# A Proofs for the iSAEM Algorithm

## A.1 Proof of Lemma 2

**Lemma.** *Assume A3,A4. For all* $\mathbf{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1} \langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \tag{16}$$

**Proof** Using A3 and the fact that we can exchange integration with differentiation and the Fisher's identity, we obtain

$$\begin{aligned} \nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \Big( \nabla_{\boldsymbol{\theta}} \, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathsf{L}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big) \\ &= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \Big( \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big) \\ &= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \, \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \left( \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \right) , \end{aligned} \tag{17}$$

Consider the following vector map:

$$\mathbf{s} \to \nabla_{\boldsymbol{\theta}} L(\mathbf{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}(\mathbf{s})} = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{s} \, .$$

Taking the gradient of the above map *w.r.t.* $\mathbf{s}$ and using assumption A3, we show that:

$$\mathbf{0} = -\, \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \Big( \underbrace{\nabla_{\boldsymbol{\theta}}^2 \big( \psi(\boldsymbol{\theta}) + \mathrm{r}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) \,|\, \mathbf{s} \rangle \big)}_{= \mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \boldsymbol{\theta})} |_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}(\mathbf{s})} \Big) \, \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s}) \, .$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{B}(\mathbf{s})(\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))) \, ,$$

where we recall $\mathrm{B}(\mathbf{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big( \mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \overline{\boldsymbol{\theta}}(\mathbf{s})) \Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top$. The proof of (16) follows directly from the assumption A4. □

## A.2 Proof of Theorem 1

Beforehand, We present two intermediary Lemmas important for the analysis of the incremental update of the iSAEM algorithm. The first one gives a characterization of the quantity $\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}]$:

**Lemma.** *Assume A1. The update (1) is equivalent to the following update on the resulting statistics*

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1} \big( \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \big) \, .$$

*Also:*

$$\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E}[\overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \left(1 - \frac{1}{n}\right) \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)} \right] + \frac{1}{n} \mathbb{E}[\eta_{i_k}^{(k+1)}] \, ,$$

*where* $\overline{\mathbf{s}}^{(k)}$ *is defined by* (3) *and* $\tau_i^k = \max\{k' : i_{k'} = i, \; k' < k\}$.

**Proof** From update (1), we have:

$$\begin{aligned} \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= \tilde{S}^{(k)} - \hat{\mathbf{s}}^{(k)} + \frac{1}{n} \left( \tilde{S}_{i_k}^{(k+1)} - \tilde{S}_{i_k}^{(\tau_i^k)} \right) \\ &= \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{S}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n} \left( \tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k+1)} \right) \, . \end{aligned}$$

Since $\tilde{S}_{i_k}^{(k+1)} = \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)}) + \eta_{i_k}^{(k+1)}$ we have

$$\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{S}^{(k)} - \overline{\mathbf{s}}^{(k)} - \frac{1}{n} \left( \tilde{S}_{i_k}^{(\tau_i^k)} - \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)}) \right) + \frac{1}{n} \eta_{i_k}^{(k+1)} \, .$$

Taking the full expectation of both side of the equation leads to:

$$\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E}[\overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right] - \frac{1}{n}\mathbb{E}[\mathbb{E}[\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k]] + \frac{1}{n}\mathbb{E}[\eta_{i_k}^{(k+1)}] \;.$$

Since we have $\mathbb{E}[\tilde{S}_i^{(\tau_i^k)}|\mathcal{F}_k] = \frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)}$ and $\mathbb{E}\left[\overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k\right] = \overline{\mathbf{s}}^{(k)}$, we conclude the proof of the Lemma. □

We also derived the following auxiliary Lemma which sets an upper bound for the quantity $\mathbb{E}[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2]$:

**Lemma 7.** *For any $k \geq 0$ and consider the iSAEM update in (1), it holds that*

$$\mathbb{E}[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \leq 4\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{2\,\mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^n \mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2\right] + 2\frac{c_\eta}{M_k} + 4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \;.$$

**Proof** Applying the iSAEM update yields:

$$\begin{aligned}
\mathbb{E}[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] =& \mathbb{E}[\|\tilde{S}^{(k)} - \hat{\boldsymbol{s}}^{(k)} - \frac{1}{n}(\tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k)})\|^2] \\
\leq& 4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] + 4\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\
&+ \frac{2}{n^2}\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] + 2\frac{c_\eta}{M_k} \;.
\end{aligned}$$

The last expectation can be further bounded by

$$\frac{2}{n^2}\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{2\,\mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \;,$$

where (a) is due to Lemma 1 and which concludes the proof of the Lemma.

□

**Theorem.** *Assume A1-A5. Consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$ for any $k \leq \mathsf{K_m}$ where $\mathsf{K_m}$ is a positive integer. Let $\{\gamma_k = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of stepsizes, $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1+6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = c_1\overline{L}/n$. Then:*

$$\upsilon_{\max}^{-2}\sum_{k=0}^{\mathsf{K_m}} \tilde{\alpha}_k \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K_m})})] + \sum_{k=0}^{\mathsf{K_m}-1} \tilde{\Gamma}_k \mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \;.$$

**Proof** Under the smoothness of the Lyapunov function $V$ (cf. Lemma 1), we can write:

$$V(\hat{\boldsymbol{s}}^{(k+1)}) \leq V(\hat{\boldsymbol{s}}^{(k)}) + \gamma_{k+1}\langle \tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \;.$$

Taking the expectation on both sides yields:

$$\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k+1)})\right] \leq \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)})\right] + \gamma_{k+1}\mathbb{E}\left[\langle \tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \;.$$

Using Lemma 3, we obtain:

$$
\begin{aligned}
&\mathbb{E}\left[\langle \tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] \\
=&\mathbb{E}\left[\langle \bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] + \left(1-\frac{1}{n}\right)\mathbb{E}\left[\langle \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] \\
&+ \frac{1}{n}\mathbb{E}\left[\langle \eta_{i_k}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] \\
\overset{(a)}{\leq} & -\upsilon_{\min}\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \left(1-\frac{1}{n}\right)\mathbb{E}\left[\langle \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] \\
&+ \frac{1}{n}\mathbb{E}\left[\langle \eta_{i_k}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] \\
\overset{(b)}{\leq} & -\upsilon_{\min}\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{1-\frac{1}{n}}{2\beta}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \frac{\beta(n-1)+1}{2n}\mathbb{E}\left[\left\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\right\|^2\right] + \frac{1}{2n}\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
\overset{(a)}{\leq} & \left(\upsilon_{\max}^2\frac{\beta(n-1)+1}{2n} - \upsilon_{\min}\right)\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{1-\frac{1}{n}}{2\beta}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \frac{1}{2n}\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \,,
\end{aligned}
$$

where (a) is due to the growth condition (2) and (b) is due to Young's inequality (with $\beta \to 1$). Note $a_k = \gamma_{k+1}\left(\upsilon_{\min} - \upsilon_{\max}^2\frac{\beta(n-1)+1}{2n}\right)$ and

$$
\begin{aligned}
a_k\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \leq &\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \\
&+ \frac{\gamma_{k+1}(1-\frac{1}{n})}{2\beta}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{\gamma_{k+1}}{2n}\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \,.
\end{aligned}
\tag{18}
$$

We now give an upper bound of $\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right]$ using Lemma 7 and plug it into (18):

$$
\begin{aligned}
&(a_k - 2\gamma_{k+1}^2\,\mathrm{L}_V)\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\
\leq &\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] \\
&+ \gamma_{k+1}\left(\frac{1}{2\beta}(1-\frac{1}{n}) + 2\gamma_{k+1}\,\mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \gamma_{k+1}\left(\gamma_{k+1}\,\mathrm{L}_V + \frac{1}{2n}\right)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
&+ \frac{\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2] \,.
\end{aligned}
\tag{19}
$$

Next, we observe that

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2]\right) \,,
$$

where the equality holds as $i_k$ and $j_k$ are drawn independently. For any $\beta > 0$, it holds

