Talk 1: Distributed Optimization and Statistical Learning via ADMM (I)

WEN Hao

2021-4-29

Main Resource: Chapter 7 of [1]

1 Recall of basic ADMM

A general ADMM optimization problem is formulated as

$$\begin{array}{ll} \text{minimize} & f(x)+g(z) \\ \text{subject to} & Ax+Bz=c \end{array}$$

The augmented Lagrangian of this problem is given by

$$\mathcal{L}_\rho(x,z,y)=f(x)+g(z)+\langle y, Ax+Bz-c\rangle+\frac{1}{\rho}\|Ax+Bz-c\|^2.$$

The iterations are given by

$$\begin{aligned} x^{k+1} &= \arg\min_x \left\{\mathcal{L}_\rho(x,z^k,y^k)\right\} \\ z^{k+1} &= \arg\min_z \left\{\mathcal{L}_\rho(x^{k+1},z,y^k)\right\} \\ y^{k+1} &= y^k+\rho(Ax^{k+1}+Bz^{k+1}-c) \end{aligned}$$

Convergence of ADMM: under the conditions

- f, g closed, proper convex;
- $\mathcal{L}_0(x,z,y)$ has a saddle point,

as $k \to \infty$ one has

- feasibility: $Ax^k + By^k - c \to 0$
- objective: $f(x^k) + g(y^k) \to p^*$
- dual: $y^k \to y^*$

2 Consensus Problem

Assume we have global variable $x \in \mathbb{R}^n$ and "split" (or distributed) objective function

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

e.g. x can be (global) model parameters, DNN weights (and biases, etc.), f_i can be the loss function associated with the i-th block ("client") of data. The optimization problem is

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x)$$

Problem: f is NOT block-separable.

Solution: add a common global variable $z \in \mathbb{R}^n$, so that the optimization problem is formulated as (equivalent to)

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{N} f_i(x_i) \\ &\text{subject to} && x_i - z = 0, \quad i = 1, \cdots, N \end{aligned}$$

which is called a **consensus problem**. One has the augmented Lagrangian

$$\mathcal{L}_\rho(x_1, \cdots, x_N, z, y) = \sum_{i=1}^{N} \left[f_i(x_i) + \langle y_i, x_i - z \rangle + \frac{\rho}{2} \|x_i - z\|^2 \right],$$

and ADMM iterations

$$\left(x = (x_1^T, \cdots, x_N^T)^T, x^{k+1} = \arg\min_x \left\{ \sum_{i=1}^{N} \left[f_i(x_i) + \langle y_i^k, x_i - z^k \rangle + \frac{\rho}{2} \|x_i - z^k\|^2 \right] \right\} \right)$$

$$
\begin{aligned}
\rightsquigarrow \quad x_i^{k+1} &= \arg\min_{x_i} \left\{ f_i(x_i) + \langle y_i^k, x_i - z^k \rangle + \frac{\rho}{2} \|x_i - z^k\|^2 \right\} \\
z^{k+1} &= \arg\min_{z} \left\{ \sum_{i=1}^{N} \left[f_i(x_i^{k+1}) + \langle y_i^k, x_i^{k+1} - z \rangle + \frac{\rho}{2} \|x_i^{k+1} - z\|^2 \right] \right\} \\
&= \arg\min_{z} \left\{ \frac{N\rho}{2} \|z\|^2 - \langle z, \sum_{i=1}^{N} (y_i^k + \rho x_i^{k+1}) \rangle + \cdots \right\} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i^k}{\rho} + x_i^{k+1} \right) \\
y_i^{k+1} &= y_i^k + \rho(x_i^{k+1} - z^{k+1})
\end{aligned}
$$

