A  BI-MAML Algorithm

**Algorithm 2** Biphasic MAML (BI-MAML)

**Input:** Loss functions \( \{f_i(w)\}_{i \in [M]} \), MAML parameter \( \alpha \), step size \( \beta \), tolerance level \( \varepsilon_0, \varepsilon \).

1: **initialize** \( w(0) \in \mathbb{R}^d \) arbitrarily
2: **for** \( t \in \mathbb{N} \cup \{0\} \) **do**
3:  
4:  
5:  
6:  
7:  
8:  
9: **end**

B  Proof of Proposition 3.1

**Proof.** Recall the MAML algorithm with update Eq. (3.1), i.e.,

\[
    w^+ = w - \beta \nabla F(w),
\]

and that \( \nabla F_i(w) = (I_d - \alpha \nabla^2 f_i(w))\nabla f_i(w - \alpha \nabla f_i(w)). \) Expand the terms to get

\[
    \nabla F(w) = E_{i \sim p} [(I_d - \alpha \nabla^2 f_i(w))\nabla f_i(w - \alpha \nabla f_i(w))]
\]

\[
    = E_{i \sim p} \nabla f_i(w) - \alpha \nabla f_i(w) - \alpha E_{i \sim p} \nabla^2 f_i(w) \nabla f_i(w - \alpha \nabla f_i(w))
\]

\[
    = E_{i \sim p} (I_d - \alpha \nabla^2 f_i(w)) \nabla f_i(w) - \alpha E_{i \sim p} \nabla^2 f_i(w)(I_d - \alpha \nabla^2 f_i(w)) \nabla f_i(w)
\]

\[
    = E_{i \sim p} (I_d - \alpha \nabla^2 f_i(w))(I_d - \alpha \nabla^2 f_i(w)) \nabla f_i(w)
\]

\[
    = E_{i \sim p} A_i(w) \nabla f_i(w),
\]

where the first equality follows from definition, the third equality follows from mean value theorem. Here \( \tilde{w}_i \) is a value between \( w \) and \( \nabla f_i(w) \) such that mean value theorem holds. The formula can be further recast into

\[
    \nabla F(w) = E_{i \sim p} \nabla f_i(w) - \alpha (\nabla^2 f_i(w) + \nabla^2 f_i(\tilde{w}_i)) \nabla f_i(w) + \alpha^2 \nabla^2 f_i(w) \nabla^2 f_i(\tilde{w}_i) \nabla f_i(w)
\]

\[
    = \nabla f(w) - E_{i \sim p} [(\alpha (\nabla^2 f_i(w) + \nabla^2 f_i(\tilde{w}_i)) - \alpha^2 \nabla^2 f_i(w) \nabla^2 f_i(\tilde{w}_i)) \nabla f_i(w)].
\]

If we think of the infinitesimal step size \( \beta \to 0 \), we obtain an ODE that represents the gradient flow on \( F(w) \):

\[
    \dot{w} = -\nabla F(w)
\]

\[
    = -\nabla f(w) + E_{i \sim p} [(\alpha (\nabla^2 f_i(w) + \nabla^2 f_i(\tilde{w}_i)) - \alpha^2 \nabla^2 f_i(w) \nabla^2 f_i(\tilde{w}_i)) \nabla f_i(w)]
\]

We define a shorthand \( B_i(w) \) for notational convenience. \( \square \)

C  Proof of the Convergent Upper Bound

**Lemma C.1.** If the loss function \( f_i(w) \) satisfies Assumptions 3.2 and 3.3 and \( \alpha < \frac{1}{2L} \), then it holds that

\[
    \nabla f(w)^T \nabla^2 f(w) E_{i \sim p} [B_i(w) \nabla f_i(w)] \leq \frac{5}{4} L^2 \alpha (L^2 \alpha^2 + 2L^2 \alpha + 2) \| \nabla f(w) \|^2 + \frac{\sigma^2}{2}. \quad (C.1)
\]