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\big\rangle\Big] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 - 2\gamma_{k+1}\big\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\big\rangle\Big] \\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 + \gamma_{k+1}\beta\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2\Big] ,
\end{aligned}
$$

where the last inequality is due to Young's inequality. Subsequently, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\tau_i^{k+1})}\|^2] \\
\leq&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n^2}\sum_{i=1}^n \mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2\Big] .
\end{aligned}
$$

Observe that $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)})$. Applying Lemma 7 yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\tau_i^{k+1})}\|^2] \\
\leq&\big(\gamma_{k+1}^2 + \frac{n-1}{n}\frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\Big[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\Big] + \sum_{i=1}^n \mathbb{E}\Big[\frac{1-\frac{1}{n}+\gamma_{k+1}\beta}{n}\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2\Big] \\
\leq&4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\Big[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2\Big] + 2\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
&+4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+\sum_{i=1}^n \mathbb{E}\Big[\frac{1-\frac{1}{n}+\gamma_{k+1}\beta+\frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})}{n}\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2\Big] .
\end{aligned}
$$

Let us define

$$
\Delta^{(k)} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2] .
$$

From the above, we get

$$
\begin{aligned}
\Delta^{(k+1)} \leq& \big(1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})\big)\Delta^{(k)} + 4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\Big[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2\Big] \\
&+ 2\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] + 4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] .
\end{aligned}
$$

Setting $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1+6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \overline{L}}$, $\beta = \frac{c_1\overline{L}}{n}$, $c_1(k\alpha - 1) \geq c_1(\alpha-1) \geq 6$, $\alpha \geq 8$, we observe that

$$
1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta}) \leq 1 - \frac{c_1(k\alpha-1)-4}{k\alpha n c_1} \leq 1 - \frac{2}{k\alpha n c_1} ,
$$

which shows that $1-\frac{1}{n}+\gamma_{k+1}\beta+\frac{2\gamma_{k+1}\,\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})\in(0,1)$ for any $k>0$. Denote $\Lambda_{(k+1)}=\frac{1}{n}-\gamma_{k+1}\beta-\frac{2\gamma_{k+1}\,\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1}+\frac{1}{\beta})$ and note that $\Delta^{(0)}=0$, thus the telescoping sum yields:

$$
\begin{aligned}
\Delta^{(k+1)} \leq\ & 4\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}\left(1-\Lambda_{(j)}\right)\left(\gamma_{\ell+1}^2+\frac{\gamma_{\ell+1}}{\beta}\right)\mathbb{E}[\|\overline{\boldsymbol{s}}^{(\ell)}-\hat{\boldsymbol{s}}^{(\ell)}\|^2] \\
&+2\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}\left(1-\Lambda_{(j)}\right)\left(\gamma_{\ell+1}^2+\frac{\gamma_{\ell+1}}{\beta}\right)\mathbb{E}\left[\left\|\eta_{i_\ell}^{(\ell)}\right\|^2\right] \\
&+4\sum_{\ell=0}^{k}\prod_{j=\ell+1}^{k}\left(1-\Lambda_{(j)}\right)\left(\gamma_{\ell+1}^2\right. \\
&\left.+\frac{\gamma_{\ell+1}}{\beta}\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^\ell)}-\overline{\mathbf{s}}^{(\ell)}\right\|^2\right]\ .
\end{aligned}
$$

Note $\omega_{k,\ell}=\prod_{j=\ell+1}^{k}\left(1-\Lambda_{(j)}\right)$ Summing on both sides over $k=0$ to $k=\mathsf{K}_{\mathsf{m}}-1$ yields:

$$
\begin{aligned}
&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\Delta^{(k+1)} \\
=&4\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)\omega_{k,1}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2]+2\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)\omega_{k,1}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
+&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}4\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)\omega_{k,1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)}-\overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
\leq&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\frac{4\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2]+\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\frac{2\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
+&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\frac{4\left(\gamma_{k+1}^2+\frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)}-\overline{\mathbf{s}}^{(k)}\right\|^2\right]\ .
\end{aligned}
\tag{20}
$$

We recall (19) where we have summed on both sides from $k=0$ to $k=\mathsf{K}_{\mathsf{m}}-1$:

$$
\begin{aligned}
&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\left(a_k-2\gamma_{k+1}^2\,\mathrm{L}_V\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)}-\hat{\boldsymbol{s}}^{(k)}\|^2] \\
\leq&\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(0)})-V(\hat{\boldsymbol{s}}^{(K)})\right] \\
+&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\gamma_{k+1}\left(\frac{1}{2\beta}(1-\frac{1}{n})+2\gamma_{k+1}\,\mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)}-\overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
+&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\gamma_{k+1}\left(\gamma_{k+1}\,\mathrm{L}_V+\frac{1}{2n}\right)\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\
+&\sum_{k=0}^{\mathsf{K}_{\mathsf{m}}-1}\frac{\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}{n^2}\Delta^{(k)}\ .
\end{aligned}
\tag{21}
$$

Plugging (20) into (21) results in:

$$\sum_{k=0}^{\mathsf{K_m}-1} \tilde{\alpha}_k \mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \sum_{k=0}^{\mathsf{K_m}-1} \tilde{\beta}_k \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right]$$
$$\leq \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K)})\right] + \sum_{k=0}^{\mathsf{K_m}-1} \tilde{\Gamma}_k \mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \;,$$

where

$$\tilde{\alpha}_k = a_k - 2\gamma_{k+1}^2 \,\mathrm{L}_V - \frac{\gamma_{k+1}^2 \,\mathrm{L}_V \,\mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \;,$$
$$\tilde{\beta}_k = \gamma_{k+1}\left(\frac{1}{2\beta}(1-\frac{1}{n}) + 2\gamma_{k+1}\,\mathrm{L}_V\right) - \frac{\gamma_{k+1}^2 \,\mathrm{L}_V \,\mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{4\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \;,$$
$$\tilde{\Gamma}_k = \gamma_{k+1}\left(\gamma_{k+1}\,\mathrm{L}_V + \frac{1}{2n}\right) + \frac{\gamma_{k+1}^2 \,\mathrm{L}_V \,\mathrm{L}_{\mathbf{s}}^2}{n^2} \frac{2\big(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\big)}{\Lambda_{(k+1)}} \;,$$

and

$$a_k = \gamma_{k+1}\left(\upsilon_{\min} - \upsilon_{\max}^2 \frac{\beta(n-1)+1}{2n}\right) \;,$$
$$\Lambda_{(k+1)} = \frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1}\,\mathrm{L}_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \;,$$
$$c_1 = \upsilon_{\min}^{-1}, \alpha = \max\{8, 1+6\upsilon_{\min}\}, \overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}, \gamma_{k+1} = \frac{1}{k\alpha c_1 \overline{L}}, \beta = \frac{c_1 \overline{L}}{n} \;.$$

When, for any $k > 0$, $\tilde{\alpha}_k \geq 0$, we have by Lemma 2 that:

$$\sum_{k=0}^{\mathsf{K_m}} \tilde{\alpha}_k \mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \upsilon_{\max}^2 \sum_{k=0}^{\mathsf{K_m}} \tilde{\alpha}_k \mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \;,$$

which yields an upper bound of the gradient of the Lyapunov function $V$ along the path of the iSAEM update and concludes the proof of the Theorem. □

## B Proofs for the vrTTEM and the fiTTEM Algorithms

### B.1 Proofs of Auxiliary Lemmas ( Lemma 4, Lemma 5 and Lemma 6)

**Lemma.** *Consider the vrTTEM update (2) with $\rho_k = \rho$, it holds for all $k > 0$*

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq & 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2 \mathrm{L}_{\mathbf{s}}^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] \\ & + 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - \tilde{S}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] ,\end{aligned}$$

*where we recall that $\ell(k)$ is the first iteration number in the epoch that iteration $k$ is in.*

