Appendix
High-Performance Distributed ML at Scale through Parameter Server Consistency Models

1 Comparison with GraphLab

We compare our MF SGD implementation using Petuum-ESSPTable with with GraphLab's synchronous engine on Netflix data set (Fig. 1). GraphLab's asynchronous engine performs worse than synchronous engine and is not shown. The experiments were run on 8 nodes, each with 64 cores and 128GB memory, connected via 1Gbps ethernet. We use $\lambda = 0.05$, and the step-sizes are calibrated for both systems to be equal.

We observe that Petuum-ESSP converges significantly faster than GraphLab in real time due to Petuum-ESSP's high throughput. Each iteration in Petumm's MF takes about 15.3 seconds, while GraphLab takes about 97 seconds. Therefore over 1850 seconds experiment run time Petuum executes 120 iterations compared with GraphLab's 19 iterations. Notice that Petuum's first recorded objective occurs at about iteration 16, which has objective value comparable to GraphLab's last recorded objective (right-most end of blue curve) at iteration 19. We believe this large difference comes from Petuum's relaxed consistency model which allows computation to proceed with staler parameters, resulting in much higher throughput.

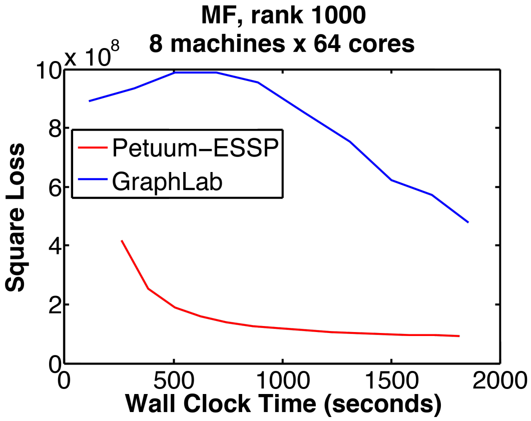

Figure 1: MF Convergence curve for Petuum-ESSPTable and GraphLab.

2 Proof of Theorems

Theorem 1 (SGD under VAP, convergence in expectation) Given convex function $f(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x})$ such that components f_t are also convex. We search for minimizer $\mathbf{x}^*$ via gradient descent on each component ∇f_t with step-size $\breve{\eta}_t$ close to $\eta_t = \frac{\eta}{\sqrt{t}}$ such that the update $\hat{\mathbf{u}}_t = -\breve{\eta}_t \nabla f_t(\breve{\mathbf{x}}_t)$ is computed on noisy view $\breve{\mathbf{x}}_t$. The VAP bound follows the decreasing v_t described above. Under suitable conditions (f_t are L-Lipschitz and bounded diameter $D(x\|x') \leq F^2$),

$$R[X] := \sum_{t=1}^{T} f_t(\breve{\mathbf{x}}_t) - f(\mathbf{x}^*) = \mathcal{O}(\sqrt{T})$$

and thus $\frac{R[X]}{T} \to 0$ as $T \to \infty$.

Proof. We will use real-time sequence $\hat{\mathbf{x}}_t$ defined by

$$\hat{\mathbf{x}}_t := \mathbf{x}_0 + \sum_{t'=1}^{t} \hat{\mathbf{u}}_{t'}$$

$$\begin{aligned}
R[X] &= \sum_{t=1}^{T} f_t(\breve{\mathbf{x}}_t) - f(\mathbf{x}^*) \\
&\leq \sum_{t=1}^{T} \langle \nabla f_t(\breve{\mathbf{x}}_t), \breve{\mathbf{x}}_t - \mathbf{x}^* \rangle && (f_t \text{ are convex}) \\
&= \sum_{t=1}^{T} \langle \breve{\boldsymbol{g}}_t, \breve{\mathbf{x}}_t - \mathbf{x}^* \rangle
\end{aligned}$$

where $\breve{\boldsymbol{g}}_t := \nabla f_t(\breve{\mathbf{x}}_t)$. From Lemma A.1 below we have

$$R[X] \leq \sum_{t=1}^{T} \frac{1}{2}\breve{\eta}_t ||\breve{\boldsymbol{g}}_t||^2 + \frac{D(\mathbf{x}^*||\hat{\mathbf{x}}_t) - D(\mathbf{x}^*||\hat{\mathbf{x}}_{t+1})}{\breve{\eta}_t} + \langle \breve{\mathbf{x}}_t - \hat{\mathbf{x}}_t, \breve{\boldsymbol{g}}_t \rangle$$