Simplify notations by letting

$$
\begin{cases}
\overline{x}^k := \frac{1}{N} \sum_{i=1}^{N} x_i^k \\
\overline{y}^k := \frac{1}{N} \sum_{i=1}^{N} y_i^k
\end{cases}
$$

then one has the following observations

$$
z^{k+1} = \frac{1}{N\rho} \sum_{i=1}^{N} y_i^k + \frac{1}{N} \sum_{i=1}^{N} x_i^{k+1} = \frac{\overline{y}^k}{\rho} + \overline{x}^{k+1}
$$

and

$$
\begin{aligned}
& y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - z^{k+1}) = y_i^k + \rho(x_i^{k+1} - \frac{\overline{y}^k}{\rho} - \overline{x}^{k+1}) \\
\Rightarrow \quad & \overline{y}^{k+1} = \overline{y}^k + \rho(\overline{x}_i^{k+1} - \frac{\overline{y}^k}{\rho} - \overline{x}^{k+1}) = 0 \\
\Rightarrow \quad & \overline{z}^{k+1} = \frac{0}{\rho} + \overline{x}^{k+1} = \overline{x}^{k+1}
\end{aligned}
$$

Then one can rewrite the iterations as

$$
\begin{aligned}
& x_i^{k+1} = \arg\min_{x_i} \left\{ f_i(x_i) + \langle y_i^k, x_i - \overline{x}^k \rangle + \frac{\rho}{2} \|x_i - \overline{x}^k\|^2 \right\} \\
& (z^{k+1} = \overline{x}^{k+1})
\end{aligned}
$$

$$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - \overline{x}^{k+1})$$

This can be further simplified by setting $u_i = \dfrac{y_i}{\rho}$:

$$\begin{aligned}
x_i^{k+1} &= \operatorname*{arg\,min}_{x_i} \left\{ f_i(x_i) + \frac{\rho}{2}\|x_i - \overline{x}^k + u_i^k\|^2 \right\} = \mathrm{prox}_{f_i,\rho}(\overline{x}^k - u_i^k) \\
(z^{k+1} &= \overline{x}^{k+1}) \\
u_i^{k+1} &= u_i^k + (x_i^{k+1} - \overline{x}^{k+1})
\end{aligned}$$

Statistical Interpretation for the ADMM iterations for consensus problem: at iteration $k+1$, assume x_i has prior distribution

$$x_i \sim N(\overline{x}^k - u_i^k, \rho I_n)$$

or equivalently

$$p(x_i) = \det(2\pi\rho I)^{-1/2} \exp\left(-\frac{1}{2}\|x_i - \overline{x}^k + u_i^k\|_{\rho I}^2 \right)$$

Remark 2.1 *As stated in [1], the bias of the mean of the above normal distribution, which is $-u_i^k$, can be interpreted as the "price" of "client" i disagreeing with the consensus $\overline{x}^k$ in the previous (the k -th) iteration. As for why the "price" is $-u_i^k$, note that the previous scaled dual update u_i^k is augmented by the bias of the k-th x_i-update, hence the accumulation of the biases of x_i-updates.*

Let

$$f_i(x_i) = \mathrm{NLL}(x_i) = -\log \mathrm{LH}(x_i)$$

be the negative log likelihood function[1] of x_i (w.r.t. the data (or observations) at the i-th "client"). Then the max a posteriori estimates (MAP) of the parameters x_i are

$$\begin{aligned}
\mathrm{MAP}(x_i) &= \operatorname*{arg\,max}_{x_i} \{p(x_i) \cdot \mathrm{LH}(x_i)\} \\
&= \operatorname*{arg\,max}_{x_i} \left\{ \exp(-f_i(x_i)) \cdot \det(2\pi\rho I)^{-1/2} \cdot \exp\left(-\frac{\rho}{2}\|x_i - \overline{x}_i^k + u_i^k\|^2 \right) \right\} \\
&= \operatorname*{arg\,min}_{x_i} \left\{ f_i(x_i) + \frac{\rho}{2}\|x_i - \overline{x}_i^k + u_i^k\|^2 \right\} = x_i^{k+1}
\end{aligned}$$

i.e. the $(k+1)$-th update of x_i are just the MAP of x_i with given prior distribution.

[1] for a good visualization of NLL, ref. https://ljvmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/

Remark 2.2 *Most federated learning optimization algorithms fall into paradigm of this basic consensus problem (with inexact inner minimization loops), including FedAvg [2], FedOpt(FedAdam, FedAdagrad, ...) [3], etc.*

3 Consensus with Regularization

Consider the problem

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \boxed{\text{g(z)}} \\ \text{subject to} \quad & x_i - z = 0, \quad i = 1, \cdots, N \end{aligned}$$

regularization on consensus

with regularization term $g(z)$ in the objective function.