**Proof** Beforehand, we provide a rewiriting of the quantity $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}$ that will be useful throughout this proof:

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) \\ &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\boldsymbol{\mathcal{S}}^{(k+1)}) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right) .\end{aligned} \tag{22}$$

We observe, using the identity (22), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\boldsymbol{\mathcal{S}}^{(k+1)}\|^2]+2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))}-\tilde{S}^{(k)}\|^2]. \tag{23}$$

For the latter term, we obtain its upper bound as

$$\begin{aligned}&\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \\ =&\mathbb{E}\Big[\|\frac{1}{n}\sum_{i=1}^n (\overline{\boldsymbol{s}}_i^{(k)} - \tilde{S}_i^{\ell(k)}) - (\overline{\boldsymbol{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))})\|^2\Big] \\ \overset{(a)}{\leq}&\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \overset{(b)}{\leq} \mathrm{L}_{\mathbf{s}}^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] ,\end{aligned}$$

where $(a)$ uses the variance inequality and $(b)$ uses Lemma 1. Substituting into (23) proves the lemma. □

**Lemma.** *Consider the fiTTEM update (3) with $\rho_k = \rho$. It holds for all $k > 0$ that*

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq & 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2 \frac{\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\ & + 2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - \tilde{S}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] ,\end{aligned}$$

*where $\mathrm{L}_{\mathbf{s}}$ is the smoothness constant defined in Lemma 1.*

**Proof** Beforehand, we provide a rewiriting of the quantity $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}$ that will be useful throughout this proof:

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) \\ &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\boldsymbol{\mathcal{S}}^{(k+1)}) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)} - (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})\right]\right) .\end{aligned} \tag{24}$$

We observe, using the identity (24), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+2\rho^2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)}-\boldsymbol{\mathcal{S}}^{(k+1)}\|^2]+2(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))}-\tilde{S}^{(k)}\|^2] . \tag{25}$$

For the latter term, we obtain its upper bound as

$$\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] = \mathbb{E}\Big[\|\frac{1}{n}\sum_{i=1}^{n}\big(\overline{\mathbf{s}}_i^{(k)} - \overline{\boldsymbol{\mathcal{S}}}_i^{(k)}\big) - \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)\|^2\Big]$$
$$\stackrel{(a)}{\leq} \mathbb{E}[\|\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \;,$$

where $(a)$ uses the variance inequality. We can further bound the last expectation using Lemma 1:

$$\mathbb{E}[\|\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\overline{\mathbf{s}}_i^{(k)} - \overline{\mathbf{s}}_i^{(t_i^k)}\|^2] \stackrel{(a)}{\leq} \frac{\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \;.$$

Substituting into (25) proves the lemma. □

**Lemma.** *Considering a decreasing stepsize $\gamma_k \in (0,1)$ and a constant $\rho \in (0,1)$, we have*

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \leq \frac{\rho}{1-\rho}\sum_{\ell=0}^{k}(1-\gamma_\ell)^2(\boldsymbol{\mathcal{S}}^{(\ell)} - \tilde{S}^{(\ell)}) \;,$$

*where $\boldsymbol{\mathcal{S}}^{(k)}$ is defined either by Line 2 (vrTTEM ) or Line 3 (fiTTEM ).*

**Proof** We begin by writing the two-timescale update:

$$\begin{aligned} \tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big) \;, \\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \;, \end{aligned} \tag{26}$$

where $\boldsymbol{\mathcal{S}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(t_i^k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)$ according to (3). Denote $\delta^{(k+1)} = \hat{\mathbf{s}}^{(k+1)} - \tilde{S}^{(k+1)}$. Then from (26), doing the subtraction of both equations yields:

$$\delta^{(k+1)} = (1-\gamma_{k+1})\delta^{(k)} + \frac{\rho}{1-\rho}(1-\gamma_{k+1})(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k+1)}) \;.$$

Using the telescoping sum and noting that $\delta^{(0)} = 0$, we have

$$\delta^{(k+1)} \leq \frac{\rho}{1-\rho}\sum_{\ell=0}^{k}(1-\gamma_{\ell+1})^2(\boldsymbol{\mathcal{S}}^{(\ell+1)} - \tilde{S}^{(\ell+1)}) \;.$$

□

### B.2 Additional Intermediary Result

**Lemma 8.** *At iteration $k+1$,the drift term of update (3), with $\rho_{k+1} = \rho$, is equivalent to the following :*

$$\begin{aligned} \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} =& \rho(\hat{\mathbf{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}) + \rho\eta_{i_k}^{(k+1)} + \rho\left[\big(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]\right] \\ &+ (1-\rho)\left(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\right) \;, \end{aligned}$$

*where we recall that $\eta_{i_k}^{(k+1)}$, defined in (12), which is the gap between the MC approximation and the expected statistics.*

**Proof** Using the fiTTEM update $\tilde{S}^{(k+1)} = (1-\rho)\tilde{S}^{(k)} + \rho \boldsymbol{\mathcal{S}}^{(k+1)}$ where $\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})$ leads to the following decomposition:

$$
\begin{aligned}
&\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \\
=&(1-\rho)\tilde{S}^{(k)} + \rho\left(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})\right) - \hat{\boldsymbol{s}}^{(k)} + \rho\overline{\mathbf{s}}^{(k)} - \rho\overline{\mathbf{s}}^{(k)} \\
=&\rho(\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}) + \rho(\tilde{S}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(k)}) + (1-\rho)\left(\tilde{S}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right) + \rho\left(\overline{\boldsymbol{\mathcal{S}}}^{(k)} - \overline{\mathbf{s}}^{(k)} + (\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)})\right) \\
=&\rho(\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}) + \rho\eta_{i_k}^{(k+1)} - \rho\left[(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]\right] \\
&+(1-\rho)\left(\tilde{S}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right) \ ,
\end{aligned}
$$

where we observe that $\mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}] = \overline{\mathbf{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}$ and which concludes the proof.

*Important Note:* Note that $\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not equal to $\eta_{i_k}^{(k+1)}$, defined in (12), which is the gap between the MC approximation and the expected statistics. Indeed $\tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not computed under the same model as $\overline{\mathbf{s}}_{i_k}^{(k)}$. □

### B.3 Proof of Theorem 2

**Theorem.** *Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K}_\mathsf{m}$ where $\mathsf{K}_\mathsf{m}$ is a positive integer. Let $\{\gamma_{k+1} = 1/(k^a\overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of stepsizes, $\overline{L} = \max\{\mathrm{L}_\mathbf{s}, \mathrm{L}_V\}$, $\rho = \mu/(c_1\overline{L}n^{2/3})$, $m = nc_1^2/(2\mu^2 + \mu c_1^2)$ and a constant $\mu \in (0,1)$. Then:*

$$
\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] \leq \frac{2n^{2/3}\overline{L}}{\mu \mathsf{P}_\mathsf{m} \upsilon_{\min}^2 \upsilon_{\max}^2}\left(\mathbb{E}[\Delta V] + \sum_{k=0}^{\mathsf{K}_\mathsf{m}-1} \tilde{\eta}^{(k+1)} + \chi^{(k+1)}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\right) \ .
$$

**Proof** Using the smoothness of $V$ and update (2), we obtain:

$$
\begin{aligned}
V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) + \left\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle + \frac{\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \\
&\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1}\left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \ .
\end{aligned} \tag{27}
$$

Denote $\mathsf{H}_{k+1} := \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}$ the drift term of the fiTTEM update in (7) and $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. Taking expectations on both sides show that

$$
\begin{aligned}
&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \\
\overset{(a)}{\leq}&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] - \gamma_{k+1}\rho\mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] \\
&+\frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
\overset{(b)}{\leq}&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\rho\mathbb{E}\left[\left\langle \mathsf{h}_k \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] \\
&-\gamma_{k+1}\rho\mathbb{E}\left[\left\langle \eta_{i_k}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\right\rangle\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
\overset{(c)}{\leq}&\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \left(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2\right)\mathbb{E}\left[\|\mathsf{h}_k\|^2\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\
&-\gamma_{k+1}\rho\mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \ ,
\end{aligned} \tag{28}
$$

where we have used (22) in $(a)$ and $\mathbb{E}\left[\boldsymbol{\mathcal{S}}^{(k+1)}\right] = \overline{\mathbf{s}}^{(k)} + \mathbb{E}[\eta_{i_k}^{(k+1)}]$ in $(b)$, the growth condition in Lemma 2 and Young's inequality with the constant equal to $1$ in $(c)$.

