Alternating Direction Method of Multipliers (ADMM)

Илья Трофимов

13.10.2016

Машинное обучение и большие данные
ФИВТ, осень 2016

ADMM

Задача условной оптимизации

$$x^* = \operatorname*{argmin}_x f(x)$$

при ограничении $Ax = b$

$$x \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$$

ADMM

Задача условной оптимизации

$$x^* = \underset{x}{\operatorname{argmin}} f(x)$$

при ограничении $Ax = b$

$$x \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$$

Построим Лагранжиан

$$L(x, y) = f(x) + y^T(Ax - b)$$

$y \in \mathbb{R}^p$ - множители Лагранжа

ADMM

Двойственная задача

$$y^* = \operatorname*{argmax}_y g(y)$$

$$g(y) = \inf_x L(x, y)$$

ADMM

Двойственная задача

$$y^* = \operatorname*{argmax}_y g(y)$$

$$g(y) = \inf_x L(x, y)$$

Условная оптимизация сводится к безусловной

$$x^* = \operatorname*{argmin}_x L(x, y^*)$$

Dual Ascent

Идея: делать шаги по градиенту y

Dual Ascent

Идея: делать шаги по градиенту y

$$x^+(y) = \operatorname*{argmin}_x L(x, y)$$

$$g(y) = L(x^+(y), y)$$

$$\begin{aligned}
\nabla g(y) &= \frac{\partial L}{\partial x}(x^+(y), y)\frac{dx^+}{dy} &+ \quad \frac{\partial L}{\partial y}(x^+(y), y) \\
&= 0 &+ \quad Ax^+(y) - b \\
&= Ax^+(y) - b
\end{aligned}$$

Dual Ascent

Идея: делать шаги по градиенту y

$$x^+(y) = \operatorname*{argmin}_x L(x, y)$$

$$g(y) = L(x^+(y), y)$$

$$\begin{aligned}\nabla g(y) &= \frac{\partial L}{\partial x}(x^+(y), y)\frac{dx^+}{dy} + \frac{\partial L}{\partial y}(x^+(y), y) \\ &= 0 + Ax^+(y) - b \\ &= Ax^+(y) - b\end{aligned}$$

Итерация алгоритма Dual Ascent:

$$x^{k+1} \leftarrow \operatorname*{argmin}_x L(x, y^k)$$

$$y^{k+1} \leftarrow y^k + \alpha_k(Ax^{k+1} - b)$$

Dual Decomposition

Если $f(\cdot)$ сепарабельная

$$f(x) = \sum_{i=1}^{N} f_i(x_i), \quad x = (x_1, \ldots, x_N)$$

то Лагранжиан тоже сепарабелен

$$L(x, y) = \sum_{i=1}^{N} L(x_i, y) - y^T b$$

$$L_i(x_i, y) = \sum_{i=1}^{N} f_i(x_i) + y^T (A_i x_i - b)$$

Dual Decomposition

Если $f(\cdot)$ сепарабельная

$$f(x) = \sum_{i=1}^{N} f_i(x_i), \quad x = (x_1, \ldots, x_N)$$

то Лагранжиан тоже сепарабелен

$$L(x, y) = \sum_{i=1}^{N} L(x_i, y) - y^T b$$

$$L_i(x_i, y) = \sum_{i=1}^{N} f_i(x_i) + y^T (A_i x_i - b)$$

Итерация алгоритма Dual decomposition:

$$x_i^{k+1} \leftarrow \operatorname*{argmin}_{x_i} L_i(x_i, y^k), \quad i = 1, \ldots, N$$

$$y^{k+1} \leftarrow y^k + \alpha_k (\sum_{i=1}^{N} A_i x^{k+1} - b)$$

Method of Multipliers

На практике лучше работает обобщенный Лагранжиан

$$L_\rho(x, y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|^2$$

Method of Multipliers

На практике лучше работает обобщенный Лагранжиан

$$L_\rho(x, y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|^2$$

Который является "обычным Лагранжианом" для задачи

$$x^* = \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2)\|Ax - b\|^2\right)$$

при ограничении $Ax = b$

$x \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$

Method of Multipliers

На практике лучше работает обобщенный Лагранжиан

$$L_\rho(x, y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|^2$$

Который является "обычным Лагранжианом" для задачи

$$x^* = \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2)\|Ax - b\|^2\right)$$

при ограничении $Ax = b$

$x \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$

Итерация алгоритма Method of Multipliers:

$$x^{k+1} \leftarrow \underset{x}{\operatorname{argmin}}\, L_\rho(x, y^k)$$

$$y^{k+1} \leftarrow y^k + \rho(Ax^{k+1} - b)$$

Method of Multipliers

Почему $\alpha_k = \rho$?