We now bound each term:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{2}\breve{\eta}_t ||\breve{\boldsymbol{g}}_t||^2 &\leq \sum_{t=1}^{T} \frac{1}{2}\breve{\eta}_t L^2 && \text{(Lipschitz assumption)} \\
&= \sum_{t=r+1}^{T} \frac{1}{2} \frac{\eta}{\sqrt{t-r}} L^2 + const && (r > 0 \text{ is the finite clock drift in VAP}) \\
&= \frac{1}{2}\eta L^2 \sum_{t=r+1}^{T} \frac{1}{\sqrt{t-r}} + const \\
&\leq \frac{1}{2}\eta L^2 \int_{t=r+1}^{T} \frac{1}{\sqrt{t-r}} dt + const \\
&\leq \frac{1}{2}\eta L^2 (\sqrt{T-r} - 1) + const \\
&= \mathcal{O}(\sqrt{T})
\end{aligned}$$

where the clock drift comes from the fact that $\breve{\eta}_t$ is not exactly $\eta_t = \frac{\eta}{\sqrt{t}}$ in VAP.

$$\begin{aligned}
\sum_{t=1}^{T} \frac{D(\mathbf{x}^*||\hat{\mathbf{x}}_t) - D(\mathbf{x}^*||\hat{\mathbf{x}}_{t+1})}{\breve{\eta}_t} &= \frac{D(\mathbf{x}^*||\hat{\mathbf{x}}_1)}{\breve{\eta}_1} - \frac{D(\mathbf{x}^*||\hat{\mathbf{x}}_{T+1})}{\breve{\eta}_T} + \sum_{t=2}^{T} \left[D(\mathbf{x}^*||\hat{\mathbf{x}}_t) \left(\frac{1}{\breve{\eta}_t} - \frac{1}{\breve{\eta}_{t-1}} \right) \right] \\
&\leq \frac{F^2}{\eta} + 0 + \frac{F^2}{\eta} \sum_{t=2}^{T} \left[\sqrt{t-k} - \sqrt{t-r} \right] \quad \text{(clock drift)} \\
&\leq \frac{F^2}{\eta} + \frac{F^2}{\eta} \int_{t=\max(k,r)}^{T} \left(\sqrt{t-k} - \sqrt{t-r} \right) dt + const \\
&= \frac{F^2}{\eta} + \frac{F^2}{\eta} \left[(t-k)^{3/2} - (t-r)^{3/2} \right]_{max(k,r)}^{T} + const \\
&= \frac{F^2}{\eta} + \frac{F^2}{\eta} \left[(T-k)^{3/2} - (T-r)^{3/2} \right] + const \\
&= \frac{F^2}{\eta} + \frac{F^2}{\eta} \left[\left(T^{\frac{3}{2}} + \frac{3}{2} k T^{\frac{1}{2}} + \mathcal{O}(\sqrt{T}) \right) \right. \\
&\quad \left. - \left(T^{\frac{3}{2}} + \frac{3}{2} r T^{\frac{1}{2}} + \mathcal{O}(\sqrt{T}) \right) \right] + const \quad \text{(binomial expansion)} \\
&= \mathcal{O}(\sqrt{T})
\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^{T}\langle\breve{\mathbf{x}}_t-\hat{\mathbf{x}}_t,\breve{\boldsymbol{g}}_t\rangle &\leq \sum_{t=1}^{T}||\breve{\mathbf{x}}_t-\hat{\mathbf{x}}_t||_2||\breve{\boldsymbol{g}}_t||_2\\ &\leq \sum_{t=1}^{T}\sqrt{d}v_tL \quad \text{(using eq.(2) from main text)}\\ &= \sqrt{d}L\sum_{t=1}^{T}\frac{v_0}{\sqrt{t}}\\ &= \sqrt{d}Lv_0\sqrt{T}=\mathcal{O}(\sqrt{T})\end{aligned}$$

Together, we have $R[X]\leq\mathcal{O}(\sqrt{T})$ as desired.

□

Lemma A.1 For $\mathbf{x}^*,\breve{\mathbf{x}}_t\in X$, and $X=\mathbb{R}^d$,

$$\langle\breve{\boldsymbol{g}}_t,\breve{\mathbf{x}}_t-\mathbf{x}^*\rangle=\frac{1}{2}\breve{\eta}_t||\breve{\boldsymbol{g}}_t||^2+\frac{D(\mathbf{x}^*||\hat{\mathbf{x}}_t)-D(\mathbf{x}^*||\hat{\mathbf{x}}_{t+1})}{\breve{\eta}_t}+\langle\breve{\mathbf{x}}_t-\hat{\mathbf{x}}_t,\breve{\boldsymbol{g}}_t\rangle$$

where $D(x||x'):=\frac{1}{2}||x-x'||^2$.