**Proof.** Another upper bound for the third term on the right-hand side of Eq. (3.6) can be derived through relaxing its difference with the quadratic form

\[
    \nabla f(w)^T \nabla^2 f(w) E_{i \sim p} [B_i(w) \nabla f_i(w)] - \nabla f(w)^T \nabla^2 f(w) E_{i \sim p} [B_i(w) \nabla f_i(w)]
\]

\[
    = E_{i \sim p} [\nabla f(w)^T \nabla^2 f(w) B_i(w) (\nabla f_i(w) - \nabla f_i(w))]
\]

\[
    \leq \frac{1}{2} E_{i \sim p} \| B_i(w)^T \nabla^2 f(w) \nabla f_i(w) \|^2 + \frac{1}{2} E_{i \sim p} \| \nabla f_i(w) - \nabla f_i(w) \|^2,
\]

\[
    \leq \frac{5}{4} L^2 \alpha (L^2 \alpha^2 + 2L^2 \alpha + 2) \| \nabla f(w) \|^2 + \frac{\sigma^2}{2}.
\]
where the last inequality follows from Young’s inequality. This provides yet another upper bound after rearranging the terms as follows:

\[
\nabla f(w)^T \nabla^2 f(w) E_{i \sim p}[B_i(w) \nabla f_i(w)] \leq \nabla f(w)^T \nabla^2 f(w) E_{i \sim p}[B_i(w) \nabla f(w)] + \frac{L^2}{2} \max_i \|B_i(w)\|^2 \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}
\]

\[
\leq \left( L \max_i \|B_i(w)\| + \frac{L^2}{2} \max_i \|B_i(w)\|^2 \right) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}.
\]

The first and second inequality are due to Assumptions 3.2 and 3.4. Recall that

\[
B_i(w) = \alpha (\nabla^2 f_i(w) + \nabla^2 f_i(\tilde{w}_i)) - \alpha^2 \nabla^2 f_i(w) \nabla^2 f_i(\tilde{w}_i),
\]

and it is not hard to see that \( \max_i \|B_i(w)\| \leq 2\alpha L + \alpha^2 L^2 \). Hence we conclude that

\[
\nabla f(w)^T \nabla^2 f(w) E_{i \sim p}[B_i(w) \nabla f_i(w)] \leq \left( L \max_i \|B_i(w)\| + \frac{L^2}{2} \max_i \|B_i(w)\|^2 \right) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}
\]

\[
\leq \frac{1}{2} \mu L^2 \alpha (L + 2)(L^3 \alpha + 2L^2 \alpha + 2) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}
\]

\[
\leq \frac{5}{4} \mu L^2 \alpha (L^3 \alpha + 2L^2 \alpha + 2) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2},
\]

where the last inequality follows from \( \alpha < \frac{1}{2L} \).

**Proof of Lemma 3.7**

**Proof.** Plug Eq. (C.1) into Eq. (3.6) to get

\[
\frac{d}{dt} \frac{1}{2} \|\nabla f(w)\|^2 \leq -\nabla f(w)^T \nabla^2 f(w) \nabla f(w) + \frac{5}{4} \mu L^2 \alpha (L^3 \alpha + 2L^2 \alpha + 2) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}
\]

\[
\leq -\left( \mu - \frac{5}{4} \mu L^2 \alpha (L^3 \alpha + 2L^2 \alpha + 2) \right) \|\nabla f(w)\|^2 + \frac{\sigma^2}{2}.
\]

**Theorem C.2.** If it holds that

\[
\alpha < \min \left\{ \frac{3}{5} \mu^{1/3} L^{-5/3}, \frac{1}{15} \mu^{1/2} L^{-2}, \frac{1}{15} \mu L^{-2} \right\},
\]

then \( \|\nabla f(w(t))\|^2 \) under (3.2) is upper bounded by a function \( y(t) \) that is exponentially convergent to

\[
\frac{\sigma^2}{2} \mu - \frac{5}{4} \mu L^2 \alpha (L^3 \alpha + 2L^2 \alpha + 2) < \frac{\sigma^2}{\mu}
\]

as \( t \to \infty \).