Method of Multipliers

Почему $\alpha_k = \rho$?
Для решения (x^*, y^*) должно выполняться

$$Ax^* - b = 0 \qquad \nabla f(x^*) + A^T y^* = 0$$

Method of Multipliers

Почему $\alpha_k = \rho$?
Для решения (x^*, y^*) должно выполняться

$$Ax^* - b = 0 \qquad \nabla f(x^*) + A^T y^* = 0$$

$$\begin{aligned} 0 &= \nabla_x L_\rho(x^{k+1}, y^k) \\ &= \nabla f(x^{k+1}) + A^T(y^k + \rho(Ax^{k+1} - b)) \\ &= \nabla f(x^{k+1}) + A^T y^{k+1} \end{aligned}$$

ADMM

Задача условной оптимизации

$$\underset{x,z}{\operatorname{argmin}} \left(f(x) + g(z)\right)$$

при ограничении $Ax + Bz = c$

$$x \in \mathbb{R}^n, z \in \mathbb{R}^m, A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m}, c \in \mathbb{R}^p$$

ADMM

Построим обобщенный Лагранжиан

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|^2$$

$y \in \mathbb{R}^p$ - множители Лагранжа

Итерация алгоритма ADMM:

$$x^{k+1} \leftarrow \operatorname*{argmin}_x L_\rho(x, z^k, y^k)$$
$$z^{k+1} \leftarrow \operatorname*{argmin}_z L_\rho(x^{k+1}, z, y^k)$$
$$y^{k+1} \leftarrow y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

Нормированная форма ADMM

Введем обозначение для невязки $r = Ax + Bz - c$

$$\begin{aligned} y^T + (\rho/2)\|r\|^2 &= (\rho/2)\|r + (1/\rho)y\|^2 - (1/2\rho)\|y\|^2 \\ &= (\rho/2)\|r + u\|^2 - (\rho/2)\|u\|^2 \end{aligned}$$

где $u = (1/\rho)y$ - нормированный множитель Лагранжа

Нормированная форма ADMM

Введем обозначение для невязки $r = Ax + Bz - c$

$$\begin{aligned} y^T + (\rho/2)\|r\|^2 &= (\rho/2)\|r + (1/\rho)y\|^2 - (1/2\rho)\|y\|^2 \\ &= (\rho/2)\|r + u\|^2 - (\rho/2)\|u\|^2 \end{aligned}$$

где $u = (1/\rho)y$ - нормированный множитель Лагранжа

Итерация алгоритма ADMM, нормированная форма:

$$x^{k+1} \leftarrow \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2)\|Ax + Bz^k - c + u^k\|^2 \right)$$

$$z^{k+1} \leftarrow \underset{z}{\operatorname{argmin}} \left(g(z) + (\rho/2)\|Ax^{k+1} + Bz - c + u^k\|^2 \right)$$

$$u^{k+1} \leftarrow u^k + Ax^{k+1} + Bz^{k+1}$$

Сходимость ADMM

При выполнении условий:

1. Функции f, g - собственные выпуклые функции, а также для каждого $\alpha \in \mathbb{R}$, множество $\{x \in \mathrm{dom}(f) \mid f(x) \leq \alpha\}$ замкнуто
2. У Лагранжиана существует седловая точка $L_0(x^*, z^*, y) \leq L_0(x^*, z^*, y^*) \leq L_0(x, z, y^*)$

Можно доказать результаты про сходимость:

1. Невязка стремится к нулю $r^k \to 0$
2. Целевая функции стремится к минимуму $f(x^k) + g(z^k) \to min$
3. Двойственная переменная сходится $y^k \to y^*$