Proof.

$$\begin{aligned}D(\mathbf{x}^*||\hat{\mathbf{x}}_t)-D(\mathbf{x}^*||\hat{\mathbf{x}}_{t+1}) &= \frac{1}{2}||\mathbf{x}^*-\hat{\mathbf{x}}_t+\hat{\mathbf{x}}_t-\hat{\mathbf{x}}_{t+1}||^2-\frac{1}{2}||\mathbf{x}^*-\hat{\mathbf{x}}_t||^2\\ &= \frac{1}{2}||\mathbf{x}^*-\hat{\mathbf{x}}_t+\breve{\eta}_t\breve{\boldsymbol{g}}_t||^2-\frac{1}{2}||\mathbf{x}^*-\hat{\mathbf{x}}_t||^2\\ &= \frac{1}{2}\breve{\eta}_t||\breve{\boldsymbol{g}}_t||^2-\breve{\eta}_t\langle\hat{\mathbf{x}}_t-\mathbf{x}^*,\breve{\boldsymbol{g}}_t\rangle\end{aligned}$$

Divide both sides by $\breve{\eta}_t$ gets the desired answer.

□

Lemma 4 $\bar{u}_t\leq\frac{\eta}{\sqrt{t}}L$ *and* $\gamma_t:=||\boldsymbol{\gamma}_t||_2\leq P(2s+1)$.

Proof. $||\mathbf{u}_t||_2=||-\eta_t\nabla f_t||_2\leq\frac{\eta}{\sqrt{t}}L$ since f is L-Lipschitz. Therefore $\bar{u}_t=\frac{1}{P(2s+1)}\sum_{t'\in\mathcal{W}_t}||\mathbf{u}_{t'}||_2\leq\frac{\eta}{\sqrt{t}}L$ since $|\mathcal{W}_t|\leq P(2s+1)$.

If $\bar{u}_t=0$, then $\boldsymbol{\gamma}_t=\mathbf{0}$ and the lemma holds trivially. For $\bar{u}_t>0$. $\boldsymbol{\gamma}_t=\frac{1}{\bar{u}_t}(\tilde{\mathbf{x}}_t-\mathbf{x}_t)=\frac{1}{\bar{u}_t}\sum_{t'\in\mathcal{S}_t}\mathbf{u}_{t'}$. Thus $||\boldsymbol{\gamma}_t||_2=\frac{1}{\bar{u}_t}||\sum_{t'\in\mathcal{S}_t}\mathbf{u}_{t'}||_2\leq\frac{1}{\bar{u}_t}\sum_{t'\in\mathcal{S}_t}||\mathbf{u}_{t'}||_2\leq\frac{1}{\bar{u}_t}\sum_{t'\in\mathcal{W}_t}||\mathbf{u}_{t'}||_2=P(2s+1)$. □

Theorem 5 *(SGD under SSP, convergence in probability) Given convex function $f(\boldsymbol{x})=\sum_{t=1}^{T}f_t(\boldsymbol{x})$ such that components f_t are also convex. We search for minimizer $\boldsymbol{x}^*$ via gradient descent on each component ∇f_t under SSP with staleness s and P workers. Let $\boldsymbol{u}_t:=-\eta_t\nabla_t f_t(\tilde{\boldsymbol{x}}_t)$ with $\eta_t=\frac{\eta}{\sqrt{t}}$. Under suitable conditions (f_t are L-Lipschitz and bounded divergence $D(x||x')\leq F^2$), we have*

$$P\left[\frac{R\left[X\right]}{T}-\frac{1}{\sqrt{T}}\left(\eta L^2+\frac{F^2}{\eta}+2\eta L^2\mu_\gamma\right)\geq\tau\right]\leq\exp\left\{\frac{-T\tau^2}{2\bar{\eta}_T\sigma_\gamma+\frac{2}{3}\eta L^2(2s+1)P\tau}\right\}$$

where $R[X]:=\sum_{t=1}^{T}f_t(\tilde{x}_t)-f(x^*)$, and $\bar{\eta}_T=\frac{\eta^2L^4(\ln T+1)}{T}=o(T)$.

Proof. From lemma A.1, substitute $\breve{\mathbf{x}}_t$ with $\tilde{x}_t$ we have

$$\begin{aligned} R\left[X\right] &\leq \sum_{t=1}^{T} \langle \tilde{g}_t, \tilde{x}_t - x^* \rangle \\ &= \sum_{t=1}^{T} \frac{1}{2}\eta_t \left\|\tilde{g}_t\right\|^2 + \frac{D\left(x^*\|x_t\right) - D\left(x^*\|x_{t+1}\right)}{\eta_t} + \langle \tilde{x}_t - x_t, \tilde{g}_t \rangle \\ &\leq \eta L^2 \sqrt{T} + \frac{F^2}{\eta}\sqrt{T} + \sum_{t=1}^{T} \langle \bar{u}_t \boldsymbol{\gamma}_t, \tilde{g}_t \rangle \\ &\leq \eta L^2 \sqrt{T} + \frac{F^2}{\eta}\sqrt{T} + \sum_{t=1}^{T} \frac{\eta}{\sqrt{t}} L^2 \gamma_t \end{aligned}$$