The ADMM iterations:

$$\begin{aligned} x_i^{k+1} &= \arg\min_{x_i} \left\{ f_i(x_i) + \langle y_i^k, x_i - z^k \rangle + \frac{\rho}{2} \|x_i - z^k\|^2 \right\} \\ z^{k+1} &= \arg\min_{z} \left\{ g(z) + \sum_{i=1}^{N} \left(\langle y_i^k, x_i^{k+1} - z \rangle + \frac{\rho}{2} \|x_i^{k+1} - z\|^2 \right) \right\} \\ y_i^{k+1} &= y_i^k + \rho(x_i^{k+1} - z^{k+1}) \end{aligned}$$

has no analytic expression in general

One similarly has the following reductions by letting $u_i = \dfrac{y_i}{\rho}$:

$$\begin{aligned} z^{k+1} &= \arg\min_{z} \left\{ g(z) + \sum_{i=1}^{N} \left(\frac{\rho}{2} \|z\|^2 - \langle \rho x_i^{k+1} + y_i^k, z \rangle + \cdots \right) \right\} \\ &= \arg\min_{z} \left\{ g(z) + N \left(\frac{\rho}{2} \|z\|^2 - \langle \rho \overline{x}^{k+1} + \overline{y}^k, z \rangle + \cdots \right) \right\} \\ &= \arg\min_{z} \left\{ g(z) + \frac{N\rho}{2} \|z - \overline{x}^{k+1} - \frac{\overline{y}^k}{\rho}\|^2 \right\} \\ &= \arg\min_{z} \left\{ g(z) + \frac{N\rho}{2} \|z - \overline{x}^{k+1} - \overline{u}^k\|^2 \right\} = \text{prox}_{g, N\rho}(\overline{x}^{k+1} + \overline{u}^k) \\ x_i^{k+1} &= \arg\min_{x_i} \left\{ f_i(x_i) + \frac{\rho}{2} \|x_i - z^k + u_i^k\|^2 \right\} = \text{prox}_{f_i, \rho}(z^k - u_i^k) \\ u_i^{k+1} &= u_i^k + (x_i^{k+1} - z^{k+1}) \end{aligned}$$

Example 3.1

(1) $g(z) = \lambda\|z\|_1, \lambda > 0$, *then*

$$\begin{aligned} z^{k+1} &= \arg\min_z \left\{ \lambda\|z\|_1 + \frac{N\rho}{2}\|z - \overline{x}^{k+1} - \overline{u}^k\|^2 \right\} \\ &= \mathrm{S}_{\lambda/N\rho}(\overline{x}^{k+1} + \overline{u}^{k+1}) \quad \leftarrow \quad \textbf{soft thresholding} \end{aligned}$$

(2) $g(z) = I_{\mathbb{R}^n_+}(z)$ *the indicator function of* $\mathbb{R}^n_+$, *then*

$$\begin{aligned} z^{k+1} &= \arg\min_z \left\{ I_{\mathbb{R}^n_+}(z) + \frac{N\rho}{2}\|z - \overline{x}^{k+1} - \overline{u}^k\|^2 \right\} \\ &= (\overline{x}^{k+1} + \overline{u}^{k+1})_+ \end{aligned}$$

4 General Form Consensus

Now consider even more general setting:

$x_i \in \mathbb{R}^{n_i}, z \in \mathbb{R}^n$,
x_i consists of a selection of components of z,
i.e. $\forall i \in [1, N], \forall j \in [1, n_i], \exists \mathcal{G}(i,j)$ s.t. $(x_i)_j = z_{\mathcal{G}(i,j)}$

This general setting is of interest in cases where $n_i \ll n$, e.g. large global model and small local model (a small part of global params related to local data, corr. to vertical split of data?)