Furthermore, for $k+1 \leq \ell(k)+m$ (*i.e.*, $k+1$ is in the same epoch as $k$), we have

$$
\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] = \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} + \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))} \,|\, \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\big\rangle\Big] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 \\
&-2\gamma_{k+1}\big\langle \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))} \,|\, \rho(\mathsf{h}_k - \eta_{i_k}^{(k+1)}) + (1-\rho)(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)})\big\rangle\Big] \\
\leq&\mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta}\|\mathsf{h}_k\|^2 \\
&+\frac{\gamma_{k+1}\rho}{\beta}\|\eta_{i_k}^{(k+1)}\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\Big] \,,
\end{aligned}
$$

where we first used (22) and the last inequality is due to Young's inequality.

Consider the following sequence

$$
R_k := \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)}) + b_k\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2] \,,
$$

where $b_k := \overline{b}_{k \bmod m}$ is a periodic sequence where:

$$
\overline{b}_i = \overline{b}_{i+1}(1+\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2) + \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2, \;\; i = 0,1,\ldots,m-1 \;\text{ with }\; \overline{b}_m = 0 \,.
$$

Note that $\overline{b}_i$ is decreasing with $i$ and this implies

$$
\overline{b}_i \leq \overline{b}_0 = \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\,\frac{(1+\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2)^m - 1}{\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2}, \;\; i = 1,2,\ldots,m \,.
$$

For $k+1 \leq \ell(k)+m$, we have the following inequality

$$
\begin{aligned}
R_{k+1} \leq\, &\mathbb{E}\Big[V(\hat{\boldsymbol{s}}^{(k)}) - \big(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2\big)\,\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\mathsf{H}_{k+1}\|^2\Big] \\
&+ \gamma_{k+1}\mathbb{E}\left[\rho\left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \\
&+ b_{k+1}\mathbb{E}\left[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2\|\mathsf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta}\|\mathsf{h}_k\|^2\right] \\
&+ b_{k+1}\mathbb{E}\left[\frac{\gamma_{k+1}\rho}{\beta}\|\eta_{i_k}^{(k+1)}\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \,.
\end{aligned}
$$

And using Lemma 4 we obtain:

$$
\begin{aligned}
R_{k+1} \leq\, &\mathbb{E}\Big[V(\hat{\boldsymbol{s}}^{(k)}) - \big(\gamma_{k+1}\rho\upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2 - \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\big)\,\|\mathsf{h}_k\|^2 + \gamma_{k+1}^2\rho^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\,\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big] \\
&+ b_{k+1}\mathbb{E}\left[(1+\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2 + (\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\|\mathsf{h}_k\|^2\right] \\
&+ \gamma_{k+1}\mathbb{E}\left[(\rho + \rho^2\gamma_{k+1}\,\mathrm{L}_V)\left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho-(1-\rho)^2\gamma_{k+1}\,\mathrm{L}_V)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \\
&+ b_{k+1}\mathbb{E}\left[(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\|\eta_{i_k}^{(k+1)}\|^2 + (\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2)\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \,.
\end{aligned}
$$

Rearranging the terms yields:

$$
\begin{aligned}
R_{k+1} \leq\, &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\big(\rho\upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)\big)\mathbb{E}[\|\mathsf{h}_k\|^2] \\
&+ \Big(\underbrace{b_{k+1}(1+\gamma\beta + 2\gamma^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2) + \gamma^2\rho^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}_{=b_k \;\;\text{since } k+1 \leq \ell(k)+m}\Big)\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big] + \tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)} \,,
\end{aligned}
$$

where

$$\tilde{\eta}^{(k+1)} = \left(\gamma_{k+1}(\rho + \rho^2\gamma_{k+1}\,\mathrm{L}_V) + b_{k+1}(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\right)\mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right]$$
$$\chi^{(k+1)} = \left(b_{k+1}(\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2) - \gamma_{k+1}(1-\rho-(1-\rho)^2\gamma_{k+1}\,\mathrm{L}_V)\right)$$
$$\tilde{\chi}^{(k+1)} = \chi^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \ .$$

This leads, using Lemma 2, that for any $\gamma_{k+1}$, $\rho$ and $\beta$ such that $\rho\upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2) > 0$,

$$\begin{aligned}
&\upsilon_{\max}^2\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\
\leq &\frac{R_k - R_{k+1}}{\gamma_{k+1}\big(\rho\upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)\big)} \\
&+ \frac{\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}}{\gamma_{k+1}\big(\rho\upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)\big)} \ .
\end{aligned}$$

We first remark that

$$\begin{aligned}
&\gamma_{k+1}\big(\rho\upsilon_{\min} + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2\,\mathrm{L}_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)\big) \\
&\geq \frac{\gamma_{k+1}\rho}{c_1}\big(1 - \gamma_{k+1}c_1\rho\,\mathrm{L}_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1)\big) \ ,
\end{aligned}$$

where $c_1 = \upsilon_{\min}^{-1}$. By setting $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = \frac{c_1\overline{L}}{n^{1/3}}$, $\rho = \frac{\mu}{c_1\overline{L}n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2+\mu c_1^2}$ and $\{\gamma_{k+1}\}$ any sequence of decreasing stepsizes in $(0,1)$, it can be shown that there exists $\mu \in (0,1)$, such that the following lower bound holds

$$\begin{aligned}
&1 - \gamma_{k+1}c_1\rho\,\mathrm{L}_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1) \\
\geq &1 - \frac{\mu}{n^{\frac{2}{3}}} - \bar{b}_0\big(\frac{n^{\frac{1}{3}}}{\overline{L}} + \frac{2\mu}{\overline{L}n^{\frac{2}{3}}}\big) \\
\geq &1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mathrm{L}_V\,\mu^2}{c_1^2 n^{\frac{4}{3}}}\frac{(1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m - 1}{\gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}\big(\frac{n^{\frac{1}{3}}}{\overline{L}} + \frac{2\mu}{\overline{L}n^{\frac{2}{3}}}\big) \\
\overset{(a)}{\geq} &1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mu}{c_1^2}(\mathrm{e}-1)\big(1 + \frac{2\mu}{n}\big) \geq 1 - \mu - \mu(1+2\mu)\frac{\mathrm{e}-1}{c_1^2} \overset{(b)}{\geq} \frac{1}{2} \ ,
\end{aligned}$$

where the simplification in (a) is due to

$$\frac{\mu}{n} \leq \gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2 \leq \frac{\mu}{n} + \frac{2\mu^2}{c_1^2 n^{\frac{4}{3}}} \leq \frac{\mu c_1^2 + 2\mu^2}{c_1^2}\frac{1}{n} \ \text{ and } \ (1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m \leq \mathrm{e} - 1.$$

and the required $\mu$ in (b) can be found by solving the quadratic equation.

Finally, these results yield:

$$\upsilon_{\max}^2\sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \frac{2(R_0 - R_{\mathsf{K}_\mathsf{m}})}{\upsilon_{\min}\rho} + 2\sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\frac{\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}}{\upsilon_{\min}\rho} \ .$$

Note that $R_0 = \mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)})]$ and if $\mathsf{K}_\mathsf{m}$ is a multiple of $m$, then $R_{\mathsf{max}} = \mathbb{E}[V(\hat{\boldsymbol{s}}^{(\mathsf{K}_\mathsf{m})})]$. Under the latter condition, we have

$$\sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \frac{2n^{2/3}\overline{L}}{\mu\upsilon_{\min}^2\upsilon_{\max}^2}\mathbb{E}[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K}_\mathsf{m})})] + \frac{2n^{2/3}\overline{L}}{\mu\upsilon_{\min}^2\upsilon_{\max}^2}\sum_{k=0}^{\mathsf{K}_\mathsf{m}-1}\left[\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}\right] \ .$$

This concludes our proof.