Where the last step uses the fact

$$\begin{aligned} \langle \bar{u}_t \boldsymbol{\gamma}_t, \tilde{g}_t \rangle &\leq \bar{u}_t ||\boldsymbol{\gamma}_t||_2 ||\tilde{g}_t||_2 \\ &\leq \gamma_t \frac{\eta}{\sqrt{t}} L^2 && \text{(Lemma 4)} \end{aligned}$$

Dividing T on both sides,

$$\frac{R\left[X\right]}{T} - \frac{\eta L^2}{\sqrt{T}} - \frac{F^2}{\eta\sqrt{T}} \quad \leq \quad \frac{\sum_{t=1}^{T} \frac{\eta}{\sqrt{t}} L^2 \gamma_t}{T} \tag{1}$$

Let $a_t := \frac{\eta}{\sqrt{t}} L^2(\gamma_t - \mu_\gamma)$. Notice that a_t zero-mean, and $|a_t| \leq \eta L^2 \max_t(\gamma_t) \leq \eta L^2(2s+1)P$. Also, $\frac{1}{T}\sum_{t=1}^{T} var(a_t) = \frac{1}{T}\sum_{t=1}^{T} \frac{\eta^2}{t} L^4 \sigma_\gamma < \frac{\eta^2 L^4 \sigma_\gamma}{T}(\ln T + 1) = \bar{\eta}_T \sigma_\gamma$ where $\bar{\eta}_T = \frac{\eta^2 L^4 (\ln T + 1)}{T}$. Bernstein's inequality gives, for $\tau > 0$,

$$P\left(\frac{\sum_{t=1}^{T} \frac{\eta}{\sqrt{t}} L^2 \gamma_t - \frac{\eta}{\sqrt{t}} L^2 \mu_\gamma}{T} \geq \tau\right) \leq \exp\left\{\frac{-T\tau^2}{2\bar{\eta}_T \sigma_\gamma + \frac{2}{3}\eta L^2 (2s+1) P \tau}\right\} \tag{2}$$

Note the following identity:

$$\sum_{i=a}^{b} \frac{1}{\sqrt{i}} \leq 2\sqrt{b - a + 1} \tag{3}$$

Thus

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\eta}{\sqrt{t}} L^2 \mu_\gamma \leq \frac{2\eta L^2 \mu_\gamma}{\sqrt{T}} \tag{4}$$

Plugging eq. 1 and 4 to eq. 2, we have

$$P\left[\frac{R\left[X\right]}{T} - \frac{1}{\sqrt{T}}\left(\eta L^2 + \frac{F^2}{\eta} + 2\eta L^2 \mu_\gamma\right) \geq \tau\right] \leq \exp\left\{\frac{-T\tau^2}{2\bar{\eta}_T \sigma_\gamma + \frac{2}{3}\eta L^2 (2s+1) P \tau}\right\}$$

□

We need the following Lemma to prove Theorem 2 and 6.

Lemma A.2 Let $\boldsymbol{\Omega^*}$ be the hessian of the loss at optimum $\mathbf{x}^*$, then

$$\boldsymbol{g}_t := \nabla f(\tilde{\mathbf{x}}_t) = (\tilde{\mathbf{x}}_t - \mathbf{x}^*)\boldsymbol{\Omega^*} + \mathcal{O}(\rho_t^2)$$

for $\tilde{\mathbf{x}}_t$ close to the optimum such that $\mathcal{O}(\rho_t) = \mathcal{O}(||\tilde{\mathbf{x}}_t - \mathbf{x}^*||)$ is small. Here $\boldsymbol{\Omega^*} = \nabla^2 f(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*}$ is the Hessian at the optimum

Proof. Using Taylor's theorem and expanding around $\mathbf{x}^*$,

$$\begin{aligned} f(\tilde{\mathbf{x}}_t) &= f(\mathbf{x}^*) + (\tilde{\mathbf{x}}_t - \mathbf{x}^*)^T \left.\nabla f(\mathbf{x})\right|_{\mathbf{x}=\mathbf{x}^*} \\ &\quad + \frac{1}{2}(\tilde{\mathbf{x}}_t - \mathbf{x}^*)^T \boldsymbol{\Omega^*} (\tilde{\mathbf{x}}_t - \mathbf{x}^*) + \mathcal{O}(||\tilde{\mathbf{x}}_t - \mathbf{x}^*||^3) \\ &= f(\mathbf{x}^*) + \frac{1}{2}(\tilde{\mathbf{x}}_t - \mathbf{x}^*)^T \boldsymbol{\Omega^*} (\tilde{\mathbf{x}}_t - \mathbf{x}^*) + \mathcal{O}(||\tilde{\mathbf{x}}_t - \mathbf{x}^*||^3) \end{aligned}$$

where the last step uses $\nabla f(\mathbf{x}) = 0$ at $\mathbf{x}^*$. Taking gradient w.r.t. $\tilde{\mathbf{x}}_t$,

$$\begin{aligned} \nabla f(\tilde{\mathbf{x}}_t) &= (\tilde{\mathbf{x}}_t - \mathbf{x}^*)^T \boldsymbol{\Omega^*} + \mathcal{O}(||\tilde{\mathbf{x}}_t - \mathbf{x}^*||^2) \\ &= (\tilde{\mathbf{x}}_t - \mathbf{x}^*)^T \boldsymbol{\Omega^*} + \mathcal{O}(\rho_t^2) \end{aligned}$$