**Proof.** If \( y(t) \) is the solution of an IVP

\[
\dot{y} \leq -\left( \mu - \frac{5}{4} \mu L^2 \alpha (L^3 \alpha + 2L^2 \alpha + 2) \right) y + \frac{\sigma^2}{2}
\]

with initial condition \( y(0) = \|\nabla f(w(0))\|^2 \), then \( \|\nabla f(w(t))\|^2 \leq y(t) \) for any \( t \geq 0 \). Moreover, it is an ODE of the following form: \( \dot{y} = -\zeta y + \gamma \), which is a simple first-order separable ODE that permits a family of solutions

\[
y(t) = (e^{-\zeta(t+\alpha)} + \gamma) / \zeta
\]
under the condition \( y(0) > \gamma/\zeta \). In our case, \( \zeta = \mu - \frac{5}{4}L^2\alpha(L^3\alpha^2 + 2L^2\alpha + 2) \), \( \gamma = \frac{L^2}{2} \), and the constant \( c_0 \) depends on initial condition \( y(0) \). Consequently, we have \( y \) converges to \( \gamma/\zeta \) exponentially whenever \( \zeta > 0 \). The following theorem provides sufficient conditions for convergence.

We derive sufficient conditions for the quadratic inequality \( \frac{1}{2}\mu - \frac{5}{4}L^2\alpha(L^3\alpha^2 + 2L^2\alpha + 2) > 0 \), i.e.,

\[
\frac{5}{4}L^5\alpha^3 < \frac{\mu}{6}, \quad \frac{5}{2}L^4\alpha^2 < \frac{\mu}{6}, \quad \frac{5}{2}L^2\alpha < \frac{\mu}{6}.
\]

The sufficient conditions reduce to

\[
\alpha < \min \left\{ \frac{\sqrt{2}}{15} 2^{1/3} L^{-5/3}, \frac{1}{15} 2^{1/2} L^{-2}, \frac{1}{15} \mu L^{-2} \right\}
\]

and we have

\[
\frac{\gamma}{\zeta} < \frac{\sigma^2/2}{\mu/2} = \frac{\sigma^2}{\mu}.
\]

\[\Box\]

**Lemma C.3.** Suppose the loss function \( f_i(w) \) satisfies Assumptions 3.2 and 3.4, then for any \( w \in \mathbb{R}^d \) such that \( \|\nabla f(w)\| \leq G \), it holds that \( \|\nabla F(w)\| \leq (1 + 2\alpha L + \alpha^2 L^2)G + (2\alpha L + \alpha^2 L^2)\sigma \).

**Proof.** Recall that \( \nabla F_i(w) = A_i(w)\nabla f_i(w - \alpha \nabla f_i(w)) \). Apply mean value theorem to \( \nabla f_i(w - \alpha \nabla f_i(w)) \) to get

\[
\nabla f_i(w) - \alpha \nabla^2 f_i(\tilde{w}) \nabla f_i(w) = A_i(\tilde{w}) \nabla f_i(w),
\]

where \( \tilde{w} \) lies between \( w \) and \( w - \alpha \nabla f_i(w) \). Consequently, \( \nabla F_i(w) = A_i(w)A_i(\tilde{w}) \nabla f_i(w) \). Further notice that

\[
\|\nabla F(w)\| = \| \mathbb{E}_{i \sim \rho} \nabla F_i(w) \|
\]

\[
= \| \mathbb{E}_{i \sim \rho}[\nabla f_i(w) + (\nabla F_i(w) - \nabla f_i(w))] \|
\]

\[
\leq \| \mathbb{E}_{i \sim \rho} \nabla f_i(w) \| + \| \mathbb{E}_{i \sim \rho}[(I - A_i(w)A_i(\tilde{w}))\nabla f_i(w)] \|
\]

\[
\leq \| \nabla f(w) \| + \mathbb{E}_{i \sim \rho} \|I_d - A_i(w)A_i(\tilde{w})\| \|\nabla f_i(w)\|.
\]

The second equality follows from separating the difference between \( \nabla F(w) \) and \( \nabla f(w) \). The third inequality is due to Eq. (C.2) and triangular inequality. The last inequality is due to Cauchy-Schwarz inequality, and the product of the two norms can be handled separately. Expand \( A_i(w), A_i(\tilde{w}) \) and bound the first term by a constant to get

\[
\|I_d - A_i(w)A_i(\tilde{w})\| = \| \alpha^2 \nabla^2 f_i(w) \nabla^2 f_i(\tilde{w}) - \alpha \nabla^2 f_i(w) - \alpha \nabla^2 f_i(\tilde{w}) \| \leq 2\alpha L + \alpha^2 L^2.
\]