□

### B.4 Proof of Theorem 3

**Theorem.** *Assume A1-A5. Consider the fiTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K_m}$ where $\mathsf{K_m}$ be a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of positive stepsizes, $\alpha = \max\{2, 1+2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = 1/(\alpha n)$, $\rho = 1/(\alpha c_1 \overline{L} n^{2/3})$ and $c_1(k\alpha - 1) \geq c_1(\alpha-1) \geq 2$, $\alpha \geq 2$. Then:*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2] \leq \frac{4\alpha \overline{L} n^{2/3}}{\mathsf{P_m} \upsilon_{\min}^2 \upsilon_{\max}^2} \left( \mathbb{E}\big[\Delta V\big] + \sum_{k=0}^{\mathsf{K_m}-1} \Xi^{(k+1)} + \Gamma^{(k+1)} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \right) .$$

**Proof** Using the smoothness of $V$ and update (3), we obtain:

$$\begin{aligned} V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) + \big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \big\rangle + \frac{\mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \\ &\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1} \big\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \big\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 . \end{aligned} \tag{29}$$

Denote $\mathsf{H}_{k+1} := \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}$ the drift term of the fiTTEM update in (7) and $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. Using Lemma 8 and the additional following identity:

$$\mathbb{E}\left[ \big(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}] \right] = 0 \,, \tag{30}$$

we have:

$$\begin{aligned} &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \\ \leq &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\rho \mathbb{E}[\langle \mathsf{h}_k \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle - \gamma_{k+1} \mathbb{E}\left[ \big\langle \rho \mathbb{E}[\eta_{i_k}^{(k+1)} | \mathcal{F}_k] + (1-\rho)\mathbb{E}[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}] \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \big\rangle \right] \\ &+ \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\mathsf{H}_{k+1}\|^2 \\ \overset{(a)}{\leq} &- \upsilon_{\min} \gamma_{k+1}\rho \mathbb{E}[\|\mathsf{h}_k\|^2] - \gamma_{k+1} \mathbb{E}\left[ \left\| \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\|^2 \right] - \frac{\gamma_{k+1}\rho^2}{2} \xi^{(k+1)} - \frac{\gamma_{k+1}(1-\rho)^2}{2} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &+ \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\mathsf{H}_{k+1}\|^2 \\ \overset{(b)}{\leq} &- (\upsilon_{\min}\gamma_{k+1}\rho + \gamma_{k+1}\upsilon_{\max}^2) \mathbb{E}[\|\mathsf{h}_k\|^2] - \frac{\gamma_{k+1}\rho^2}{2} \xi^{(k+1)} - \frac{\gamma_{k+1}(1-\rho)^2}{2} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &+ \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\mathsf{H}_{k+1}\|^2 \,, \end{aligned}$$

where $\xi^{(k+1)} = \mathbb{E}[\|\mathbb{E}[\eta_{i_k}^{(k+1)} | \mathcal{F}_k]\|^2]$.

**Bounding** $\mathbb{E}\left[\|\mathsf{H}_{k+1}\|^2\right]$ Using Lemma 5, we obtain:

$$\begin{aligned} &\gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \mathrm{L}_V) \mathbb{E}[\|\mathsf{h}_k\|^2] \\ \leq &\mathbb{E}\left[ V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)}) \right] + \tilde{\xi}^{(k+1)} + \left( (1-\rho)^2 \gamma_{k+1}^2 \mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &+ \frac{\gamma_{k+1}^2 \mathrm{L}_V \rho^2 \mathrm{L}_{\mathbf{s}}^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \,, \end{aligned} \tag{31}$$

where $\tilde{\xi}^{(k+1)} = \gamma_{k+1}^2 \rho^2 \mathrm{L}_V \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] - \frac{\gamma_{k+1}\rho^2}{2} \xi^{(k+1)}$. Next, we observe that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{n} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \right) , \tag{32}$$

where the equality holds as $i_k$ and $j_k$ are drawn independently. Then,

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
&= \mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\big\rangle\Big] .
\end{aligned}$$

Note that $\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) = -\gamma_{k+1}\mathsf{H}_{k+1}$ and that in expectation we recall that $\mathbb{E}[\mathsf{H}_{k+1}|\mathcal{F}_k] = \rho\mathsf{h}_k + \rho\mathbb{E}[\eta_{i_k}^{(k+1)}|\mathcal{F}_k] + (1-\rho)\mathbb{E}[\tilde{S}^{(k)} - \hat{\boldsymbol{s}}^{(k)}]$ where $\mathsf{h}_k = \hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}$. Thus, for any $\beta > 0$, it holds

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\big\rangle\Big] \\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + (1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] \\
&+ \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\Big] ,
\end{aligned}$$

where the last inequality is due to Young's inequality. Plugging this into (32) yields:

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
=&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\big\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\big\rangle\Big] \\
\leq&\mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + (1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] \\
&+ \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}\left[\Big\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\Big\|^2\right]\Big] .
\end{aligned}$$

Subsequently, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] \\
\leq&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n^2}\sum_{i=1}^n \mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathsf{h}_k\|^2 \\
&+ \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] + \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}\left[\Big\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\Big\|^2\right]\Big] .
\end{aligned}$$

We now use Lemma 5 on $\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 = \gamma_{k+1}^2\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2$ and obtain:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] \\
\leq&\left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\
&+ \sum_{i=1}^n\left(\frac{\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{n} + \frac{(n-1)(1+\gamma_{k+1}\beta)}{n^2}\right)\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2\right] \\
&+ \gamma_{k+1}(1-\rho)^2\left(2\gamma_{k+1} + \frac{1}{\beta}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \left(2\gamma_{k+1}^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] \\
\leq&\left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \sum_{i=1}^n\left(\frac{1-\frac{1}{n}+\gamma_{k+1}\beta+\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{n}\right)\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2\right] \\
&+ \gamma_{k+1}(1-\rho)^2\left(2\gamma_{k+1} + \frac{1}{\beta}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \left(2\gamma_{k+1}^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\Big\|\eta_{i_k}^{(k+1)}\Big\|^2] .
\end{aligned}$$

Let us define

$$\Delta^{(k)} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \; .$$

From the above, we get

$$\begin{aligned}\Delta^{(k+1)} \leq & \left(1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2\right)\Delta^{(k)} + \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\ & + \gamma_{k+1}(1-\rho)^2\left(2\gamma_{k+1} + \frac{1}{\beta}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \gamma_{k+1}\left(2\gamma_{k+1} + \frac{\rho^2}{\beta}\right)\mathbb{E}[\left\|\eta_{i_k}^{(k+1)}\right\|^2] \; .\end{aligned}$$

Setting $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{2, 1+2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \overline{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we observe that

$$1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2 \leq 1 - \frac{1}{n} + \frac{1}{\alpha k n} + \frac{1}{\alpha^2 c_1^2 k^2 n^{\frac{4}{3}}} \leq 1 - \frac{c_1(k\alpha - 1) - 1}{k\alpha n c_1} \leq 1 - \frac{1}{k\alpha n c_1}$$

which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2 \in (0,1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} = \frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

$$\begin{aligned}\Delta^{(k+1)} \leq & \sum_{\ell=0}^{k}\omega_{k,\ell}\left(2\gamma_{\ell+1}^2\rho^2 + \frac{\gamma_{\ell+1}^2\rho^2}{\beta}\right)\mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(\ell)} - \hat{\boldsymbol{s}}^{(\ell)}\right\|^2\right] \\ & + \sum_{\ell=0}^{k}\omega_{k,\ell}\gamma_{\ell+1}(1-\rho)^2\left(2\gamma_{\ell+1} + \frac{1}{\beta}\right)\mathbb{E}\left[\left\|\tilde{S}^{(\ell)} - \hat{\boldsymbol{s}}^{(\ell)}\right\|^2\right] + \sum_{\ell=0}^{k}\omega_{k,\ell}\gamma_{\ell+1}\tilde{\epsilon}^{(\ell+1)} \; ,\end{aligned}$$

where $\omega_{k,\ell} = \prod_{j=\ell+1}^{k}\left(1 - \Lambda_{(j)}\right)$ and $\tilde{\epsilon}^{(\ell+1)} = \left(2\gamma_{k+1} + \frac{\rho^2}{\beta}\right)\mathbb{E}[\left\|\eta_{i_k}^{(k+1)}\right\|^2]$.