Let $\widetilde{z}_i \in \mathbb{R}^{n_i}$ be s.t. $(\widetilde{z}_i)_j = z_{\mathcal{G}(i,j)}$, then the general form consensus problem is formulated as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^N f_i(x_i) \\ \text{subject to} \quad & x_i - \widetilde{z}_i = 0 \end{aligned}$$

The augmented Lagrangian:

$$\mathcal{L}_\rho = \sum_{i=1}^N \left(f_i(x_i) + \langle y_i, x_i - \widetilde{z}_i \rangle + \frac{\rho}{2}\|x_i - \widetilde{z}_i\|^2 \right)$$

ADMM iterations:

$$x_i^{k+1} = \arg\min_{x_i} \left\{ f_i(x_i) + \langle y_i^k, x_i \rangle + \frac{\rho}{2} \|x_i - \widetilde{z}_i^k\|^2 \right\}$$

$$z^{k+1} = \arg\min_{z} \left\{ \sum_{i=1}^{N} \left(-\langle y_i^k, \widetilde{z}_i \rangle + \frac{\rho}{2} \|x_i^{k+1} - \widetilde{z}_i\|^2 \right) \right\}$$

$$y_i^{k+1} = y_i^k + \rho(x_i^{k+1} - \widetilde{z}_i^{k+1})$$

Rewrite

$$\begin{aligned}
z^{k+1} &= \arg\min_{z} \left\{ \sum_{i=1}^{N} \left(\frac{\rho}{2} \|\widetilde{z}_i\|^2 - \rho \langle x_i^{k+1} + \frac{1}{\rho} y_i^k, \widetilde{z}_i \rangle + \cdots \right) \right\} \\
&= \arg\min_{z} \left\{ \sum_{i=1}^{N} \frac{\rho}{2} \|\widetilde{z}_i - x_i^{k+1} - \frac{1}{\rho} y_i^k \|^2 \right\} \\
&= \arg\min_{z} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left((\widetilde{z}_i)_j - (x^{k+1})_j - \frac{1}{\rho} (y_i^k)_j \right)^2 \right\} \\
&\qquad \left(\text{since } \sum_{i=1}^{N} \sum_{j=1}^{n_i} = \sum_{g=1}^{n} \sum_{\mathcal{G}(i,j)=g} \right) \\
&= \arg\min_{z} \left\{ \sum_{g=1}^{n} \left[\sum_{\mathcal{G}(i,j)=g} \left(z_g - (x_i^{k+1})_j - \frac{1}{\rho} (y_i^k)_j \right)^2 \right] \right\} \\
&\qquad (\text{write } k_g = \#\{(i,j) | \mathcal{G}(i,j) = g\}) \\
&= \arg\min_{z} \left\{ \sum_{g=1}^{n} \left[k_g \cdot z_g^2 - 2 \sum_{\mathcal{G}(i,j)=g} \left((x_i^{k+1})_j + \frac{1}{\rho} (y_i^k)_j \right) + \cdots \right] \right\} \\
\Rightarrow \quad z_g^{k+1} &= \frac{1}{k_g} \sum_{\mathcal{G}(i,j)=g} \left((x_i^{k+1})_j + \frac{1}{\rho} (y_i^k)_j \right) \leftarrow \text{local average}
\end{aligned}$$

For the dual y-update, locally one has

$$\sum_{\mathcal{G}(i,j)=g} (y_i^{k+1})_j = \sum_{\mathcal{G}(i,j)=g} (y_i^k)_j + \rho \left(\sum_{\mathcal{G}(i,j)=g} (x_i^{k+1})_j - \sum_{\mathcal{G}(i,j)=g} (\widetilde{z}_i^{k+1})_j \right)$$

$$
\begin{aligned}
&= \sum_{\mathcal{G}(i,j)=g} (y_i^k)_j + \rho \left(\sum_{\mathcal{G}(i,j)=g} (x_i^{k+1})_j - k_g \cdot z_g^{k+1} \right) \\
&= \sum_{\mathcal{G}(i,j)=g} (y_i^k)_j + \rho \left(\sum_{\mathcal{G}(i,j)=g} (x_i^{k+1})_j - \sum_{\mathcal{G}(i,j)=g} \left((x_i^{k+1})_j + \frac{1}{\rho} (y_i^k)_j \right) \right) \\
&= 0 \\
\Rightarrow \quad & z_g^{k+1} = \frac{1}{k_g} \sum_{\mathcal{G}(i,j)=g} (x_i^{k+1})_j
\end{aligned}
$$