Пример. Условная оптимизация

$$x^* = \operatorname*{argmin}_x f(x)$$

при ограничении $x \in \mathcal{C}$

Пример. Условная оптимизация

$$x^* = \operatorname*{argmin}_x f(x)$$

при ограничении $x \in \mathcal{C}$

Индикаторная функция множества $\mathcal{C}$:

$$g(x) = 0, \text{ если } x \in \mathcal{C}, \text{ иначе } +\infty$$

Пример. Условная оптимизация

Итерация алгоритма ADMM, нормированная форма:

$$x^{k+1} \leftarrow \operatorname*{argmin}_x \left(f(x) + (\rho/2)\|x - z^k + u^k\|^2 \right)$$

$$z^{k+1} \leftarrow \operatorname*{argmin}_z \left(g(z) + (\rho/2)\|x^{k+1} - z + u^k\|^2 \right)$$

$$u^{k+1} \leftarrow u^k + x^{k+1} - z^{k+1}$$

Пример. Условная оптимизация

Итерация алгоритма ADMM, нормированная форма:

$$x^{k+1} \leftarrow \operatorname*{argmin}_x \left(f(x) + (\rho/2)\|x - z^k + u^k\|^2 \right)$$
$$z^{k+1} \leftarrow \operatorname*{argmin}_z \left(g(z) + (\rho/2)\|x^{k+1} - z + u^k\|^2 \right)$$
$$u^{k+1} \leftarrow u^k + x^{k+1} - z^{k+1}$$

$$x^{k+1} \leftarrow \operatorname*{argmin}_x \left(f(x) + (\rho/2)\|x - z^k + u^k\|^2 \right)$$
$$z^{k+1} \leftarrow \Pi_{\mathcal{C}}(x^{k+1} + u^k) \qquad \text{/* операция проектирования */}$$
$$u^{k+1} \leftarrow u^k + x^{k+1} - z^{k+1}$$

Пример. Отбор признаков

$$x^* = \underset{x}{\operatorname{argmin}} \|Ax - b\|^2$$

при ограничении $card(x) \leq c$

где $card(x)$ - кол-во ненулевых компонент x.

Пример. Отбор признаков

$$x^* = \underset{x}{\operatorname{argmin}} \|Ax - b\|^2$$

при ограничении $card(x) \leq c$

где $card(x)$ - кол-во ненулевых компонент x.

$$x^{k+1} \leftarrow (A^T A + \rho I)^{-1}(A^T b + \rho(z^k - u^k))$$
$$z^{k+1} \leftarrow \Pi_{\mathcal{C}}(x^{k+1} + u^k)$$
$$u^{k+1} \leftarrow u^k + x^{k+1} - z^{k+1}$$

Consensus

Часто целевая функция аддитивна и состоит из нескольких компонент:

$$f(x) = \sum_{i=1}^{N} f_i(x)$$

Consensus

Часто целевая функция аддитивна и состоит из нескольких компонент:

$$f(x) = \sum_{i=1}^{N} f_i(x)$$

Пример: функции эмпирического риска в машинном обучении

$$\text{минимизировать по } x, z \quad f(x) = \sum_{i=1}^{N} f_i(x_i)$$

$$\text{при ограничениях} \quad x_i - z = 0, \quad i = 1, ..., N$$

Consensus

$$L_\rho(x_1, ..., x_N, z, y) = \sum_{i=1}^{N} \left(f_i(x_i) + y_i^T (x_i - z) + \rho/2 ||x_i - z||^2 \right)$$

Алгоритм:

$$x_i^{k+1} \leftarrow \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + \rho/2 ||x_i - \bar{x}^k - (1/\rho) y_i^k||^2 \right) \quad (1)$$

$$y_i^{k+1} \leftarrow y_i^k + \rho(x_i^{k+1} - \bar{x}^{k+1}) \quad (2)$$

$$z^{k+1} \leftarrow \bar{x}^{k+1} + (1/\rho)\bar{y}^k \quad (3)$$

где $\bar{x}^k = \sum_{i=1}^N x_i, \quad \bar{y}^k = \sum_{i=1}^N y_i$

Consensus

$$L_\rho(x_1, ..., x_N, z, y) = \sum_{i=1}^{N} \left(f_i(x_i) + y_i^T (x_i - z) + \rho/2 ||x_i - z||^2 \right)$$