□

Theorem 6 (SGD under SSP, decreasing variance) Given the setup in Theorem 5 and assumption 1-3. Further assume that $f(\mathbf{x})$ has bounded and invertible Hessian Ω^* at optimum $\mathbf{x}^*$ and γ_t is bounded. Let $\mathrm{Var}_t := \mathbb{E}[\tilde{\mathbf{x}}_t^2] - \mathbb{E}[\tilde{\mathbf{x}}_t]^2$, $\boldsymbol{g}_t = \nabla f_t(\tilde{\mathbf{x}}_t)$ then for $\tilde{\mathbf{x}}_t$ near the optima $\mathbf{x}^*$ such that $\rho_t = ||\tilde{\mathbf{x}}_t - \mathbf{x}^*||$ and $\xi_t = ||\boldsymbol{g}_t|| - ||\boldsymbol{g}_{t+1}||$ are small:

$$\begin{aligned}\mathrm{Var}_{t+1} &= \mathrm{Var}_t - 2\eta_t cov(\mathbf{x}_t, \mathbb{E}^{\Delta_t}[\boldsymbol{g}_t]) + \mathcal{O}(\eta_t \xi_t)\\ &\quad + \mathcal{O}(\eta_t^2 \rho_t^2) + \mathcal{O}^*_{\gamma_t}\end{aligned}$$

where the covariance $cov(\boldsymbol{v}_1, \boldsymbol{v}_2) := \mathbb{E}[\boldsymbol{v}_1^T \boldsymbol{v}_2] - \mathbb{E}[\boldsymbol{v}_1^T]\mathbb{E}[\boldsymbol{v}_2]$ uses inner product. $\mathcal{O}^*_{\gamma_t}$ represents high order ($\geq$ 5th) terms involving $\gamma_t = ||\boldsymbol{\gamma}_t||_\infty$. Δ_t is a random variable capturing the randomness of update $\mathbf{u}_t$ conditioned on $\mathbf{x}_t$.

Proof. We write eq. 3 from the main text as $\tilde{\mathbf{x}_t} = \mathbf{x}_t + \boldsymbol{\delta}_t$ with $\boldsymbol{\delta}_t = \bar{u}_t \boldsymbol{\gamma}_t$. Conditioned on $\mathbf{x}_t$, we have

$$p(\tilde{\mathbf{x}}_t|\mathbf{x}_t) d\tilde{\mathbf{x}}_t = p(V_t(\boldsymbol{\delta}_t, \mathbf{x}_t)) dV_t \tag{5}$$

where V_t is a random variable representing the state of $\boldsymbol{\delta}_t$ conditioned on $\mathbf{x}_t$. We can express $\mathbb{E}^{\tilde{\mathbf{x}}_t}[f(\tilde{\mathbf{x}}_t)]$ in terms of $\mathbb{E}^{\mathbf{x}_t}$ for any function $f()$ of $\tilde{\mathbf{x}}_t$:

$$\begin{aligned}\mathbb{E}^{\tilde{\mathbf{x}}_t}[f(\tilde{\mathbf{x}}_t)] &= \int_{\tilde{\mathbf{x}}_t} f(\tilde{\mathbf{x}}_t) p(\tilde{\mathbf{x}}_t) d\tilde{\mathbf{x}}_t\\ &= \int_{\tilde{\mathbf{x}}_t}\int_{\mathbf{x}_t} f(\tilde{\mathbf{x}}_t) p(\tilde{\mathbf{x}}_t|\mathbf{x}_t) p(\mathbf{x}_t) d\mathbf{x}_t d\tilde{\mathbf{x}}_t && \text{(using eq. 5)}\\ &= \int_{\mathbf{x}_t}\int_{V_t} f(\tilde{\mathbf{x}}_t) p(V_t(\boldsymbol{\delta}_t, \mathbf{x}_t)) dV_t d\mathbf{x}_t\\ &= \mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{V_t}[f(\tilde{\mathbf{x}}_t)]\right] \end{aligned}\tag{6}$$

Similarly, we have

$$\mathbb{E}^{\tilde{\mathbf{x}}_{t+1}}[f(\tilde{\mathbf{x}}_{t+1})] = \mathbb{E}^{\mathbf{x}_{t+1}}\left[\mathbb{E}^{V_{t+1}}[f(\tilde{\mathbf{x}}_{t+1})]\right] \tag{7}$$

