m-1.$

Distributed ML

Krishna Pillutla

ntroductio

IPM

Literature Review Algorithm

ADMN

Literature Review A general framework

Approach

Experiments and Results ■ To reduce communication costs further.

Original Problem:

$$\min_{w_i,\dots,w_m} \qquad \sum_{i=1}^m f_i(w_i)$$
 subject to $w_i = w_{i+1}; i = 1,\dots,m-1.$

■ Relax Constraints $w_i = w_{i+1}$

Distributed ML

Krishna Pillutla

ntroductio

....

IPIV

Review Algorithm

ADML

Literature Review

Review

A general framework f distributed

Approach

Experiments

■ Constraint $w_i = w_{i+1}$ enforces component-wise equality $w_i[j] = w_{i+1}[j]$.

Distributed ML

Krishn Pillutla

atroduction

......

IPM

Review

Algorith

ADMN

Review A general

Approach

Experiments and Results

- Constraint $w_i = w_{i+1}$ enforces component-wise equality $w_i[j] = w_{i+1}[j]$.
- Relax to sum of elements over a set of indices being equal.

Distributed ML

Krishn Pillutl

ntroductio

IPM

Literatur Review

Algoriti

ADMIN

Literature Review A general framework for distributed optimization

Approach

Experiment and Results

- Constraint $w_i = w_{i+1}$ enforces component-wise equality $w_i[j] = w_{i+1}[j]$.
- Relax to sum of elements over a set of indices being equal.
- Divide features into sets $I_1, I_2, ..., I_k$ (not necessarily disjoint).

Distributed ML

Krishna Pillutla

ntroductio

IΡΝ

Literature Review

ADMN

Literature Review A general framework for distributed optimization

Approach

- Constraint $w_i = w_{i+1}$ enforces component-wise equality $w_i[j] = w_{i+1}[j]$.
- Relax to sum of elements over a set of indices being equal.
- Divide features into sets $I_1, I_2, ..., I_k$ (not necessarily disjoint).
- Enforce: $\sum_{j \in I_r} w_i[j] = \sum_{j \in I_r} w_{i+1}[j]$ for each set I_r .

Distributed

Approach

Represent all k scalar equations as a vector equation $S^T w_i = S^T w_{i+1}^2$.

 $^{^2}S \in \mathbb{R}^{n \times k}$ is a matrix of 0s and 1s such that $S_{ij} = 1 \Leftrightarrow i \in I_i$

Distributed ML

Krishn Pillutl

Introduction

miroductic

IPN

Literatur Review Algorithr

-A D.M.

Literati

Review A general framework

Approach

Experiments and Results

Represent all k scalar equations as a vector equation $S^T w_i = S^T w_{i+1}^2$.

■ New problem:

$$\min_{w_i,\dots,w_m} \qquad \sum_{i=1}^m f_i(w_i)$$
subject to $S^T w_i = S^T w_{i+1}; i = 1,\dots,m-1$.

 $^{^2}S \in \mathbb{R}^{n \times k}$ is a matrix of 0s and 1s such that $S_{ij} = 1 \Leftrightarrow i \in I_{\bar{l}}$

Distributed ML

Krishn Pillutla

Introductio

IPM

Literatur Review Algorithn

ΔΠΜΝ

Review
A general
framework fo

Approach

Experiments and Results Represent all k scalar equations as a vector equation $S^T w_i = S^T w_{i+1}^2$.

New problem:

$$\min_{\substack{w_i,\dots,w_m\\ \text{subject to}}} \sum_{i=1}^m f_i(w_i)$$
subject to $S^T w_i = S^T w_{i+1}; i = 1,\dots,m-1.$

Approximation of original problem: need not be solved by ADMM.

 $^{^2}S \in \mathbb{R}^{n \times k}$ is a matrix of 0s and 1s such that $S_{ij} = 1 \Leftrightarrow i \in I_j$

Distributed

Approach

■ Apply equation (4.75) of [BT97] to get:

$$w_i^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_i(w) + cw^T (SS^T) w + w^T d_i^{(t)} \}$$
 (3)

Distributed ML

Krishn Pillutl

Introductio

.....

IF N

Review

45141

Literat

A general framework distributed

Approach

Experiments and Results ■ Apply equation (4.75) of [BT97] to get:

$$w_i^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_i(w) + cw^T (SS^T) w + w^T d_i^{(t)} \}$$
 (3)

where

$$d_i^{(t)} = S(\lambda_i^{(t)} - \lambda_{i-1}^{(t)} - c(S^T w_i^{(t)} + \frac{S^T w_{i-1}^{(t)} + S^T w_{i+1}^{(t)}}{2}))$$

Distributed ML

Krishr Pillut

ntroductio

.