Алгоритм:

$$x_i^{k+1} \leftarrow \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + \rho/2 ||x_i - \bar{x}^k - (1/\rho) y_i^k||^2 \right) \quad (1)$$

$$y_i^{k+1} \leftarrow y_i^k + \rho(x_i^{k+1} - \bar{x}^{k+1}) \quad (2)$$

$$z^{k+1} \leftarrow \bar{x}^{k+1} + (1/\rho)\bar{y}^k \quad (3)$$

где $\bar{x}^k = \sum_{i=1}^N x_i, \quad \bar{y}^k = \sum_{i=1}^N y_i$
Шаги (1) и (2) распараллеливаются.

Consensus + regularization

Пример:

Consensus + regularization

Пример: многие функции эмпирического риска в машинном обучении

$$\begin{aligned} &\text{минимизировать} \quad f(x) = \sum_{i=1}^{N} f_i(x_i) + g(z) \\ &\text{при ограничениях} \quad x_i - z = 0, \quad i = 1, ..., N \end{aligned}$$

Consensus + regularization

Пример: многие функции эмпирического риска в машинном обучении

$$\begin{aligned} &\text{минимизировать} \quad f(x) = \sum_{i=1}^{N} f_i(x_i) + g(z) \\ &\text{при ограничениях} \quad x_i - z = 0, \quad i = 1, ..., N \end{aligned}$$

$f_i(x_i)$ - эмпирический риск по обучающим примерам, лежащим на машине i
$g(z)$ - регуляризация

Consensus + regularization

Алгоритм:

$$x_i^{k+1} \leftarrow \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + \rho/2 \|x_i - z^k - (1/\rho) y_i^k\|^2 \right)$$

$$z^{k+1} \leftarrow \underset{z}{\operatorname{argmin}} \left(g(z) + \sum_{i=1}^{N} \left(-(y_i^k)^T z + (\rho/2)\|x_i^{k+1} - z\|^2 \right) \right)$$

$$y_i^{k+1} \leftarrow y_i^k + \rho(x_i^{k+1} - z^{k+1})$$

Sharing

$$\text{минимизировать } \sum_{i=1}^{N} f_i(x_i) + g(\sum_{i=1}^{N} x_i)$$

где $x_i \in \mathbb{R}^n$.
$g(\cdot)$ - общая целевая функция
$f(\cdot)$ - индивидуальные штрафы

Sharing

$$\text{минимизировать} \sum_{i=1}^{N} f_i(x_i) + g(\sum_{i=1}^{N} x_i)$$

где $x_i \in \mathbb{R}^n$.
$g(\cdot)$ - общая целевая функция
$f(\cdot)$ - индивидуальные штрафы

В форме ADMM

$$\text{минимизировать} \quad \sum_{i=1}^{N} f_i(x_i) + g(\sum_{i=1}^{N} z_i)$$

$$\text{при ограничениях} \quad x_i - z_i = 0, \quad i = 1, ..., N$$

где $x_i \in \mathbb{R}^n, z_i \in \mathbb{R}^n$

Sharing

Зачем это все нужно?

Sharing

Зачем это все нужно?

Если размерность x велика, то

$$x = (1, ...2, 7, ..., 1, 3, ..., 9)$$
$$x_1 = (1, ...2, 0, ..., 0, 0, ..., 0)$$
$$x_2 = (0, ...0, 7, ..., 1, 0, ..., 0)$$
$$x_3 = (0, ...0, 0, ..., 0, 3, ..., 9)$$
$$x = x_1 + x_2 + x_3$$

Sharing

Зачем это все нужно?
Если размерность x велика, то

$$x = (1, ...2, 7, ..., 1, 3, ..., 9)$$
$$x_1 = (1, ...2, 0, ..., 0, 0, ..., 0)$$
$$x_2 = (0, ...0, 7, ..., 1, 0, ..., 0)$$
$$x_3 = (0, ...0, 0, ..., 0, 3, ..., 9)$$
$$x = x_1 + x_2 + x_3$$

$g(x)$ - эмпирический риск
$f_i(x_i)$ - регуляризатор по подмножеству переменных (блочно-сепарабельный)
Например: $f_i(x_i) = \lambda \|x_i\|_p$

