Theorem. *Let* v_i*'s be independent Rademacher variables and* $f : \mathbb{R}^d \to \mathbb{R}$ *be* L*-smooth. Then, we have the following bounds for the control variate forward gradient* $\boldsymbol{h}_{\mathbf{v},\epsilon}(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^d$*:*

$$\|\mathbb{E}[\boldsymbol{h}_{\mathbf{v},\epsilon}(\boldsymbol{\theta})] - \nabla f(\boldsymbol{\theta})\| \leq \frac{\epsilon L d^{3/2}}{2}, \tag{1}$$

and

$$\mathbb{E}[\|\boldsymbol{h}_{\mathbf{v},\epsilon}(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|^2] \leq \frac{\epsilon^2 L^2 d^3}{4} + 2(d-1)\frac{\epsilon L d^{3/2}}{2}\left\|\nabla f(\boldsymbol{\theta}) - \nabla \hat{f}(\boldsymbol{\theta})\right\| + (d-1)\left\|\nabla f(\boldsymbol{\theta}) - \nabla \hat{f}(\boldsymbol{\theta})\right\|^2 \tag{2}$$

Proof. Since f is L-smooth, we have

$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})| \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \quad (\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d). \tag{3}$$

This inequality with $\boldsymbol{\theta} + \epsilon\mathbf{v}$ and $\boldsymbol{\theta}$ yields

$$\begin{aligned} |f(\boldsymbol{\theta} + \epsilon\mathbf{v}) - f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}) \cdot \epsilon\mathbf{v}| &\leq \frac{L}{2}\|\epsilon\mathbf{v}\|^2 & (4)\\ &= \frac{\epsilon^2 L d}{2}. & (5)\end{aligned}$$

We denote $D_{\mathbf{v},\epsilon}(\boldsymbol{\theta}) := \frac{f(\boldsymbol{\theta}+\epsilon\mathbf{v}) - f(\boldsymbol{\theta})}{\epsilon} - \nabla f(\boldsymbol{\theta}) \cdot \mathbf{v}$. Note that $|D_{\mathbf{v},\epsilon}(\boldsymbol{\theta})| \leq \epsilon L d/2$. Using this inequality, we can show inequality (1) as follows:

$$\begin{aligned} \|\mathbb{E}[\boldsymbol{h}_{\mathbf{v},\epsilon}(\boldsymbol{\theta})] - \nabla f(\boldsymbol{\theta})\| &= \|\mathbb{E}[\boldsymbol{g}_{\mathbf{v},\epsilon}(\boldsymbol{\theta}) - \hat{\boldsymbol{g}_{\mathbf{v}}}(\boldsymbol{\theta}) + \mathbb{E}[\hat{\boldsymbol{g}_{\mathbf{v}}}(\boldsymbol{\theta})]] - \mathbb{E}[\boldsymbol{g}_{\mathbf{v}}(\boldsymbol{\theta})]\| & (6)\\ &= \|\mathbb{E}[\boldsymbol{g}_{\mathbf{v},\epsilon}(\boldsymbol{\theta}) - \boldsymbol{g}_{\mathbf{v}}(\boldsymbol{\theta})]\| & (7)\\ &\leq \mathbb{E}[\|\boldsymbol{g}_{\mathbf{v},\epsilon}(\boldsymbol{\theta}) - \boldsymbol{g}_{\mathbf{v}}(\boldsymbol{\theta})\|] & (8)\\ &= \mathbb{E}[\left\|\frac{f(\boldsymbol{\theta} + \epsilon\mathbf{v}) - f(\boldsymbol{\theta})}{\epsilon}\mathbf{v} - (\nabla f(\boldsymbol{\theta}) \cdot \mathbf{v})\mathbf{v}\right\|] & (9)\\ &= \mathbb{E}[|D_{\mathbf{v},\epsilon}(\boldsymbol{\theta})|\,\|\mathbf{v}\|] & (10)\\ &\leq \frac{\epsilon L d}{2} d^{1/2} & (11)\\ &= \frac{\epsilon L d^{3/2}}{2}. & (12)\end{aligned}$$

Next, we show inequality (2). To simplify the notation, we omit the argument $\boldsymbol{\theta}$. Then, we have

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{h}_{\mathbf{v},\epsilon} - \nabla f\|^2] &= \mathbb{E}[\|(\boldsymbol{g}_{\mathbf{v},\epsilon} - \boldsymbol{g}_{\mathbf{v}}) + (\boldsymbol{h}_{\mathbf{v}} - \nabla f)\|^2] && (13)\\
&= \mathbb{E}[\|D_{\mathbf{v},\epsilon}\mathbf{v} + (\boldsymbol{h}_{\mathbf{v}} - \nabla f)\|^2] && (14)\\
&= \mathbb{E}[\|D_{\mathbf{v},\epsilon}\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}\cdot(\boldsymbol{h}_{\mathbf{v}} - \nabla f)] + \mathbb{E}[\|\boldsymbol{h}_{\mathbf{v}} - \nabla f\|^2] && (15)\\
&= \mathbb{E}[{D_{\mathbf{v},\epsilon}}^2\|\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}\cdot(\boldsymbol{g}_{\mathbf{v}} - \hat{\boldsymbol{g}}_{\mathbf{v}} + \nabla\hat{f} - \nabla f)] + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (16)\\
&= \mathbb{E}[{D_{\mathbf{v},\epsilon}}^2\|\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}\cdot(\boldsymbol{g}_{\mathbf{v}} - \hat{\boldsymbol{g}}_{\mathbf{v}})] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}]\cdot(\nabla\hat{f} - \nabla f) + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (17)\\
&= \mathbb{E}[{D_{\mathbf{v},\epsilon}}^2\|\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}\cdot((\nabla f - \nabla\hat{f})\cdot\mathbf{v})\mathbf{v}] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}]\cdot(\nabla\hat{f} - \nabla f) + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (18)\\
&= \mathbb{E}[{D_{\mathbf{v},\epsilon}}^2\|\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}((\nabla f - \nabla\hat{f})\cdot\mathbf{v})\|\mathbf{v}\|^2] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}]\cdot(\nabla\hat{f} - \nabla f) + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (19)\\
&= d\mathbb{E}[{D_{\mathbf{v},\epsilon}}^2] + 2d\mathbb{E}[D_{\mathbf{v},\epsilon}((\nabla f - \nabla\hat{f})\cdot\mathbf{v})] + 2\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}]\cdot(\nabla\hat{f} - \nabla f) + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (20)\\
&= d\mathbb{E}[{D_{\mathbf{v},\epsilon}}^2] + 2(d-1)\mathbb{E}[D_{\mathbf{v},\epsilon}\mathbf{v}]\cdot(\nabla f - \nabla\hat{f}) + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (21)\\
&\le d(\frac{\epsilon L d}{2})^2 + 2(d-1)\frac{\epsilon L d^{3/2}}{2}\left\|\nabla f - \nabla\hat{f}\right\| + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2 && (22)\\
&= \frac{\epsilon^2 L^2 d^3}{4} + 2(d-1)\frac{\epsilon L d^{3/2}}{2}\left\|\nabla f - \nabla\hat{f}\right\| + (d-1)\left\|\nabla f - \nabla\hat{f}\right\|^2. && (23)
\end{aligned}
$$

□