IPM

Review

A D. A A A

Literatu Review

A general framework distributed

Approach

Experiment and Results

■ Apply equation (4.75) of [BT97] to get:

$$w_i^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_i(w) + cw^T (SS^T) w + w^T d_i^{(t)} \}$$
 (3)

where

$$d_i^{(t)} = S(\lambda_i^{(t)} - \lambda_{i-1}^{(t)} - c(S^T w_i^{(t)} + \frac{S^T w_{i-1}^{(t)} + S^T w_{i+1}^{(t)}}{2}))$$

Dual Update:

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \frac{c}{2} (S^T w_i^{(t+1)} - S^T w_{i+1}^{(t+1)})$$
 (4)

Distributed

Approach

Apply equation (4.75) of [BT97] to get:

$$w_i^{(t+1)} = \underset{w}{\operatorname{argmin}} \{ f_i(w) + cw^T (SS^T) w + w^T d_i^{(t)} \} \quad (3)$$

where

$$d_i^{(t)} = S(\lambda_i^{(t)} - \lambda_{i-1}^{(t)} - c(S^T w_i^{(t)} + \frac{S^T w_{i-1}^{(t)} + S^T w_{i+1}^{(t)}}{2}))$$

Dual Update:

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \frac{c}{2} (S^T w_i^{(t+1)} - S^T w_{i+1}^{(t+1)})$$
 (4)

Communication Cost

Algorithm

Distributed ML

Krishna Pillutla

ntroduction

meroducen

Literature Review

Review Algorithr

ADMI

Review
A general framework for distributed

Approach

Experiments and Results

Algorithm 4 ADMM with reduced communication

- 1: Initialise $w^{(0)}, w_i^{(0)}, \lambda_i = \mathbf{0}_k = \lambda_{i-1}$
- 2: **for** t = 0, 1, 2, ... **do**
- 3: Solve for $w_i^{(t+1)}$ as per equation 3
- 4: Communicate models with neighbours and obtain $w_{i-1}^{(t+1)}$ and $w_{i+1}^{(t+1)}$
- 5: Update λ_i and λ_{i-1} by equation 4
- 6: end for
- 7: **return** $w_i^{(t+1)}$ or ParameterMixing (w_i^{t+1})

Distributed ML

Krishna Pillutla

ntroductio

....

Literatur Review

Algorit

ADMI

Review
A general framework

Approach

Experiments

■ Key Observation: Parameter Mixing step in the end can be skipped

Distributed ML

Pillutl

ntroductio

IPIV

Review Algorithn

ADMN

Literature Review A general framework for distributed

Approach

- Key Observation: Parameter Mixing step in the end can be skipped
- Can reduce communication cost to be smaller than Simple Parameter Mixing

Distributed ML

Krishna Pillutla

atroduction

miroductic

IPIV

Literature Review

ADMN

Literatur Review

A general framework for distributed optimization Approach

Experiments and Results

1 Introduction

2 IPM

- Literature Review
- Algorithm

3 ADMM

- Literature Review
- A general framework for distributed optimization
- Approach
- 4 Experiments and Results

Results: Synthetic Datasets

Distributed ML

Krishna Pillutla

Introductio

IPN

Literature Review Algorithn

ADMM

Literatur

A general framework for distributed optimization Approach

- Synthetic datasets with 4-10 features
- 200-4000 training and testing examples
- $S \in \mathbb{R}^{n \times 3}$
- Solved with cvx as black-box solver
- Stopping based on duality gap

Results: Synthetic Datasets

Distributed ML

Krishna Pillutla

Introductio

IPN

Literatur Review

ADMM

ADMM

A general framework for distributed optimization

Experiments and Results

Table: Performance with Synthetic Dataset TestFinal4.mat with 4 nodes

Method	Objective value	Test Accuracy
Full problem	0.4891	0.8000
PM	0.6009	0.7350
Our method	0.5128	0.7900

Results: low k values work

Distributed ML

Krishna Pillutla

ntroduction

.

IPIV

Review

а Білі

Literatu

Review

framework fo distributed optimization

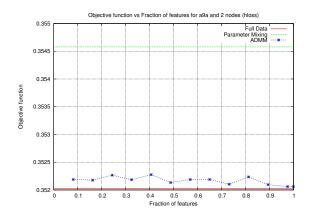


Figure: Dataset a9a split into two nodes: objective function value

Results: low k values work

Distributed ML

Krishna

ntroductio

IPM

Literatur Review

ADMN

Literatur

A general framework fo distributed optimization

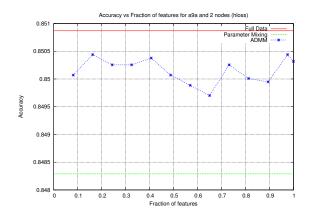


Figure: Dataset a9a split into two nodes: test accuracy

Results: Larger k means a better approximation

Distributed ML

Krishn:

ntroductio

IPM

Review

ADMI.