Hence the iterations simplifies to

$$
\begin{aligned}
x_i^{k+1} &= \underset{x_i}{\arg\min} \left\{ f_i(x_i) + \langle y_i^k, x_i \rangle + \frac{\rho}{2} \| x_i - \widetilde{z}_i^k \|^2 \right\} \\
&= \underset{x_i}{\arg\min} \left\{ f_i(x_i) + \frac{\rho}{2} \| x_i - \widetilde{z}_i^k + u_i^k \|^2 \right\} = \operatorname{prox}_{f_i,\rho}(\widetilde{z}_i^k - u_i^k) \\
z_g^{k+1} &= \frac{1}{k_g} \sum_{\mathcal{G}(i,j)=g} (x_i^{k+1})_j \\
u_i^{k+1} &= u_i^k + (x_i^{k+1} - \widetilde{z}_i^{k+1})
\end{aligned}
$$

5 General Form Consensus with Regularization

General form consensus + consensus with regularization:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \boxed{\text{g(z)}} \\
\text{subject to} \quad & x_i - \widetilde{z}_i = 0
\end{aligned}
$$

ADMM iterations:

$$
\begin{aligned}
x_i^{k+1} &= \underset{x_i}{\arg\min} \left\{ f_i(x_i) + \langle y_i^k, x_i - \widetilde{z}_i^k \rangle + \frac{\rho}{2} \| x_i - \widetilde{z}_i^k \|^2 \right\} \\
&= \operatorname{prox}_{f_i,\rho}(\widetilde{z}_i^k - u_i^k) \\
z^{k+1} &= \underset{z}{\arg\min} \left\{ g(z) + \sum_{i=1}^{N} \left(-\langle y_i^k, \widetilde{z}_i^k \rangle + \frac{\rho}{2} \| x_i^{k+1} - \widetilde{z}_i \|^2 \right) \right\}
\end{aligned}
$$

$$= \mathrm{prox}_{g,k_g\rho}(v_{\mathcal{G}})$$

$$\text{where } (v_{\mathcal{G}}) = (v_1, \cdots, v_n)^T, \text{ s.t. } v_g = \frac{1}{k_g} \sum_{\mathcal{G}(i,j)=g} \big((x_i^{k+1})_j + (u_i^k)_j\big)$$

$$u_i^{k+1} = u_i^k + (x_i^{k+1} - \tilde{z}_i^{k+1})$$

6 Sharing Problem

A sharing problem is an optimization problem formulated as

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^N f_i(x_i) + g(\sum_{i=1}^N x_i) \\ &\text{where } f_i\text{: local cost} \\ &\qquad\quad g\text{: shared cost} \end{aligned}$$

Remark 6.1 *A sharing problem is dual to a consensus problem.*

Indeed, rewrite a sharing problem in the ADMM form

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^N f_i(x_i) + g(\sum_{i=1}^N z_i) \\ &\text{subject to} \quad x_i - z_i = 0 \end{aligned}$$

Its dual function is

$$\begin{aligned} \Gamma(v_1, \cdots, v_N) &= \inf_{x,z} \left\{ \sum_{i=1}^N f_i(x_i) + g(\sum_{i=1}^N z_i) + \sum_{i=1}^N \langle v_i, x_i - z_i \rangle \right\} \\ &= \inf_x \left\{ \sum_{i=1}^N (f_i(x_i) + \langle v_i, x_i \rangle) \right\} + \inf_z \left\{ g(\sum_{i=1}^N z_i) - \sum_{i=1}^N \langle v_i, z_i \rangle \right\} \\ &= -\sup_x \left\{ \sum_{i=1}^N (-f_i(x_i) + \langle -v_i, x_i \rangle) \right\} + \inf_z \left\{ g(\sum_{i=1}^N z_i) - \sum_{i=1}^N \langle v_i, z_i \rangle \right\} \\ &= -\sum_{i=1}^N f_i^*(-v_i) + \boxed{\inf_z \left\{ g(\sum_{i=1}^N z_i) - \sum_{i=1}^N \langle v_i, z_i \rangle \right\}} \leftarrow \star \end{aligned}$$