The remaining term can be bounded by variance \( \sigma \) and gradient norm \( \|\nabla f(w)\|\):

\[
\mathbb{E}_{i \sim \rho} \|\nabla f_i(w)\| \leq \| \mathbb{E}_{i \sim \rho} \nabla f_i(w) \| + \mathbb{E}_{i \sim \rho} \| \nabla f_i(w) - \mathbb{E}_{i \sim \rho} \nabla f_i(w) \|
\]

\[
\leq \| \nabla f(w) \| + \sqrt{\mathbb{E}_{i \sim \rho} \| \nabla f_i(w) - \nabla f(w) \|^2}
\]

\[
\leq \| \nabla f(w) \| + \sigma.
\]

The second inequality follows from Jenson inequality. Combining the upper bounds together yields

\[
\|\nabla F(w)\| \leq (1 + 2\alpha L + \alpha^2 L^2)\|\nabla f(w)\| + (2\alpha L + \alpha^2 L^2)\sigma.
\]

\[\Box\]
Proof of Theorem 4.1

Proof. Since the expected loss $f$ is $\mu$-strongly convex, we always have in the first stage that
\[
\frac{d}{dt} \|\nabla f(w(t))\|^2 = \nabla f(w)\nabla^2 f(w) \dot{w} \\
= -\nabla f(w)^T \nabla^2 f(w) \nabla F(w) \\
\leq -\mu \|\nabla f(w)\|^2,
\]
where $\dot{w} = -\nabla f(w)$. It reaches a tolerant level at $\|\nabla f(w)\| \leq \varepsilon_0$, as long as
\[
t \geq \frac{1}{\mu} \log \left( \frac{\|\nabla f(w(0))\|^2}{\varepsilon_0^2} \right) \\
= \frac{2}{\mu} \log \left( \frac{\|\nabla f(w(0))\|^2}{\varepsilon_0^2} \right).
\]

Let us denote
\[
t_1 = \min \left\{ t : \|\nabla f(w(t))\|^2 \leq \varepsilon_0^2 \right\},
\]
By Lemma C.3 and the assumption $\alpha \leq \frac{1}{2L}$ we have
\[
\|\nabla F(w(t_1))\| \leq (1 + 2\alpha L + \alpha^2 L^2)\varepsilon_0 + (2\alpha L + \alpha^2 L^2)\sigma \\
\leq 9 \varepsilon_0 + \frac{5}{4}\sigma.
\]

Let us denote $K = \frac{9}{4}\varepsilon_0 + \frac{5}{4}\sigma$, and Theorem 3.8 implies that if $\alpha \leq \min\left\{ \frac{1}{2L}, \frac{7\mu}{8(16K + 9\sigma)} \right\}$ the MAML loss $F(w)$ is $\frac{\mu}{8}$-strongly convex at $w$, and the MAML ODE (3.2) after time $t_1$ is a gradient flow on a $\frac{\mu}{8}$-strongly convex loss $F(w)$. This dynamics then converges exponentially fast to an approximate stationary point $\tilde{w}$ where $\|\nabla F(\tilde{w})\| \leq \varepsilon$. Similar to the proof of Theorem 3.6, a sufficient condition for the approximate stationary point $\tilde{w}$ writes $e^{-\mu/8\cdot \|\nabla F(w(t_1))\|^2} \leq \varepsilon$, which means $w(\tau + t_1)$ is an approximate stationary point if
\[
\tau \geq \frac{8}{\mu} \log \left( \frac{\|\nabla F(w(t_1))\|^2}{\varepsilon^2} \right) \\
= \frac{16}{\mu} \log \left( \frac{9\varepsilon_0 + 5\sigma}{4\varepsilon} \right).
\]

Combine two parts together to get the major result that the bi-MAML ODE converges to an approximate stationary point $\tilde{w}(t)$ within
\[
t = \frac{1}{\mu} O \left[ \log \left( \frac{(9\varepsilon_0 + 5\sigma)\|\nabla f(w(0))\|}{4\varepsilon_0\varepsilon} \right) \right].
\]

\[
\square
\]

D Proof of Strong Convexity

Lemma D.1. Suppose the loss function $f_i(w)$ satisfies Assumptions 3.2 and 3.4, then for any $w \in \mathbb{R}^d$ such that $\|\nabla F(w)\| \leq K$ and $\alpha < \frac{1}{4L}$, it holds that $\|\nabla f(w)\| \leq \frac{16}{7}K + \frac{9}{7}\sigma$.