In the same vein, we introduce random variable Δ, conditioned on $\mathbf{x}_t$:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t) d\mathbf{x}_{t+1} = p(\Delta_t(\mathbf{u}_t, \mathbf{x}_t)) d\Delta_t \tag{8}$$

since $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{u}_t$ (eq. 2 in the main text). Here Δ is a random variable representing the state of $\mathbf{u}_t$ conditioned on $\mathbf{x}_t$. Analogous to eq. 6, we have

$$\mathbb{E}^{\mathbf{x}_{t+1}}[f(\mathbf{x}_{t+1})] = \mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{\Delta_t}[f(\mathbf{x}_{t+1})]] \tag{9}$$

for some function $f()$ of $\mathbf{x}_{t+1}$. There are a few facts we will use throughout:

$$\mathbb{E}^{\mathbf{x}_t}\left[h(\mathbf{x}_t, \bar{u}_t)\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]\right] = \mathbb{E}^{\mathbf{x}_t}[h(\mathbf{x}_t, \bar{u}_t)]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t] \qquad \text{(since } \boldsymbol{\gamma}_t \perp \mathbf{x}_t, \bar{u}_t\text{)} \tag{10}$$

$$\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{x}_t^T g(\mathbf{u}_t)]\right] = \mathbb{E}^{\mathbf{x}_t}\left[\mathbf{x}_t^T \mathbb{E}^{\Delta_t}[g(\mathbf{u}_t)]\right] \qquad (\Delta_t \text{ conditioned on } \mathbf{x}_t) \tag{11}$$

$$\mathbb{E}^{\Delta_t}[\bar{u}_{t+1}] = \bar{u}_{t+1} \tag{12}$$

where $h(\mathbf{x}_t, \bar{u}_t)$ is some function of $\mathbf{x}_t$ and $\bar{u}_t$, and similarly for $g()$. Eq. 12 follows from $\bar{u}_{t+1}$ being an average over the randomness represented by Δ_t. We can now expand Var_t:

$$\begin{aligned}\mathrm{Var}_t &= \mathbb{E}^{\tilde{\mathbf{x}}_t}[\tilde{\mathbf{x}}_t^2] - (\mathbb{E}^{\tilde{\mathbf{x}}_t}[\tilde{\mathbf{x}}_t])^2\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\tilde{\mathbf{x}}_t^2]] - (\mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\tilde{\mathbf{x}}_t]])^2 && \text{(using eq. 6)}\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\mathbf{x}_t^2 + \boldsymbol{\delta}_t^2 + 2\mathbf{x}_t^T\boldsymbol{\delta}_t]] - (\mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\mathbf{x}_t + \boldsymbol{\delta}_t]])^2 \end{aligned}\tag{13}$$

We expand each term:

$$\begin{aligned}&\mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\mathbf{x}_t^2 + \boldsymbol{\delta}_t^2 + 2\mathbf{x}_t^T\boldsymbol{\delta}_t]]\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^2 + \mathbb{E}^{V_t}[\boldsymbol{\delta}_t^2] + 2\mathbf{x}_t^T\mathbb{E}^{V_t}[\boldsymbol{\delta}_t]]\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^2] + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t^2\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t^2]] + 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T\bar{u}_t\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]]\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^2] + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t^2]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t^2] + 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T\bar{u}_t]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]\end{aligned}$$

$$\begin{aligned}&(\mathbb{E}^{\mathbf{x}_t}[\mathbb{E}^{V_t}[\mathbf{x}_t + \boldsymbol{\delta}_t]])^2\\ &= (\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t + \mathbb{E}^{V_t}[\boldsymbol{\delta}_t]])^2\\ &= (\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t + \bar{u}_t\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]])^2\\ &= (\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t] + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t])^2\\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t]^2 + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]^2\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]^2 + 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{Var}_{\,t} &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^2] + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t^2]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t^2] + 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T\bar{u}_t]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t] \\ &- \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t]^2 - \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]^2\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]^2 - 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t]\end{aligned} \tag{14}$$

Following similar procedures, we can write $\mathrm{Var}_{\,t+1}$ as

$$\begin{aligned}\mathrm{Var}_{\,t+1} &= \mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^2] + \mathbb{E}^{\mathbf{x}_{t+1}}[\bar{u}_{t+1}^2]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}^2] \\ &+ 2\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^T\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &- \mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}]^2 - \mathbb{E}^{\mathbf{x}_{t+1}}[\bar{u}_{t+1}]^2\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}]^2 \\ &- 2\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^T]\mathbb{E}^{\mathbf{x}_{t+1}}[\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}]\end{aligned} \tag{15}$$