For $\circledast$, assume $v_s \neq v_t$, and $\{z_i^*\}_{i=1}^N$ s.t. $\circledast > -\infty$. Then let $\{\widetilde{z}_i\}_{i=1}^N$ be s.t. $\widetilde{z}_i = z_i^*$ for $i \neq s, t$, $\widetilde{z}_s = z_s^* + w, \widetilde{z}_t = z_t^* - w, w \neq 0$, one has

$$\begin{aligned}
& g(\sum_{i=1}^N \widetilde{z}_i) - \sum_{i=1}^N \langle v_i, \widetilde{z}_i \rangle \\
=& g(\sum_{i=1}^N z_i^*) - \sum_{i=1}^N \langle v_i, z_i^* \rangle + \langle w, -v_s + v_t \rangle
\end{aligned}$$

One can always choose w so that $\langle w, -v_s + v_t \rangle < 0$, contradiction (with $v_s \neq v_t$ for some s, t). Hence

$$\begin{aligned}
\circledast &= \begin{cases} \inf\limits_z \left\{ g(\sum\limits_{i=1}^N z_i) - \langle v_1, \sum\limits_{i=1}^N z_i \rangle \right\}, & v_1 = \cdots = v_N \\ -\infty, & \text{otherwise} \end{cases} \\
&= \begin{cases} -g^*(v_1), & v_1 = \cdots = v_N \\ -\infty, & \text{otherwise} \end{cases}
\end{aligned}$$

i.e. the dual function is

$$\Gamma(v_1, \cdots, v_N) = \begin{cases} -g^*(v_1) - \sum\limits_{i=1}^N f_i^*(-v_i), & v_1 = \cdots = v_N \\ -\infty, & \text{otherwise} \end{cases}$$

and the dual problem is

$$\begin{aligned}
\text{minimize} \quad & g^*(v) + \sum_{i=1}^N f_i^*(-v_i) \\
\text{subject to} \quad & v_i = v
\end{aligned}$$

a consensus problem with regularization.

One can show that the dual of this consensus problem is the original sharing problem.

ADMM iterations for sharing problem:

$$x_i^{k+1} = \arg\min_{x_i} \left\{ f_i(x_i) + \frac{\rho}{2} \|x_i - z_i^k + u_i^k\|^2 \right\}$$

$$\boxed{z^{k+1} = \arg\min_{z} \left\{ g\Big(\sum_{i=1}^{N} z_i\Big) + \boxed{\frac{\rho}{2}\sum_{i=1}^{N} \|z_i - x_i^{k+1} - u_i^k\|^2} \right\}} \qquad z = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix}$$

$$u_i^{k+1} = u_i^k + (x_i^{k+1} - z_i^{k+1})$$

variables can be reduced from Nn to n

Write $a_i = u_i^k + x_i^{k+1}, \overline{z} = \dfrac{1}{N}\sum_{i=1}^{N} z_i$, then the $(k+1)$-th z-update is formulated as (equivalent to)

$$\begin{aligned} &\text{minimize} \quad g(N\overline{z}) + \frac{\rho}{2}\sum_{i=1}^{N}\|z_i - a_i\|^2 \\ &\text{subject to} \quad N\overline{z} - \sum_{i=1}^{N} z_i = 0 \end{aligned}$$

Since

$$\frac{\rho}{2}\sum_{i=1}^{N}\|z_i - a_i\|^2 \geqslant \frac{\rho}{2}\frac{\|\sum_{i=1}^{N}(z_i - a_i)\|^2}{N} = \frac{N\rho}{2}\|\overline{z} - \overline{a}\|^2$$

"=" holds only when $z_i = a_i + \overline{z} - \overline{a}$, i.e.

$$z_i^{k+1} = u_i^k + x_i^{k+1} + \overline{z}^{k+1} - \overline{u}^k - \overline{x}^{k+1}$$

Hence the constrained optimization problem of z-update is equivalent to the following unconstrained problem

$$\text{minimize} \qquad g(N\overline{z}) + \frac{N\rho}{2}\|\overline{z} - \overline{a}\|^2$$