Proof. Notice that
\[
\|\nabla f(w)\| = \| E_{i \sim p} f_i(w) \| \\
= \| E_{i \sim p} [\nabla F_i(w) + (\nabla f_i(w) - \nabla F_i(w))] \| \\
\leq \| \nabla F(w) \| + \| E_{i \sim p} (I_d - A_i(w)A_i(\tilde{w})) \nabla f_i(w) \| \\
\leq \| \nabla F(w) \| + E_{i \sim p} \| I_d - A_i(w)A_i(\tilde{w}) \| \| \nabla f_i(w) \| \\
\leq \| \nabla F(w) \| + (2\alpha L + \alpha^2 L^2) E_{i \sim p} \| \nabla f_i(w) \|,
\]

where the first inequality follows from triangular inequality and the third inequality is due to Assumption 3.2. Similarly, we have
\[
\mathbb{E}_{i \sim p} \|\nabla f_i(w)\| \leq \|\nabla f(w)\| + \mathbb{E}_{i \sim p} \|\nabla f_i(w) - \nabla f(w)\|
\]
\[
\leq \|\nabla f(w)\| + \sigma,
\]
where the first inequality is due to triangular inequality and the second one is due to Assumption 3.4. Rearrange the terms under the assumption \(\alpha < \frac{1}{2L}\) to get
\[
\|\nabla f(w)\| \leq \frac{1}{1 - 2\alpha L - \alpha^2 L^2} \|\nabla F(w)\| + \frac{2\alpha L + \alpha^2 L^2}{1 - 2\alpha L - \alpha^2 L^2} \sigma
\]
\[
\leq \frac{16}{l} K + \frac{9}{l} \sigma.
\]

Lemma D.2. Suppose \(f_i(w)\) satisfies Assumptions 3.2, 3.3 and 3.5. For any \(\alpha \leq \min\{\frac{1}{2L}, \frac{\mu}{8\kappa G}\}\) and \(w \in U(G) := \{w \in \mathbb{R}^d : \|\nabla f(w)\| \leq G\}\), we have \(\frac{\mu}{8} I_d \leq \text{Hess}(F(w)) \leq \frac{9L}{8} I_d\).

Proof. Consider \(w, u \in U(G)\), we have
\[
\|\nabla F(w) - \nabla F(u)\| = \|A(w)\nabla f(w - \alpha \nabla f(w)) - A(u)\nabla f(u - \alpha \nabla f(u))\|
\]
\[
\leq \|A(w) - A(u)\| \|\nabla f(w - \alpha \nabla f(w))\| + \|A(u)\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|,
\]
where the inequality follows from triangular inequality. For the first term, we have an upper bound
\[
\|A(w) - A(u)\| \|\nabla f(w - \alpha \nabla f(w))\|
\]
\[
= \alpha \|\nabla^2 f(w - \alpha \nabla f(w))\| \|\nabla f(w - \alpha \nabla f(w))\|
\]
\[
\leq \alpha \kappa \|\nabla w - u\| \|\nabla f(w - \alpha \nabla f(w))\|
\]
\[
= \alpha \kappa \|\nabla w - u\| A(\bar{w}) f(w)
\]
\[
\leq \alpha \kappa \|\nabla w - u\| A(\bar{w}) \|f(w)\|
\]
\[
\leq \alpha (1 - \alpha \mu) \kappa G \|\nabla w - u\|
\]
where the first inequality is due to Cauchy-Schwarz inequality, the second inequality is due to Assumption 3.5, and the second equality follows from mean value theorem, and the last inequality is due to the fact that \(\|A(\bar{w})\| = \|I_d - \alpha \nabla^2 f(\bar{w})\| \leq 1 - \alpha \mu\). Similarly, we bound the second part as
\[
\|A(u)\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
\]
\[
\leq \|A(w)\| \|\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
\]
\[
\leq (1 - \alpha \mu) \|\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
\]
\[
\leq (1 - \alpha \mu)^2 L \|w - u\|,
\]
where the last inequality follows from mean value inequality. Putting the pieces together to get, when \(\alpha \leq \min\{\frac{1}{2L}, \frac{\mu}{8\kappa G}\}\),
\[
\|\nabla F(w) - \nabla F(u)\| \leq \alpha (1 - \alpha \mu) \kappa G \|w - u\| + (1 - \alpha \mu)^2 L \|w - u\|
\]
\[
\leq \alpha \kappa G \|w - u\| + (1 - \alpha \mu)^2 L \|w - u\|
\]
\[
\leq \left(\frac{\mu}{8} + L\right) \|w - u\|
\]
\[
\leq \frac{9L}{8} \|w - u\|
\]
and therefore \(\text{Hess}(F(w)) \leq \frac{9L}{8} I_d\).
The corresponding lower bound similarly follows from triangular inequality where
\[
\|\nabla F(w) - \nabla F(u)\| = \|A(w)\nabla f(w - \alpha \nabla f(w)) - A(u)\nabla f(u - \alpha \nabla f(u))\|
\geq \|A(u)(\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
- \|(A(w) - A(u))\nabla f(w - \alpha \nabla f(w))\|.
\]