Summing on both sides over $k = 0$ to $k = \mathsf{K_m} - 1$ yields:

$$\begin{aligned}\sum_{k=0}^{\mathsf{K_m}-1}\Delta^{(k+1)} \leq & \sum_{k=0}^{\mathsf{K_m}-1}\frac{2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}}{\Lambda_{(k+1)}}\mathbb{E}[\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] \\ & + \sum_{k=0}^{\mathsf{K_m}-1}\frac{\gamma_{k+1}(1-\rho)^2\left(2\gamma_{k+1} + \frac{1}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \sum_{k=0}^{\mathsf{K_m}-1}\frac{\gamma_{k+1}}{\Lambda_{(k+1)}}\tilde{\epsilon}^{(k+1)} \; .\end{aligned}$$

We recall (31) where we have summed on both sides from $k = 0$ to $k = \mathsf{K_m} - 1$:

$$\begin{aligned}& \mathbb{E}\big[V(\hat{\mathbf{s}}^{(\mathsf{K_m})}) - V(\hat{\mathbf{s}}^{(0)})\big] \\ \leq & \sum_{k=0}^{\mathsf{K_m}-1}\left\{\gamma_{k+1}(-(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}\rho^2\,\mathrm{L}_V)\mathbb{E}[\|\mathsf{h}_k\|^2] + \gamma^2\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2\,\Delta^{(k)}\right\} \\ & + \sum_{k=0}^{\mathsf{K_m}-1}\left\{\tilde{\xi}^{(k+1)} + \left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\right\} \\ \leq & \sum_{k=0}^{\mathsf{K_m}-1}\left\{\left[-\gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}^2\rho^2\,\mathrm{L}_V + \frac{\rho^2\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right]\mathbb{E}[\|\mathsf{h}_k\|^2]\right\} \\ & + \sum_{k=0}^{\mathsf{K_m}-1}\Xi^{(k+1)} + \sum_{k=0}^{\mathsf{K_m}-1}\Gamma^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] \; ,\end{aligned} \tag{33}$$

where

$$\Xi^{(k+1)} = \tilde{\xi}^{(k+1)} + \frac{\gamma_{k+1}^3\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{\Lambda_{(k+1)}}\tilde{\epsilon}^{(k+1)}$$

and

$$\Gamma^{(k+1)} = \left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2}\right) + \frac{\gamma_{k+1}^3\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2(1-\rho)^2\left(2\gamma_{k+1}+\frac{1}{\beta}\right)}{\Lambda_{(k+1)}} \ .$$

We now analyse the following quantity

$$\begin{aligned}
&-\gamma_{k+1}(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}^2\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\\
&=\gamma_{k+1}\left[-(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right] \ .
\end{aligned} \tag{34}$$

Furthermore, we recall that $c_1=\upsilon_{\min}^{-1}$, $\alpha=\max\{2,1+2\upsilon_{\min}\}$, $\overline{L}=\max\{\mathrm{L}_{\mathbf{s}},\mathrm{L}_V\}$, $\gamma_{k+1}=\frac{1}{k}$, $\beta=\frac{1}{\alpha n}$, $\rho=\frac{1}{\alpha c_1\overline{L}n^{2/3}}$, $c_1(k\alpha-1)\geq c_1(\alpha-1)\geq 2$, $\alpha\geq 2$.Then,

$$\begin{aligned}
&\gamma_{k+1}\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\frac{1}{n}-\gamma_{k+1}\beta-\gamma_{k+1}^2\rho^2\,\mathrm{L}_{\mathbf{s}}^2}\\
\leq&\frac{1}{k\alpha^2c_1^2\overline{L}n^{4/3}}+\frac{\overline{L}(k\alpha^2c_1^2n^{4/3})^{-1}\left(\frac{2}{k^2\alpha^2c_1^2\overline{L}^2n^{4/3}}+\frac{1}{k\alpha c_1^2\overline{L}^2n^{1/3}}\right)}{\frac{1}{n}-\frac{1}{k\alpha n}-\frac{1}{k^2\alpha^2c_1^2n^{4/3}}}\\
=&\frac{1}{k\alpha^2c_1^2\overline{L}n^{4/3}}+\frac{\overline{L}\left(\frac{2}{k^2\alpha^2c_1^2\overline{L}^2n^{4/3}}+\frac{1}{k\alpha c_1^2\overline{L}^2n^{1/3}}\right)}{(k\alpha c_1n^{1/3})(k\alpha-1)c_1-1}\\
\overset{(a)}{\leq}&\frac{1}{k\alpha^2c_1^2\overline{L}n^{4/3}}+\frac{\frac{1}{k\alpha c_1^2\overline{L}n^{1/3}}\left(\frac{2}{k\alpha n}+1\right)}{2(\alpha c_1n^{1/3})-1}\\
\leq&\frac{1}{k^2\alpha c_1^2\overline{L}n^{4/3}}+\frac{1}{4k\alpha^2c_1^3\overline{L}n^{2/3}}\\
\leq&\frac{3/4}{\alpha c_1^2\overline{L}n^{2/3}} \ ,
\end{aligned} \tag{35}$$

where $(a)$ is due to $c_1(k\alpha-1)\geq c_1(\alpha-1)\geq 2$ and $k\alpha c_1n^{1/3}\geq 1$. Note also that

$$-(\upsilon_{\min}\rho+\upsilon_{\max}^2)\leq-\rho\upsilon_{\min}=-\frac{1}{\alpha c_1^2\overline{L}n^{2/3}} \ ,$$

which yields that

$$\left[-(\upsilon_{\min}\rho+\upsilon_{\max}^2)+\gamma_{k+1}\rho^2\,\mathrm{L}_V+\frac{\rho^2\gamma_{k+1}\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2+\frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right]\leq-\frac{1/4}{\alpha c_1^2\overline{L}n^{2/3}} \ .$$

Using the Lemma 2, we know that $\upsilon_{\max}^2\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2\leq\|\hat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2$ and using (35) on (33) yields:

$$\begin{aligned}
&\upsilon_{\max}^2\sum_{k=0}^{\mathsf{K_m}-1}\gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2]\\
\leq&\frac{4\alpha\overline{L}n^{2/3}}{\upsilon_{\min}^2}\left[V(\hat{\mathbf{s}}^{(0)})-V(\hat{\mathbf{s}}^{(\mathsf{K_m})})\right]\\
&+\frac{4\alpha\overline{L}n^{2/3}}{\upsilon_{\min}^2}\sum_{k=0}^{\mathsf{K_m}-1}\Xi^{(k+1)}+\sum_{k=0}^{\mathsf{K_m}-1}\Gamma^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)}-\tilde{S}^{(k)}\|^2\right] \ ,
\end{aligned}$$

proving the bound on the second order moment of the gradient of the Lyapunov function:

$$\sum_{k=0}^{\mathsf{K_m}-1} \gamma_{k+1}\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\|^2] \leq \frac{4\alpha\overline{L}n^{2/3}}{\upsilon_{\min}^2\upsilon_{\max}^2}\left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(\mathsf{K_m})})\right] + \frac{4\alpha\overline{L}n^{2/3}}{\upsilon_{\min}^2\upsilon_{\max}^2}\sum_{k=0}^{\mathsf{K_m}-1} \Xi^{(k+1)} + \sum_{k=0}^{\mathsf{K_m}-1} \Gamma^{(k+1)}\mathbb{E}\left[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\right] .$$