Another consequence is

$$(\text{for simplicity } u^{k+1} =) u_1^{k+1} = \cdots = u_N^{k+1} = \overline{u}^k + \overline{x}^{k+1} - \overline{z}^{k+1}$$

and further

$$z_i^{k+1} = \boxed{\cancel{u_i^k}} + x_i^{k+1} + \overline{z}^{k+1} - \boxed{\cancel{\overline{u}^k}} - \overline{x}^{k+1} = x_i^{k+1} + \overline{z}^{k+1} - \overline{x}^{k+1}$$

The ADMM iterations for the whole equivalent optimization problem:

$$x_i^{k+1} = \arg\min_{x_i}\left\{ f_i(x_i) + \frac{\rho}{2}\|x_i - x_i^k - \overline{z}^k + \overline{x}^k + u^k\|^2 \right\} = \text{prox}_{f_i,\rho}(x_i^k + \overline{z}^k - \overline{x}^k - u^k)$$

$$\overline{z}^{k+1} = \arg\min_{\overline{z}} \left\{ g(N\overline{z}) + \frac{N\rho}{2}\|\overline{z} - \overline{x}^{k+1} - u^k\|^2 \right\} = \mathrm{prox}_{\widetilde{g},N\rho}(\overline{x}^{k+1} + u^k)$$

$$u^{k+1} = u^k + (\overline{x}^{k+1} - \overline{z}^{k+1})$$

where $\widetilde{g}(\overline{z}) = g(N\overline{z})$.

Problems NOT discussed (and difficult)

- convergence (rate) analysis of the optimization problems, e.g. [4]
- "infeasible" problems, e.g. totally distributed cases where there's no "central collector", e.g. [5, 6, 7], or "weak" consensus [8]
- new ADMM developments, e.g. [9]
- etc.

For example, in [8], the authors considered a "weak" consensus problem

$$\text{minimize} \quad \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2}\sum_{i=1}^{N}\|x_i - \overline{x}\|^2$$

which can be reformulated as constrained optimization problems

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2}\sum_{i=1}^{N}\|x_i - z\|^2 \\ \text{subject to} \quad & Nz - \sum_{i=1}^{N} x_i = 0 \end{aligned}$$

or

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} f_i(x_i) + \frac{\lambda}{2}\sum_{i=1}^{N}\|x_i\|^2 - \frac{\lambda N}{2}\|z\|^2 \\ \text{subject to} \quad & Nz - \sum_{i=1}^{N} x_i = 0 \end{aligned}$$

which is a nonconvex sharing problem considered in [10] (Eq. (3.2)). Under certain assumptions, this latter problem is a DC (difference-of-convex) programming problem. Note the difference with the a normal consensus problem with proximal term technique, e.g. as in [11].

References

[1] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc., 2011.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Artificial Intelligence and Statistics*, pp. 1273–1282, PMLR, 2017.

[3] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive Federated Optimization," in *International Conference on Learning Representations*, 2021.

[4] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," *arXiv preprint arXiv:1907.02189*, 2019.

[5] A. Elgabli, J. Park, A. S. Bedi, M. Bennis, and V. Aggarwal, "GADMM: Fast and Communication Efficient Framework for Distributed Machine Learning," *Journal of Machine Learning Research*, vol. 21, no. 76, pp. 1–39, 2020.

[6] C. B. Issaid, A. Elgabli, J. Park, and M. Bennis, "Communication Efficient Distributed Learning with Censored, Quantized, and Generalized Group ADMM," *arXiv preprint arXiv:2009.06459*, 2020.

[7] G. França and J. Bento, "Distributed Optimization, Averaging via ADMM, and Network Topology," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1939–1952, 2020.

[8] F. Hanzely and P. Richtárik, "Federated Learning of a Mixture of Global and Local Models," *arXiv preprint arXiv:2002.05516*, 2020.

[9] J. Bai, D. Han, H. Sun, and H. Zhang, "Convergence on a Symmetric Accelerated Stochastic ADMM with Larger Stepsizes," *arXiv preprint arXiv:2103.16154*, 2021.

[10] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[11] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *Proceedings of Machine Learning and Systems* (I. Dhillon, D. Papailiopoulos, and V. Sze, eds.), vol. 2, pp. 429–450, 2020.