Sharing

$$x_i^{k+1} \leftarrow \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + (\rho/2)||x_i - z_i^k + u_i^k||^2 \right)$$

$$z_i^{k+1} \leftarrow \underset{z}{\operatorname{argmin}} \left(g(\sum_{i=1}^{N} z_i) + (\rho/2) \sum_{i=1}^{N} ||z_i - u_i^k - x_i^{k+1}||^2 \right)$$

$$u_i^{k+1} \leftarrow u_i^k + x_i^{k+1} - z_i^{k+1}$$

Задача ADMM в постановке sharing является двойственной к consensus

Разделение обучающей выборки

- По обучающим примерам - consensus

Разделение обучающей выборки

- По обучающим примерам - consensus
- По переменным - sharing

Разделение обучающей выборки

- По обучающим примерам - consensus
- По переменным - sharing
- Одновременно по примерам и переменным (Parikh, Boyd, 2011).

Применения

Применения для задач машинного обучения:

- Линейная регрессия с L1/L2 регуляризацией
- Логистическая регрессия с L1/L2 регуляризацией
- Линейный SVM
- Group LASSO
- Graphical LASSO
- Можно применять для невыпуклых целевых функций (сходимость не гарантирована). Например, для non-negative matrix factorization

Практические комментарии

На скорость сходимости существенно влияет параметр ρ.
Варианты:

1. Подобрать оптимальное ρ, сделав несколько итераций
2. Адаптивно менять ρ в зависимости от невязки (можно доказать сверхлинейную сходимость, если $\rho^k \to +\infty$)
3. $\rho = 1$ часто работает неплохо

Практические комментарии

На скорость сходимости существенно влияет параметр ρ.
Варианты:

1. Подобрать оптимальное ρ, сделав несколько итераций
2. Адаптивно менять ρ в зависимости от невязки (можно доказать сверхлинейную сходимость, если $\rho^k \to +\infty$)
3. $\rho = 1$ часто работает неплохо

Алгоритм ADMM обладает неплохой начальной сходимостью, но медленно сходится к точному решению. В машинном обучении нам нужно не найти минимум целевой функции с точность 10^{-6}, а чтобы качество прогноза на тесте было высоким. Так что OK.

Реализация ADMM на MPI

Algorithm 1 Global consensus ADMM in MPI.

initialize N processes, along with x_i, u_i, r_i, z.

repeat

1. Update $u_i := u_i + x_i - z$.
2. Update $x_i := \operatorname{argmin}_x \left(f_i(x) + (\rho/2)\|x - z + u_i\|_2^2\right)$.
3. Let $w := x_i + u_i$ and $t := \|r_i\|_2^2$.
4. *Allreduce* w and t.
5. Let $z^{\text{prev}} := z$ and update $z := \mathbf{prox}_{g, N\rho}(w/N)$.
6. **exit if** $\rho\sqrt{N}\|z - z^{\text{prev}}\|_2 \leq \epsilon^{\text{conv}}$ and $\sqrt{t} \leq \epsilon^{\text{feas}}$.
7. Update $r_i := x_i - z$.

Реализация ADMM на Hadoop

Algorithm 2 An iteration of global consensus ADMM in Hadoop/ MapReduce.

function map(key i, dataset $\mathcal{D}_i$)