We tackle each term separately:

$$\begin{aligned}\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^2] &= \mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[(\mathbf{x}_t+\mathbf{u}_t)^2]\right] && \text{(using eq. 9, 2 main text)} \\ &= \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^2] + \mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^2]\right] + 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbf{x}_t^T\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right] && \text{(using eq. 11)}\end{aligned}$$

$$\begin{aligned}&2\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^T\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &= 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[(\mathbf{x}_t+\mathbf{u}_t)^T\bar{u}_{t+1}]\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] && \text{(using eq. 9, 2 main text)} \\ &= 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{x}_t^T\bar{u}_{t+1}]\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &+ 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^T\bar{u}_{t+1}]\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &= 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbf{x}_t^T\bar{u}_{t+1}\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] && \text{(using eq. 11 and 12)} \\ &+ 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^T\bar{u}_{t+1}]\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}]\end{aligned}$$

$$\begin{aligned}-\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}]^2 &= -\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{x}_t+\mathbf{u}_t]\right]^2 \\ &= -\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t]^2 - \mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right]^2 - 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{x_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right]\end{aligned}$$

$$\begin{aligned}&-2\mathbb{E}^{\mathbf{x}_{t+1}}[\mathbf{x}_{t+1}^T]\mathbb{E}^{\mathbf{x}_{t+1}}[\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &= -2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[(\mathbf{x}_t+\mathbf{u}_t)^T]\right]\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\bar{u}_{t+1}]\right]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] \\ &= -2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^T]\right]\mathbb{E}^{\mathbf{x}_t}[\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}] - 2\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{\mathbf{x}_t}[\bar{u}_{t+1}]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}]\end{aligned}$$

Assuming stationarity for $\boldsymbol{\gamma}_t$, and thus $\bar{\boldsymbol{\gamma}} := \mathbb{E}^{V_t}[\boldsymbol{\gamma}_t] = \mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}]$, we have

$$\begin{aligned}\mathrm{Var}_{\,t+1} - \mathrm{Var}_{\,t} &= 2\left\{\mathbb{E}^{\mathbf{x}_t}\left[\mathbf{x}_t^T\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right] - \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right]\right\} \\ &- 2\left\{\mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T(\bar{u}_t-\bar{u}_{t+1})\bar{\boldsymbol{\gamma}}] - \mathbb{E}^{\mathbf{x}_t}[\mathbf{x}_t^T]\mathbb{E}^{\mathbf{x}_t}[(\bar{u}_t-\bar{u}_{t+1})\bar{\boldsymbol{\gamma}}]\right\} \\ &+ \Big\{\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^2]\right] + \mathbb{E}^{\mathbf{x}_{t+1}}[\bar{u}_{t+1}^2]\mathbb{E}^{V_{t+1}}[\boldsymbol{\gamma}_{t+1}^2] - \mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t]\right]^2 \\ &- \mathbb{E}^{\mathbf{x}_t}[\bar{u}_{t+1}]^2\bar{\boldsymbol{\gamma}}^2 - \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t^2]\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t^2] + \mathbb{E}^{\mathbf{x}_t}[\bar{u}_t]^2\mathbb{E}^{V_t}[\boldsymbol{\gamma}_t^2] \\ &+2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^T\bar{u}_{t+1}]\right]\bar{\boldsymbol{\gamma}} - 2\mathbb{E}^{\mathbf{x}_t}\left[\mathbb{E}^{\Delta_t}[\mathbf{u}_t^T]\right]\mathbb{E}^{\mathbf{x}_t}[\bar{u}_{t+1}]\bar{\boldsymbol{\gamma}}\Big\} \\ &= 2cov(\mathbf{x}_t, \mathbb{E}^{\Delta_t}[\mathbf{u}_t]) + \mathcal{O}(\eta_t\xi_t) + \mathcal{O}(\eta_t^2\rho_t^2) + \mathcal{O}^*\end{aligned}$$

where $\xi_t = ||\boldsymbol{g}_t|| - ||\boldsymbol{g}_{t+1}||$ and $\mathcal{O}^*$ are higher order terms. In the last step we use the fact that $||\boldsymbol{g}_t|| = \mathcal{O}(\rho_t)$ (lemma A.2) and thus $||\mathbf{u}_t|| = \eta_t||\nabla f(\mathbf{x}_t)||$ and $\bar{u}_t$ are both $\mathcal{O}(\eta_t\rho_t)$. Notice that $cov(\boldsymbol{v}_1, \boldsymbol{v}_2) := \mathbb{E}[\boldsymbol{v}_1^T\boldsymbol{v}_2] - \mathbb{E}[\boldsymbol{v}_1^T]\mathbb{E}[\boldsymbol{v}_2]$ uses inner product. Thus,

$$\mathrm{Var}_{\,t+1} = \mathrm{Var}_{\,t} - 2\eta_t cov(\mathbf{x}_t, \mathbb{E}^{\Delta_t}[\boldsymbol{g}_t]) + \mathcal{O}(\eta_t\xi_t) + \mathcal{O}(\eta_t^2\rho_t^2) + \mathcal{O}^* \tag{16}$$

