# A Convergence Proof Details

## A.1 Proof of Theorem 1: Joint Gradient Decay

PROOF. **Step 1: unbiased gradient decomposition.** Lemma 1 yields $G_{i+1} = G_i - \eta_g(i)\big(\sum_f w_f \widehat{\nabla L_l}(G_i; f) + \Delta_i\big)$ with $\mathbb{E}[\Delta_i] = 0$ and $\mathbb{E}\|\Delta_i\|^2 \le |F|\delta^2$, so the total variance is $\Sigma^2 = \sigma^2 + |F|\delta^2$.
**Step 2: smoothness descent inequality.** Because $\alpha_i \in [0,1]$, the composite loss $J_i = \alpha_i L_g + (1-\alpha_i)\bar{L}_l$ is $\kappa L$–smooth, giving

$$\mathbb{E}[J_{i+1}] \le \mathbb{E}[J_i] - \eta_g(i)\,\mathbb{E}\|\nabla J_i\|^2 + \tfrac{\kappa L}{2}\,\eta_g^2(i)\Sigma^2. \tag{21}$$

**Step 3: telescoping sum.** Summing the above inequality from $i = 0$ to $T-1$ gives

$$\sum_{i=0}^{T-1} \eta_g(i)\,\mathbb{E}\|\nabla J_i\|^2 \le J_0 - J_T + \tfrac{\kappa L \Sigma^2}{2} \sum_{i=0}^{T-1} \eta_g^2(i) = O(1), \tag{22}$$

since $\sum_{i=0}^{T-1} \eta_g^2(i) = O(1)$. Using the step size form $\eta_g(i) \asymp 1/i$, we can see that the constant level on the right-hand side is bounded.
**Step 4: extracting the minimum.** With $\eta_g(i) \asymp 1/i$, Inequality. (17) follows immediately by dividing both sides by $\sum_{i=0}^{T-1} \eta_g(i)$ and taking the minimum. The bounded increment $|\alpha_{i+1} - \alpha_i|$ (Oscillation of $\alpha_i$) only affects the constant $\kappa$ and does not change the $T^{-1/2}$ convergence order. □

## A.2 Proof of Corollary 1: Stability of the Joint Objective

PROOF. **Step 1: one-step descent.** $\kappa L$–smoothness together with the unbiased gradient of Lemma 1 yields

$$J_{i+1} \le J_i - \eta_g(i)\,\|\nabla J_i\|^2 + \frac{\kappa L}{2}\,\eta_g^2(i)\Sigma^2. \tag{23}$$

**Step 2: monotonicity.** Because $\sum_i \eta_g^2(i) < \infty$, there exists $i_0$ such that

$$\frac{\kappa L}{2}\eta_g(i)\Sigma^2 \le \tfrac{1}{2}\|\nabla J_i\|^2, \tag{24}$$

for all $i \ge i_0$. Plugging this into Eq.23 gives $J_{i+1} \le J_i$ for $i \ge i_0$, hence $\{J_i\}$ is eventually non-increasing.
**Step 3: lower boundedness.** Each loss term is non-negative (or contains a non-negative $\ell_2$ regularizer), so $J_i \ge 0$.
**Step 4: convergence and stability.** A monotone, lower-bounded sequence converges; denote its limit by $J_\infty$. Taking limits in Eq.23 gives $\lim_{i\to\infty} |J_{i+1} - J_i| = 0$, i.e. the objective stabilizes. □

## A.3 Proof of Theorem 2: Personalization Error Bound

PROOF. **Step 1: from single-channel to network drift.** Summing Lemma 2 over all BN channels yields

$$\|\text{shift}_{\text{net}}\| \le \frac{\|\Delta\mu_t\|}{\|\sigma\|} + \frac{\|\Delta\sigma_t^2\|}{\|\sigma^2\|} = d_t. \tag{25}$$

**Step 2: from drift to error increment.** Assume the logit mapping is $L_f$–Lipschitz; then $\|\Delta\text{logits}_t\| \le L_f\,d_t$. If the logit margin satisfies $\Pr(M \ge m_0) \ge 1-\rho$, Chebyshev gives $\Delta\text{Err}_t \le (L_f/m_0)\,d_t = c\,d_t$.
**Step 3: weighting by $(1-\alpha_t)$.** Only the personalised part of the loss is affected:

$$\text{Err}_t^{\text{JobFed}} \le \text{Err}_t^{\text{Oracle}} + (1-\alpha_t)c\,d_t. \tag{26}$$

**Step 4: bounding via $a, b$.** Since $1-\alpha_t \le 1-b$ and $\alpha_t \ge a$, $(1-\alpha_t)c\,d_t \le \frac{1-b}{a}\,d_t = \varepsilon_t$. As $\|\Delta\mu_t\|, \|\Delta\sigma_t^2\| \to 0$ during training (Theorem 1), we have $d_t \to 0$ and hence $\varepsilon_t \to 0$. □