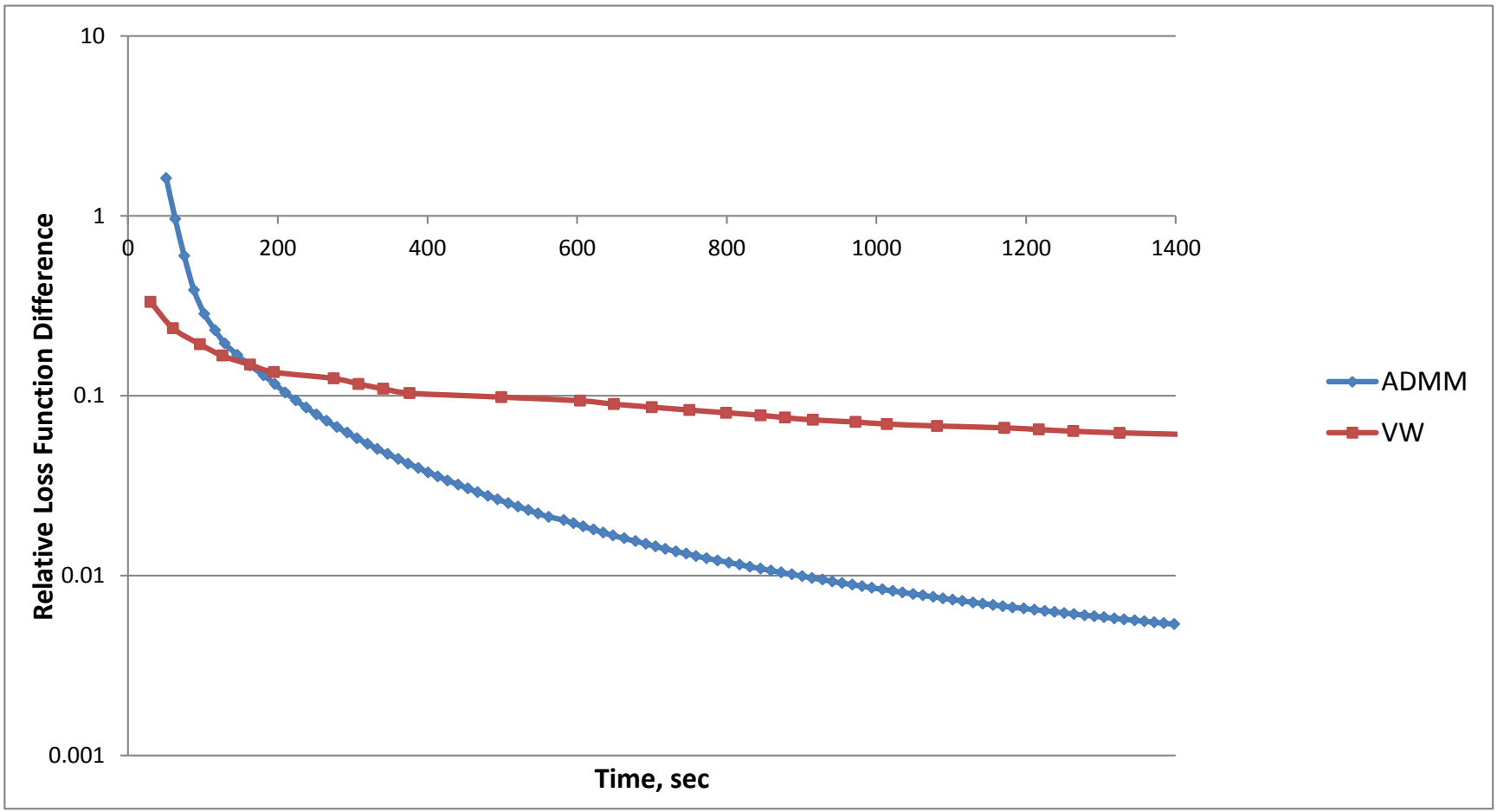
1. Read $(x_i, u_i, \hat{z})$ from HBase table.
2. Compute $z := \mathbf{prox}_{g, N\rho}((1/N)\hat{z})$.
3. Update $u_i := u_i + x_i - z$.
4. Update $x_i := \operatorname{argmin}_x \left(f_i(x) + (\rho/2)\|x - z + u_i\|_2^2\right)$.
5. *Emit* (key CENTRAL, record (x_i, u_i)).

function reduce(key CENTRAL, records $(x_1, u_1), \ldots, (x_N, u_N)$)