□

Theorem 2 (SGD under VAP, bounded variance) Assuming $f(\mathbf{x})$, $\breve{\eta}_t$, and v_t similar to theorem 1, and $f(\mathbf{x})$ has bounded and invertible Hessian, Ω^* defined at optimal point $\mathbf{x}^*$. Let $\mathrm{Var}_{\,t} := \mathbb{E}[\breve{\mathbf{x}}_t^2] - \mathbb{E}[\breve{\mathbf{x}}_t]^2$ ($\mathrm{Var}_{\,t}$ is the sum of component-wise variance[1]), and $\breve{\boldsymbol{g}}_t = \nabla f_t(\breve{\mathbf{x}}_t)$ is the gradient, then:

$$\mathrm{Var}_{\,t+1} = \mathrm{Var}_{\,t} - 2cov(\hat{\mathbf{x}}_t, \mathbb{E}^{\Delta_t}[\breve{\boldsymbol{g}}_t]) + \mathcal{O}(\delta_t) + \mathcal{O}(\breve{\eta}_t^2\rho_t^2) + \mathcal{O}_{\delta_t}^*$$

near the optima $\mathbf{x}^*$. The covariance $cov(\boldsymbol{v}_1, \boldsymbol{v}_2) := \mathbb{E}[\boldsymbol{v}_1^T\boldsymbol{v}_2] - \mathbb{E}[\boldsymbol{v}_1^T]\mathbb{E}[\boldsymbol{v}_2]$ uses inner product. $\delta_t = ||\boldsymbol{\delta}_t||_\infty$ and $\boldsymbol{\delta}_t = \breve{\mathbf{x}}_t - \hat{\mathbf{x}}_t$. $\rho_t = ||\breve{\mathbf{x}}_t - \mathbf{x}^*||$. Δ_t is a random variable capturing the randomness of update $\hat{\mathbf{u}}_t = -\eta_t\breve{\boldsymbol{g}}_t$ conditioned on $\hat{\mathbf{x}}_t$.

[1] $\mathrm{Var}_{\,t} = \sum_{i=1}^d \mathbb{E}[\breve{x}_{ti}^2] - \mathbb{E}[\breve{x}_{ti}]^2$

Proof. The proof is similar to the proof of Theorem 6. Starting off with $\breve{\mathbf{x}}_t = \hat{\mathbf{x}}_t + \boldsymbol{\delta}_t$, we define V_t, Δ_t analogously. We have

$$\begin{aligned} \mathrm{Var}_{\,t} &= \mathbb{E}^{\hat{\mathbf{x}}_t}[\hat{\mathbf{x}}_t^2] + \mathbb{E}^{\hat{\mathbf{x}}_t}[\mathbb{E}^{V_t}[\boldsymbol{\delta}_t^2]] + 2\mathbb{E}^{\hat{\mathbf{x}}_t}[\hat{\mathbf{x}}_t^T \mathbb{E}^{V_t}[\boldsymbol{\delta}_t]] \\ &\quad - \mathbb{E}^{\hat{\mathbf{x}}_t}[\hat{\mathbf{x}}_t]^2 - \mathbb{E}^{\hat{\mathbf{x}}_t}[\mathbb{E}^{V_t}[\boldsymbol{\delta}_t^2]] - 2\mathbb{E}^{\hat{\mathbf{x}}_t}[\hat{\mathbf{x}}_t]\mathbb{E}^{\hat{\mathbf{x}}_t^T}[\mathbb{E}^{V_t}[\boldsymbol{\delta}_t]] \end{aligned}$$

Similar algebra as in Theorem 6 leads to

$$\begin{aligned} \mathrm{Var}_{\,t+1} - \mathrm{Var}_{\,t} &= 2cov(\hat{\mathbf{x}}_t, \mathbb{E}^{\Delta_t}[\hat{\mathbf{u}}_t]) + 2cov(\hat{\mathbf{x}}_t, \mathbb{E}^{V_t}[\boldsymbol{\delta}_t] - \mathbb{E}^{\Delta_t}[\mathbb{E}^{V_{t+1}}[\boldsymbol{\delta}_{t+1}]]) \\ &\quad + \mathcal{O}(\delta_t^2) + \mathcal{O}(\breve{\eta}_t^2 \rho_t^2) + \mathcal{O}(\breve{\eta}_t \delta_t) + \mathcal{O}^* \\ &= -2cov(\hat{\mathbf{x}}_t, \mathbb{E}^{\Delta_t}[\breve{\boldsymbol{g}}_t]) + \mathcal{O}(\delta_t) + \mathcal{O}(\breve{\eta}_t^2 \rho_t^2) + \mathcal{O}^*_{\delta_t} \end{aligned}$$

where $\delta_t = ||\boldsymbol{\delta}_t||_\infty$. This is the desired result in the theorem statement. □