$\square$

# C Practical Implementations of Two-Timescale EM Methods

## C.1 Application on GMM

### C.1.1 Explicit Updates

We first recognize that the constraint set for $\boldsymbol{\theta}$ is given by

$$\Theta = \Delta^M \times \mathbb{R}^M .$$

Using the partition of the sufficient statistics as $S(y_i, z_i) = (S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ and the fact that $\mathbb{1}_{\{M\}}(z_i) = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{m\}}(z_i)$, the complete data log-likelihood can be expressed as in (2) with

$$\begin{aligned}
&s_{i,m}^{(1)} = \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \textstyle\sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\} , \\
&s_{i,m}^{(2)} = \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m \ , \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M \ ,
\end{aligned} \tag{36}$$

and $\psi(\boldsymbol{\theta}) = -\left\{ \log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{\mu_M^2}{2\sigma^2} \right\}$. We also define for each $m \in [\![1, M]\!]$, $j \in [\![1, 3]\!]$, $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of the conditional expected value $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i = m\}} | y = y_i]$:

$$z_{i,m} \sim \mathbb{P}\left( z_i = m | y_i; \boldsymbol{\theta} \right) \tag{37}$$

where $m \in [\![1, M]\!]$, $i \in [n]$ and $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\mu}) \in \Theta$.

In particular, given iteration $k+1$, the computation of the approximated quantity $\tilde{S}_{i_k}^{(k)}$ during Incremental-step updates, see (8) can be written as

$$\tilde{S}_{i_k}^{(k)} = \Big( \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:= \tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}) y_{i_k}, \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1}) y_{i_k}}_{:= \tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:= \tilde{\mathbf{s}}_{i_k}^{(3)}(\boldsymbol{\theta}^{(k)})} \Big)^\top . \tag{38}$$

Recall that we have used the following regularizer:

$$\mathrm{r}(\boldsymbol{\theta}) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log\left( 1 - \sum_{m=1}^{M-1} \omega_m \right) , \tag{39}$$

It can be shown that the regularized M-step evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1} \big( s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon \big)^\top \\ \big( (s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)} \big)^\top \\ \big( 1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta \big)^{-1} \big( s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)} \big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{40}$$

where we have defined for all $m \in [\![1, M]\!]$ and $j \in [\![1, 3]\!]$ , $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$.

### C.1.2 Model Assumptions (GMM example)

We use the GMM example to illustrate the required assumptions.

Many practical models can satisfy the compactness of the sets as in Assumption A1 For instance, the GMM example satisfies (11) as the sufficient statistics are composed of indicator functions and observations as defined Section C.1 Equation (36).

Assumptions A2 and A3 are standard for the curved exponential family models. For GMM, the following (strongly convex) regularization $\mathrm{r}(\boldsymbol{\theta})$ ensures A3:

$$\mathrm{r}(\boldsymbol{\theta}) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log\left( 1 - \sum_{m=1}^{M-1} \omega_m \right) ,$$

since it ensures $\boldsymbol{\theta}^{(k)}$ is unique and lies in $\text{int}(\Delta^M) \times \mathbb{R}^M$. We remark that for A2, it is possible to define the Lipschitz constant $\text{L}_p$ independently for each data $y_i$ to yield a refined characterization.

Again, A4 is satisfied by practical models. For GMM, it can be verified by deriving the closed form expression for $\text{B}(\mathbf{s})$ and using A1.

Under A1 and A3, we have $\|\hat{\boldsymbol{s}}^{(k)}\| < \infty$ since $\mathsf{S}$ is compact and $\hat{\boldsymbol{\theta}}^{(k)} \in \text{int}(\Theta)$ for any $k \geq 0$ which thus ensure that the EM methods operate in a closed set throughout the optimization process.

### C.1.3 Algorithms updates

In the sequel, recall that, for all $i \in [n]$ and iteration $k$, the computed statistic $\tilde{S}_{i_k}^{(k)}$ is defined by (38). At iteration $k$, the several E-steps defined by (1) or (2) and (3) leads to the definition of the quantity $\hat{\mathbf{s}}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{\mathbf{s}}^{(0)} = n^{-1} \sum_{i=1}^n \overline{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

**Batch EM (EM):** for all $i \in [n]$, compute $\overline{\mathbf{s}}_i^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1} \sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(k)} .$$

where $\overline{\mathbf{s}}_i^{(k)}$ are computed using the exact conditional expected balue $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y = y_i]$:

$$\widetilde{\omega}_m(y_i; \boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y = y_i] = \frac{\omega_m \exp(-\frac{1}{2}(y_i - \mu_i)^2)}{\sum_{j=1}^M \omega_j \exp(-\frac{1}{2}(y_i - \mu_j)^2)} ,$$

**Incremental EM (iEM):** draw an index $i_k$ uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \frac{1}{n}\big(\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)}\big) = n^{-1} \sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(\tau_i^k)} .$$

**batch SAEM (SAEM):** draw an index $i_k$ uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1}\tilde{S}^{(k)} .$$

where $= \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (38).

**Incremental SAEM (iSAEM):** draw an index $i_k$ uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)} + \frac{1}{n}(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)})\big) .$$

**Variance Reduced Two-Timescale EM (vrTTEM):** draw an index $i_k$ uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1 - \rho) + \rho(\tilde{S}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big))\big) .$$

**Fast Incremental Two-Timescale EM (fiTTEM):** draw an index $i_k$ uniformly at random on $[n]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1 - \rho) + \rho(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big))\big) .$$

Finally, the $k$-th update reads $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ where the function $\boldsymbol{s} \to \overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (40).

### C.2 Deformable Template Model for Image Analysis

#### C.2.1 Model and Updates

The complete model belongs to the curved exponential family, see [2], which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$
\begin{aligned}
S_1(z) &= \frac{1}{n}\sum_{i=1}^{n} S_1(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top y_i \ , \\
S_2(z) &= \frac{1}{n}\sum_{i=1}^{n} S_2(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top \left(\mathbf{K}_p^{z_i}\right) \ , \\
S_3(z) &= \frac{1}{n}\sum_{i=1}^{n} S_3(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i \ ,
\end{aligned}
\tag{41}
$$

where for any pixel $u \in \mathbb{R}^2$ and $j \in [\![1, k_g]\!]$ we denote:

$$
\mathbf{K}_p^{z_i}(x_u, j) = \mathbf{K}_p^{z_i}(x_u - \phi_i(x_u, z_i), p_j) \ .
$$

Finally, the Two-Timescale M-step yields the following parameter updates:

$$
\bar{\boldsymbol{\theta}}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n}\hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z) \end{pmatrix} \ ,
\tag{42}
$$

where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via the SA-step (7) and using the MC approximation of the sufficient statistics $\big(S_1(z), S_2(z), S_3(z)\big)$ defined in (41).

#### C.2.2 Numerical Applications

For the inference of the template, we use the Matlab code (online SAEM) used in [24] and implement our own batch, incremental, Variance reduced and Fast Incremental variants. The hyperparameters are kept the same and reads as follows $M = 400$, $\gamma_k = 1/k^{0.6}$ and $p = 16$. The number of landmarks for the template is $k_p = 15$ points and for the deformation $k_g = 6$ points. Both have Gaussian kernels with respectively standard deviation of $0.12$ and $0.3$. The standard deviation of the measurement errors is set to $0.1$.

For the simulation part, we use the Carlin and Chib MCMC procedure, see [9]. Refer to [24] for more details.

