Variables:

$$N := \text{number of data points} \tag{1}$$
$$P := \text{number of machines} \tag{2}$$
$$n_p := \text{number of data points on machine } p \tag{3}$$

Objective function:

$$\underbrace{\left(\frac{1}{N}\sum_{k=1}^{N} f_k(x)\right)}_{\text{loss}} + \underbrace{\left(\frac{\lambda}{2}||x||_2^2\right)}_{\text{regularizer}} \tag{4}$$

$$\mathcal{L}(x,\mu) = \underbrace{\left(\frac{1}{N}\sum_{i=1}^{P}\sum_{j=1}^{n_p} f_{ij}(x_i)\right)}_{\text{loss}} + \underbrace{\left(\frac{\lambda}{2P}\sum_{i=1}^{P}||x_i||_2^2\right)}_{\text{regularizer}} + \underbrace{\left(\sum_{i=1}^{P}\sum_{j=i+1}^{P}\mu_{ij}^T(x_i - x_j)\right)}_{\text{undirected equality constraints}} \tag{5}$$

Taking the gradient with respect to x_i and μ_{ij} we obtain:

$$\nabla_{x_i}\mathcal{L}(x,\mu) = \frac{1}{N}\sum_{j=1}^{n_p}\nabla_{x_i} f_{ij}(x_i) + \frac{\lambda}{P}x_i + \left(\sum_{j=i+1}^{P}\mu_{ij}\right) - \left(\sum_{j=1}^{i-1}\mu_{ji}\right) \tag{6}$$
$$\nabla_{\mu_{ij}}\mathcal{L}(x,\mu) = x_i - x_j \qquad \forall i < j \tag{7}$$

Notice in equation 6 that we subtract the μ_{ji} terms. This is because for $j < i$, the variable x_i occurs on the negative side of the sum $\sum_{i=1}^{P}\sum_{j=i+1}^{P}\mu_{ij}^T(x_i - x_j)$. Essentially we maintain the invariant that we only index μ_{ij} such that $i < j$ and hence for $j < i$ we reverse the index.

For primal and dual "learning rates" η_t and γ_t respectively we obtain the update equations:

$$x_i^{(t+1)} \leftarrow x_i^{(t)} - \eta_t\left(\frac{1}{N}\sum_{j=1}^{n_p}\nabla_{x_i} f_{ij}(x_i)\big|_{x_i^{(t)}} + \frac{\lambda}{P}x_i^{(t)} + \left(\sum_{j=i+1}^{P}\mu_{ij}\right) - \left(\sum_{j=1}^{i-1}\mu_{ji}\right)\right) \tag{8}$$
$$\mu_{ij}^{(t+1)} \leftarrow \mu_{ij}^{(t)} + \gamma_t(x_i^{(t+1)} - x_j^{(t+1)}) \qquad \forall i < j \tag{9}$$

On each machine i we store an array:

$$u_i[j] = \mu_{ij} \text{ if } i < j \text{ else } -\mu_{ji}$$

The update equations become:

$$x_i^{(t+1)} \leftarrow x_i^{(t)} - \eta_t\left(\frac{1}{N}\sum_{j=1}^{n_p}\nabla_{x_i} f_{ij}(x_i)\big|_{x_i^{(t)}} + \frac{\lambda}{P}x_i^{(t)} + \sum_{j=1}^{P}u_i^{(t)}[j]\right) \tag{10}$$
$$u_i^{(t+1)}[j] \leftarrow u_i^{(t)}[j] + \gamma_t(x_i^{(t+1)} - x_j^{(t+1)}) \tag{11}$$

0.1 Sanity Check

Suppose $x_1 = 1$ and $x_2 = 2$ and $u_i[j] = 0$ and we go can update the $u_i[j]$ values. Then on machine $i = 1$ we get:

$$u_1[2] \leftarrow 0 + \gamma_t(x_1 - x_2) = -1$$

which when added to the x_1 update equation will cause x_1 to increase (we subtract the gradient) in the direction of x_2. Conversely on machine $i = 2$ we have:

$$u_2[1] \leftarrow 0 + \gamma_t(x_2 - x_1) = 1$$

which when added to the x_2 update equation will cause x_2 to decrease in the direction of x_1.

1 Simplified Update Equations

It is possible to simplify the notation slightly further. Instead of tracking the individual lagrangian multipliers μ_{ij} in some complex array $u_i[j]$ we can instead just track the sum:

$$\sigma_i := \left(\sum_{j=i+1}^{P} \mu_{ij}\right) - \left(\sum_{j=1}^{i-1} \mu_{ji}\right) \tag{12}$$

$$\tag{13}$$

Plugging the sum into the primal update we obtain:

$$x_i^{(t+1)} \leftarrow x_i^{(t)} - \eta_t \left(\frac{1}{N} \sum_{j=1}^{n_p} \nabla_{x_i} f_{ij}(x_i)\big|_{x_i^{(t)}} + \frac{\lambda}{P} x_i^{(t)} + \sigma_i \right) \tag{14}$$

To derive the dual update we do the following:

$$\begin{aligned}
\sigma_i^{(t+1)} &= \left(\sum_{j=i+1}^{P} \leftarrow \mu_{ij}^{(t)} + \gamma_t (x_i^{(t+1)} - x_j^{(t+1)}) \right) - \left(\sum_{j=1}^{i-1} \leftarrow \mu_{ji}^{(t)} + \gamma_t (x_j^{(t+1)} - x_i^{(t+1)}) \right) && (15) \\
&= \left(\sum_{j=i+1}^{P} \mu_{ij}^{(t)} - \sum_{j=1}^{i-1} \mu_{ji}^{(t)} \right) + \gamma_t \left(\sum_{j=i+1}^{P} (x_i^{(t+1)} - x_j^{(t+1)}) - \sum_{j=1}^{i-1} (x_j^{(t+1)} - x_i^{(t+1)}) \right) && (16) \\
&= \sigma_i^{(t)} + \gamma_t \sum_{j=1}^{P} (x_i^{(t+1)} - x_j^{(t+1)}) && (17)
\end{aligned}$$

This leads to the following simplified update equations.

$$x_i^{(t+1)} \leftarrow x_i^{(t)} - \eta_t \left(\frac{1}{N} \sum_{j=1}^{n_p} \nabla_{x_i} f_{ij}(x_i)\big|_{x_i^{(t)}} + \frac{\lambda}{P} x_i^{(t)} + \sigma_i \right) \tag{18}$$

$$\sigma_i^{(t+1)} \leftarrow \sigma_i^{(t)} + \gamma_t \sum_{j=1}^{P} (x_i^{(t+1)} - x_j^{(t+1)}) \tag{19}$$

1.1 Sanity check

Suppose $x_1 = 1$ and $x_2 = 2$ and all $\sigma_i = 0$ then the σ updates will:

$$
\begin{aligned}
\sigma_1^{(t+1)} &\leftarrow 0 + \lambda\left((x_1 - x_1) + (x_1 - x_2)\right) && (20)\\
&= \lambda(x_1 - x_2) && (21)\\
&= -\lambda && (22)\\
\sigma_2^{(t+1)} &\leftarrow 0 + \lambda\left((x_2 - x_1) + (x_2 - x_2)\right) && (23)\\
&= \lambda(x_2 - x_1) && (24)\\
&= \lambda && (25)
\end{aligned}
$$

Which will cause x_1 to increase (we subtract the gradient) and x_2 to decrease.