Assumption 1 (β-smoothness): For any parameter W_1, W_2, each loss function F_i satisfies $F_i(W_1) \leq F_i(W_2) + (W_1 - W_2)^T \nabla F_i(W_2) + \beta/2 \parallel W_1 - W_2 \parallel_2^2$. It also implies that $\parallel \nabla F_i(W_1) - \nabla F_i(W_2) \parallel \leq \beta(W_1 - W_2)$.
Assumption 2 (μ-strongly convex): For any parameter W_1, W_2, each loss function F_i satisfies $F_i(W_1) \geq F_i(W_2) + (W_1 - W_2)^T \nabla F_i(W_2) + \mu/2 \parallel W_1 - W_2 \parallel_2^2$. It also implies that $\parallel \nabla F_i(W_1) - \nabla F_i(W_2) \parallel \geq \mu(W_1 - W_2)$.
Assumption 3 (Bounded gradient): The stochastic gradient is bounded as $\mathbb{E}(\parallel \nabla F_i(W_i^t) \parallel_2^2) \leq \sigma^2$.
Theorem 1: There exist two constants $A = 1 - 2\mu\eta, B = \eta\sigma^2/(2\mu)$. The gap between the optimal W^* and W_i^t follows: $\mathbb{E}(\parallel W_i^t - W^* \parallel_2^2) \leq A^t \parallel W_i^0 - W^* \parallel^2 + B$.
Proof: $\parallel W_i^{t+1} - W^* \parallel_2^2$

$=\parallel W_i^{t+1} - W_i^t + W_i^t - W^* \parallel_2^2$

$=\parallel W_i^{t+1}(M_i = 1) - W_i^t(M_i = 1) + W_i^{t+1}(M_i = 0) - W_i^t(M_i = 0) + \parallel W_i^t - W^* \parallel_2^2$

$=\parallel W_i^{t+1}(M_i = 1) - W_i^t(M_i = 1) + \parallel W_i^t - W^* \parallel_2^2$

$=\parallel -\eta \nabla F_i(W_i^t(M_i = 1)) + W_i^t - W^* \parallel_2^2$

$=\parallel \eta \nabla F_i(W_i^t(M_i = 1)) \parallel_2^2 + \parallel W_i^t - W^* \parallel_2^2 - 2\eta \nabla F_i(W_i^t(M_i = 1)(W_i^t - W^*)$

Then, we have

$\mathbb{E}(\parallel W_i^{t+1} - W^* \parallel_2^2)$

$= \eta^2 \mathbb{E}(\parallel \nabla F_i(W_i^t(M_i = 1) \parallel_2^2) + \parallel W_i^t - W^* \parallel_2^2 - 2\eta \nabla F_i(W_i^t(M_i = 1)) \mathbb{E}(\parallel W_i^t - W^* \parallel_2)$

$\leq \eta^2 \mathbb{E}(\parallel \nabla F_i(W_i^t(M_i = 1) \parallel_2^2) + \parallel W_i^t - W^* \parallel_2^2 - 2\eta(\nabla F_i\left(W_i^t(M_i = 1)\right) - \nabla F_i(W^*(M_i = 1))) \mathbb{E}(\parallel W_i^t - W^* \parallel_2)$

$\leq \eta^2 \mathbb{E}(\parallel \nabla F_i(W_i^t) \parallel_2^2) + \parallel W_i^t - W^* \parallel_2^2 - 2\eta(\nabla F_i(W_i^t) - \nabla F_i(W^*)) \mathbb{E}(\parallel W_i^t - W^* \parallel_2)$

$\leq \eta^2 \mathbb{E}(\parallel \nabla F_i(W_i^t) \parallel_2^2) + \parallel W_i^t - W^* \parallel_2^2 - 2\eta\mu \parallel W_i^t - W^* \parallel_2^2$

$\leq \eta^2\sigma^2 + (1 - 2\eta\mu) \mathbb{E}(\parallel W_i^{t+1} - W^* \parallel_2^2)$

$\leq \eta^2\sigma^2 + (1 - 2\eta\mu)(\eta^2\sigma^2 + (1 - 2\eta\mu) \mathbb{E}(\parallel W_i^{t+1} - W^* \parallel_2^2))$

…

$\leq (1 - 2\mu\eta)^t \mathbb{E}(\parallel W_i^0 - W^* \parallel^2) + \eta\sigma^2/(2\mu)$

As round t increases, the gap gradually decreases to near zero, indicating the model convergence.