When \(\alpha \leq \min\left\{\frac{1}{2L}, \frac{\mu}{8\kappa}\right\}\), the first term is lower bounded as
\[
\|A(u)(\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
\geq (1 - \alpha L)\|\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
\geq (1 - \alpha L)\mu\|(w - u) - (u - \alpha \nabla f(u))\|
\geq (1 - \alpha L)^2\mu\|w - u\|
\geq \frac{\mu}{4}\|w - u\|,
\]

where the first inequality follows from \(\lambda_{\min}(A(u)) \geq 1 - \alpha L\), the second inequality follows from Assumption 3.3, the third inequality is due to triangular inequality, and the last inequality follows from \(\alpha \leq \frac{1}{2L}\). Hence, it holds that
\[
\|\nabla F(w) - \nabla F(u)\| \geq \|A(u)(\nabla f(w - \alpha \nabla f(w)) - \nabla f(u - \alpha \nabla f(u))\|
- \|(A(w) - A(u))\nabla f(w - \alpha \nabla f(w))\|
\geq \frac{\mu}{4}\|w - u\| - \alpha(1 - \alpha \mu)\kappa G\|w - u\|
\geq \left(\frac{\mu}{4} - \frac{\mu}{8}\right)\|w - u\|
= \frac{\mu}{8}\|w - u\|,
\]

where the last inequality follows from \(\alpha \leq \frac{\mu}{8\kappa}\). Thus we obtain \(\text{Hess}(F(w)) \geq \frac{\mu}{8}\). \(\square\)

**Proof of Theorem 3.8**

**Proof.** Combining Lemmas D.1 and D.2 shows that
\[
\frac{\mu}{8}I_d \leq \text{Hess}(F(w)) \leq \frac{9L}{8}I_d,
\]

if \(w \in U(K)\) and
\[
\alpha \leq \min\left\{\frac{1}{2L}, \frac{\mu}{8\kappa}\frac{7}{16K + 9\sigma}\right\}.
\]

\(\square\)

**E Proof of Theorem 3.9**

For \(K > 0\), we define \(U(K) := \{w \in \mathbb{R}^d : \|\nabla F(w)\| \leq K\}\) and \(V(K) := \{w \in \mathbb{R}^d : f(w) - f(x^*) \leq K\}\) where \(x^*\) is the unique global minimizer of \(f\) (recall that \(f\) is \(\mu\)-strongly convex). Let \(\text{Crit}(F)\) denote the set of critical points of \(F\). The convexity of \(f\) implies that \(V(K)\) is convex. All critical points of \(F\) are contained in \(U(K)\) for any \(K > 0\); in other words
\[
\text{Crit}(F) \subseteq U(K), \quad \forall K > 0.
\]