# D Additional Experiment: Pharmacokinetics (PK) Model with Absorption Lag Time

This numerical example was conducted in order to characterize the pharmacokinetics (PK) of orally administered drug to simulated patients, using a population pharmacokinetics approach. $M = 50$ synthetic datasets were generated for $n = 5000$ patients with 10 observations (concentration measures) per patient. The goal tis to model the evolution of the concentration of the absorbed drug using a nonlinear and latent variable model.

**Model and Explicit Updates:** We consider a one-compartment PK model for oral administration with an absorption lag-time ($T^{\text{lag}}$), assuming first-order absorption and linear elimination processes. The final model includes the following variables: $ka$ the absorption rate constant, $V$ the volume of distribution, $k$ the elimination rate constant and $T^{\text{lag}}$ the absorption lag-time. We also add several covariates to our model such as $D$ the dose of drug administered, $t$ the time at which measures are taken and the weight of the patient influencing the volume $V$. More precisely, the log-volume $\log(V)$ is a linear function of the log-weight $lw70 = \log(wt/70)$. Let $z_i = (T_i^{\text{lag}}, ka_i, V_i, k_i)$ be the vector of individual PK parameters, different for each individual $i$. The final model reads:

$$y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} \quad \text{where} \quad f(t_{ij}, z_i) = \frac{D\, ka_i}{V(ka_i - k_i)}(\mathrm{e}^{-ka_i\,(t_{ij} - T_i^{\text{lag}})} - \mathrm{e}^{-k_i\,(t_{ij} - T_i^{\text{lag}})})\ , \quad (43)$$

where $y_{ij}$ is the $j$-th concentration measurement of the drug of dosage $D$ injected at time $t_{ij}$ for patient $i$. We assume in this example that the residual errors $\varepsilon_{ij}$ are independent and normally distributed with mean 0 and variance $\sigma^2$. Lognormal distributions are used for the four PK parameters.

Lognormal distributions are used for the four PK parameters:

$$\begin{aligned}
&\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2)\ , \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2)\ , \\
&\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2)\ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2)\ .
\end{aligned}$$

We recall that the complete model $(y, z)$ defined by (43) belongs to the curved exponential family, which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$S_1(z) = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad S_2(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^\top z_i, \quad S_3(z) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(t_i, z_i))^2 \quad (44)$$

where we have noted $y_i$ and $t_i$ the vector of observations and time for each patient $i$. At iteration $k$, and setting the number of MC samples to 1 for the sake of clarity, the MC sampling $z_i^{(k)} \sim p(z_i|y_i, \theta^{(k)})$ is performed using a Metropolis-Hastings procedure detailed in Algorithm 2. The quantities $\tilde{S}^{(k+1}$ and $\hat{\mathrm{s}}^{(k+1)}$ are then updated according to the different methods. Finally the maximization step yields:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} \hat{\mathrm{s}}_1^{(k+1)} \\ \hat{\mathrm{s}}_2^{(k+1)} - \hat{\mathrm{s}}_1^{(k+1)} \left(\hat{\mathrm{s}}_1^{(k+1)}\right)^\top \\ \hat{\mathrm{s}}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{z}_{\mathbf{pop}}}(\hat{\mathrm{s}}^{(k+1)}) \\ \overline{\boldsymbol{\omega}_{\boldsymbol{z}}}(\hat{\mathrm{s}}^{(k+1)}) \\ \overline{\boldsymbol{\sigma}}(\hat{\mathrm{s}}^{(k+1)}) \end{pmatrix}\ . \quad (45)$$

where $z_{\text{pop}}$ denotes the vector of fixed effects $(T_{\text{pop}}^{\text{lag}}, ka_{\text{pop}}, V_{\text{pop}}, k_{\text{pop}})$.

**Metropolis Hastings algorithm.** During the simulation step of the MISSO method, the sampling from the target distribution $\pi(z_i, \boldsymbol{\theta}) := p(z_i|y_i, \boldsymbol{\theta})$ is performed using a Metropolis Hastings (MH) algorithm [27] with proposal distribution $q(z_i, \delta)$ where $\boldsymbol{\theta} = (z_{\text{pop}}, \omega_z)$ and $\delta$ is the vector of parameters of the proposal distribution. Commonly they parameterize a Gaussian proposal. The MH algorithm is summarized in 2.

**Algorithm 2** MH aglorithm

**Input:** initialization $z_{i,0} \sim q(z_i; \boldsymbol{\delta})$
**for** $m = 1, \cdots, M$ **do**
  Sample $z_{i,m} \sim q(z_i; \boldsymbol{\delta})$
  Sample $u \sim \mathcal{U}([\![0, 1]\!])$
  Calculate the ratio $r = \frac{\pi(z_{i,m};\boldsymbol{\theta})/q(z_{i,m});\boldsymbol{\delta})}{\pi(z_{i,m-1};\boldsymbol{\theta})/q(z_{i,m-1});\boldsymbol{\delta})}$
  **if** $u < r$ **then**
    Accept $z_{i,m}$
  **else**
    $z_{i,m} \leftarrow z_{i,m-1}$
  **end if**
**end for**
**Output:** $z_{i,M}$

**Monte Carlo study:** We conduct a Monte Carlo study to showcase the benefits of our scheme. $M = 50$ datasets have been simulated using the following PK parameters values: $T^{\text{lag}}_{\text{pop}} = 1$, $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{T^{\text{lag}}} = 0.4$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. We define the mean square distance over the $M$ replicates $E_k(\ell) = \frac{1}{M}\sum_{m=1}^{M}\left(\theta_k^{(m)}(\ell) - \theta^*\right)^2$ and plot it against the epochs (passes over the data) Figure 4. Note that the MC-step (5) is performed using a Metropolis Hastings procedure since the posterior distribution under the model $\theta$ noted $p(z_i|y_i, \theta)$ is intractable due to the nonlinearity of the model (43). Figure 4 shows clear advantage of variance reduced methods (vrTTEM and fiTTEM ) avoiding the twists and turns displayed by the incremental and the batch methods.

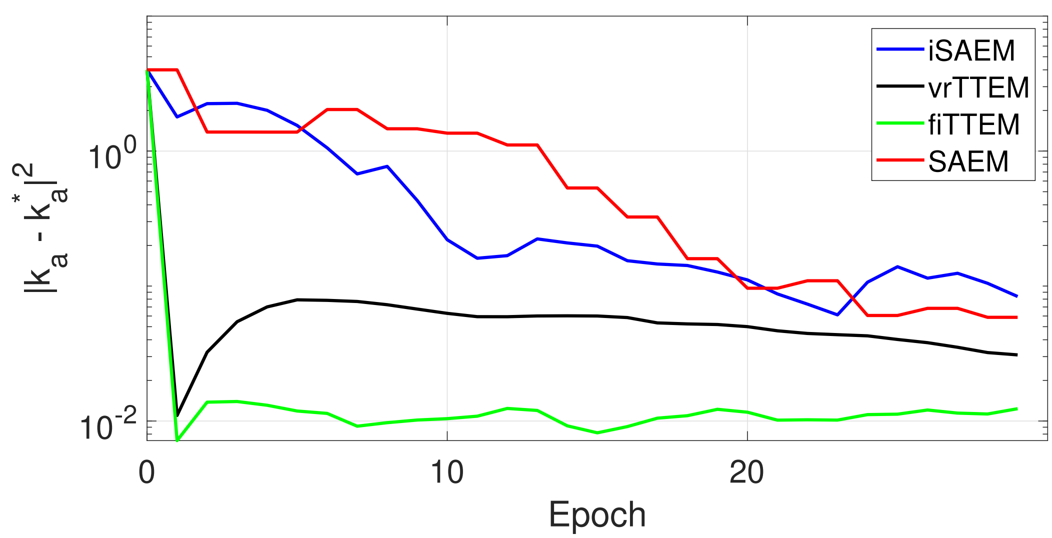


Figure 4: Precision $|ka^{(k)} - ka^*|^2$ per epoch