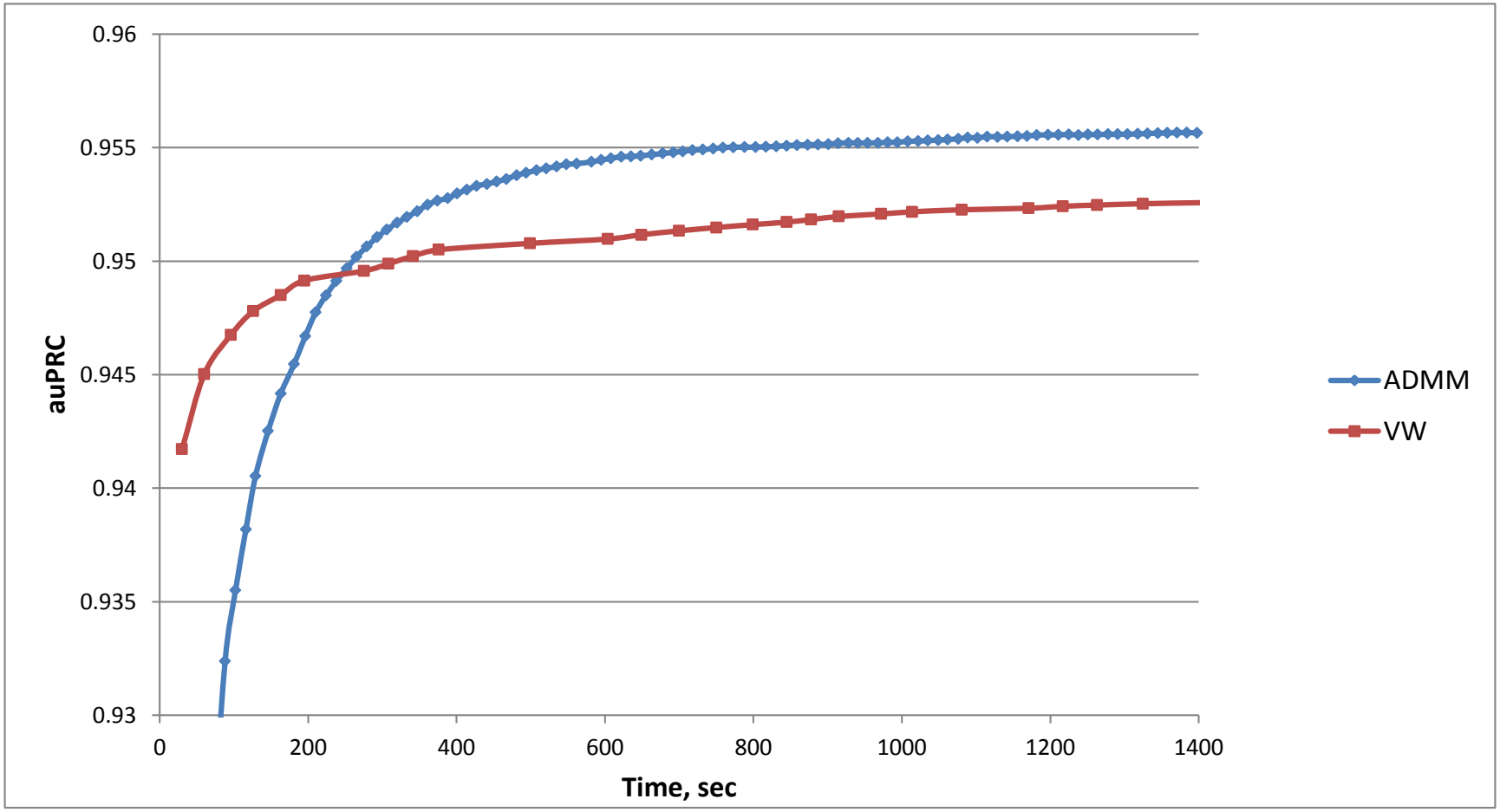
1. Update $\hat{z} := \sum_{i=1}^{N} x_i + u_i$.
2. *Emit* (key j, record $(x_j, u_j, \hat{z})$) to HBase for $j = 1, \ldots, N$.

Пример

Логистическая регрессия с L1-регуляризацией. Sharing + MPI.
Датасет "epsilon"
0.4×10^6 примеров, 2000 признаков, $N = 16$, $size = 16\,Gb$



Пример


0.96
0.955
0.95
0.945
0.94
0.935
0.93
auPRC
0
200
400
600
800
1000
1200
1400
Time, sec
ADMM
VW

Список литературы

ADMM S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, 2011.

Block Parikh, N., Boyd, S. (2011). Block Splitting for Large-Scale Distributed Learning. Neural Information Processing Systems (NIPS) Workshop on Big Learning.