**Lemma E.1.** If the loss function \(f_i(w)\) satisfies Assumptions 3.2 to 3.4, \(\alpha < \frac{1}{4L}\), then we have
\[
U(K) \subseteq V\left(\frac{1}{98\mu}\left(16K + 9\sigma\right)^2\right).
\]
Proof. Let us pick \( w \in \mathbb{R}^d \) such that \( \| \nabla F(w) \| \leq K \). Lemma D.1 implies that there exists a constant \( C_1 = \frac{16\mu}{K} + \frac{\sigma}{2} \) such that \( \| \nabla f(w) \| \leq C_1 \). Since \( f \) is \( \mu \)-strongly convex, we have

\[
f(w) \leq f(x) + \nabla f(x)^T (w - x) + \frac{1}{2\mu} \| \nabla f(w) - \nabla f(x) \|^2, \quad \forall w, x.
\]

Setting \( x \) to the global minimizer \( x^* \) of \( f \) yields

\[
f(w) \leq f(x^*) + \frac{1}{2\mu} \| \nabla f(w) \|^2 \leq f(x^*) + \frac{1}{2\mu} C_1^2 = f(x^*) + \frac{1}{98\mu} (16K + 9\sigma)^2.
\]

Therefore, we have

\[
w \in V \left( \frac{1}{98\mu} (16K + 9\sigma)^2 \right).
\]

\[\square\]

Lemma E.2. Under Assumptions 3.2 to 3.4, we have

\[
V(K) \subseteq U \left( \sigma + \sqrt{2LK} \right).
\]

Proof. Let us rewrite \( \| \nabla F(w) \| \) as below

\[
\| \nabla F(w) \| = \| \mathbb{E}_{i \sim p} (I_d - \alpha \nabla^2 f_i(w)) \nabla f_i(w) - \alpha f_i(w) \|
\]

\[
= \| \mathbb{E}_{i \sim p} (I_d - \alpha \nabla^2 f_i(w)) (I_d - \alpha \nabla^2 f_i(w)) \nabla f_i(w) \|
\]

\[
\leq \mathbb{E}_{i \sim p} \| \nabla f_i(w) \|
\]

\[
\leq (\mathbb{E}_{i \sim p} \| \nabla f_i(w) - f(w) \| + \| \nabla f_i(w) \|)
\]

\[
\leq \sqrt{\mathbb{E}_{i \sim p} \| \nabla f_i(w) - f(w) \|^2 + \| \nabla f_i(w) \|^2}
\]

\[
\leq \sigma + \| \nabla f(w) \|,
\]

where the second inequality is because of the mean value theorem. Since \( f \) is \( L \)-smooth, we have

\[
f(w) \geq f(x) + \nabla f(x)^T (w - x) + \frac{1}{2L} \| \nabla f(w) - \nabla f(x) \|^2, \quad \forall x \in \mathbb{R}^d.
\]

Since \( f \) is \( \mu \)-strongly convex, there exists a unique global minimum \( x^* \) with \( \nabla f(x^*) = 0 \). Therefore, we obtain

\[
f(w) \geq f(x^*) + \frac{1}{2L} \| \nabla f(w) \|^2.
\]

Combining the above inequality and (E.1) yields

\[
\| \nabla F(w) \| \leq \sigma + \sqrt{2L(f(w) - f(x^*))}.
\]

If \( w \in V(K) \), we get

\[
\| F(w) \| \leq \sigma + \sqrt{2LK}.
\]

\[\square\]

Combining Lemmas E.1 and E.2 gives the following corollary.

Corollary E.3. For any \( K > 0 \), if \( \alpha < \frac{1}{14\mu} \), we have the following inclusion relations

\[
\text{Crit}(F) \subseteq U(K) \subseteq V \left( \frac{1}{98\mu} (16K + 9\sigma)^2 \right) \subseteq U \left( \sigma + \sqrt{\frac{L}{\mu} (16K + 9\sigma)} \right).
\]
Corollary E.4. For any $K' \geq \left(\frac{9}{7} \sqrt{\frac{L}{\mu}} + 1 \right) \sigma$, if $\alpha < \frac{1}{2L}$, we have the following inclusion relations

$$\text{Crit}(F) \subseteq U \left( \frac{7K' - \sigma \left(9 \sqrt{\frac{L}{\mu}} + 7\right)}{16 \sqrt{\frac{L}{\mu}}} \right) \subseteq V \left( \frac{(K' - \sigma)^2}{2L} \right) \subseteq U(K')$$

Lemma E.5. Under Assumption 3.3, if $\alpha < \frac{1}{2L}$, we have $\text{Crit}(F)$ is non-empty.