ADMN

Review
A general
framework for
distributed

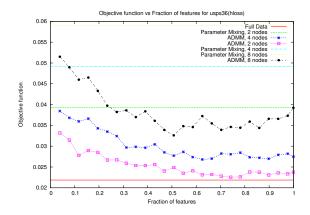


Figure: Dataset usps36: 2,4 and 8 nodes

Results: 25-100 Nodes

Distributed ML

Krishna Pillutla

ntroductio

Literatur Review

V DIVIV

Literatur

Review

framework fo distributed optimization

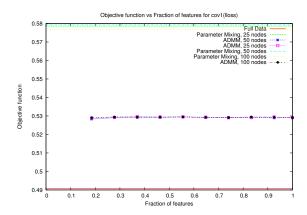


Figure: Dataset cov: 25, 50 and 100 nodes: objective value

Results: 25-100 Nodes

Distributed ML

Krishna

ntroductio

minoducti

Literatu

Algorit

ADMI

Literatur Review

A general framework for distributed optimization

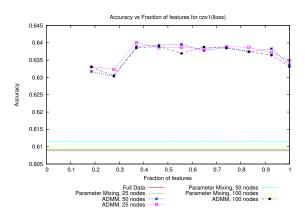


Figure: Dataset cov: 25, 50 and 100 nodes: Test Accuracy

Results: Do we need a PM step in the end?

Distributed ML

Pillutla

Introductio

IDM

Literature Review

ADMN

ADIVII Literat

Review A general framework

distributed optimization Approach

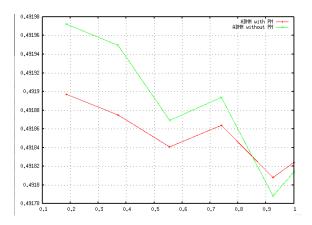


Figure: cov: 4 nodes: Comparision of objective function values ADMM with PM and ADMM without PM

Results: Do we need a PM step in the end?

Distributed ML

Krishna Pillutla

Introductio

Literatur Review

л DMM

Literatur Review

A general framework for distributed optimization

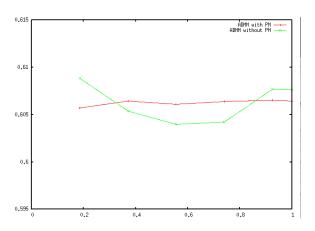


Figure: cov: 4 nodes: Comparision of test accuracy values ADMM with PM and ADMM without PM

Distributed ML

Experiments and Results

■ There is some merit in this method

Distributed ML

Krishn: Pillutla

ntroduction

Literatur Review

. _ . . .

ADIVIN

Review
A general framework for distributed optimization
Approach

- There is some merit in this method
- TODO: Theoretical Treatment

Distributed ML

Krishn: Pillutla

ntroductio

IDM

Literature Review

-A D I A I

ADMIN

Review
A general
framework for
distributed
optimization
Approach

- There is some merit in this method
- TODO: Theoretical Treatment
- TODO: Run experiments on Hadoop with larger datasets

Distributed ML

Krishn: Pillutla

ntroductio

IPM

Literatur Review Algorithr

-A D.M.A.

ADIVIII

A general framework for distributed optimization

- There is some merit in this method
- TODO: Theoretical Treatment
- TODO: Run experiments on Hadoop with larger datasets
- TODO: Grouping of features

Distributed ML

Krishn Pillutl:

ntroductio

....

Literatur Review Algorithr

-A D I A I

ADMN

Review A general

framework for distributed optimization Approach

- There is some merit in this method
- TODO: Theoretical Treatment
- TODO: Run experiments on Hadoop with larger datasets
- TODO: Grouping of features
 - Group similar features together

Distributed

- There is some merit in this method
- TODO: Theoretical Treatment
- TODO: Run experiments on Hadoop with larger datasets
- TODO: Grouping of features
 - Group similar features together
 - Clustering of features

Distributed

- There is some merit in this method
- TODO: Theoretical Treatment
- TODO: Run experiments on Hadoop with larger datasets
- TODO: Grouping of features
 - Group similar features together
 - Clustering of features
 - Sampling of data

References I

Distributed ML

Krishn Pillutl

Introductio

Literature Review

а Білі

Literatu

A general framework for distributed optimization Approach

Experiments and Results

[ACDL11] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford, *A reliable effective terascale linear learning system*, CoRR **abs/1110.4198** (2011).

[BPC+11] C S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, 2011.

[BT97] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Athena scientific optimization and computation series, Athena Scientific, 1997.

References II

Distributed ML

Pillutl

ntroductio

Literatu Review

A D. A. A.

Literatu

A general framework for distributed optimization

Experiments and Results

[HMS08] Tamir Hazan, Amit Man, and Amnon Shashua, A parallel decomposition solver for svm: Distributed dual ascend using fenchel duality.

[Man] O. L. Mangasarian, Parallel gradient distribution.

[MHM] Ryan Mcdonald, Keith Hall, and Gideon Mann, Distributed training strategies for the structured perceptron.

[MKSB13] Dhruv Mahajan, S. Sathiya Keerthi, S. Sundararajan, and Léon Bottou, A functional approximation based distributed learning algorithm, CoRR abs/1310.8418 (2013).

References III

Distributed ML

Krishna

Introductio

Literatur Review

Algoriti

ADMN

Review
A general framework for distributed optimization

Experiments and Results

[MMM⁺09] Gideon Mann, Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, and Daniel D. Walker, *Efficient large-scale distributed training of conditional maximum entropy models*, In Advances in Neural Information Processing Systems, 2009.

Distributed ML

Krishna Pillutla

Introductio

IPM

Review

л БИИ

Literature

A general framework fo distributed optimization

Experiments and Results

The End. Thank You!