Proof. First we show that $F$ is bounded from below. Since every $f_i$ is strongly convex, it is bounded from below. Recall that $F(w) = \mathbb{E}_{i \sim p} f_i(w - \alpha \nabla f_i(w))$. Therefore $F$ is also bounded from below. Let $F^* := \inf_{w \in \mathbb{R}^d} F(w)$. Pick any $v(0) \in \mathbb{R}^d$ and consider the dynamic defined by

$$\frac{dv(t)}{dt} = -\nabla F(v(t)).$$

Let $E(t) = F(v(t)) - F^*$. We have

$$\frac{dE(t)}{dt} = -\|\nabla F(v(t))\|^2.$$ 

Therefore, we get

$$t \cdot \min_{0 \leq s \leq t} \|\nabla F(v(t))\|^2 \leq \int_0^t \|\nabla F(v(s))\|^2 ds = E(0) - E(t) \leq E(0).$$

Thus we obtain

$$\min_{0 \leq s \leq t} \|\nabla F(v(t))\|^2 \leq \frac{E(0)}{t}. \quad (E.2)$$

Define another function

$$u(t) := v \left( \arg \min_{s \in [0,t]} \|\nabla F(v(t))\|^2 \right),$$

where ties can be broken arbitrarily. Eq. (E.2) implies

$$\|\nabla F(u(t))\| \leq \sqrt{\frac{E(0)}{t}}, \quad \forall t \geq 0.$$

Pick any $K \geq \left(\frac{9}{7} \sqrt{\frac{L}{\mu}} + 1 \right) \sigma$. We have

$$\|\nabla F(u(t))\| \in U(K), \quad \forall t \geq \sqrt{\frac{E(0)}{K}}.$$ 

Since $f$ is strongly convex, $V \left( \frac{(K - \sigma)^2}{2L} \right)$ is convex and non-empty. Thus $U(K)$ is non-empty and closed. Next, we show that $U(K)$ is bounded. Lemma E.1 implies $U(K) \subseteq V \left( \frac{1}{9\beta \mu} (16K + 9\sigma)^2 \right) := V_0$. Since $V_0$ is a sublevel set of $f$ and $f$ is strongly convex, therefore we get the boundedness of $V_0$, which implies the boundedness of $U(K)$. Thus $U(K)$ is compact. Define a sequence $w_n = u \left( n + \sqrt{\frac{E(0)}{K}} \right)$, where $n = 1, 2, 3, \ldots$. We have $w_n \in U(K)$. By Bolzano-Weierstrass theorem, there exists a convergent subsequence $w_{n_i}$. Let $w_0 \in U(K)$ be the limit of $w_{n_i}$. We have

$$\|\nabla F(w_0)\| = \lim_{i \to \infty} \|\nabla F(w_{n_i})\| \leq \lim_{i \to \infty} \sqrt{\frac{E(0)}{n_i + \sqrt{E(0)/K}}} = 0.$$ 

Therefore we conclude that $w_0$ is a critical point of $F$. \qed

Proof of Theorem 3.9. Since $f$ is strongly convex, $V \left( \frac{(K - \sigma)^2}{2L} \right)$ is convex and non-empty. Theorem 3.8 implies that $F$ is $\frac{\mu}{8}$-strongly convex on $U(K)$ and therefore $\frac{\mu}{8}$-strongly convex on its convex subset $V \left( \frac{(K - \sigma)^2}{2L} \right)$ (by Corollary E.4). Since $\text{Crit}(F) \neq \emptyset$ (by Lemma E.5), there is a unique critical point which is the minimizer of $F$ on $V \left( \frac{(K - \sigma)^2}{2L} \right)$. Corollary E.4 implies no critical point outside $V \left( \frac{(K - \sigma)^2}{2L} \right)$. In fact, the unique critical point is the global minimizer of $F$. \qed