SGD:

$$g_t = \hat{g}_t^{\text{SGD}} = \frac{\partial}{\partial \theta}\mathbb{E}[L(\theta)]\bigg|_{\theta=\theta_t} \tag{1}$$

SGD with momentum:

$$\begin{aligned}
\hat{g}_t^{\text{Mom}} &= \beta \hat{g}_{t-1}^{\text{Mom}} + (1-\beta) g_t &(2)\\
&= \sum_{t=1}^{T} (1-\beta)\beta^{T-t} g_t &(3)\\
\frac{\partial \hat{L}^{\text{Mom}}}{\partial \theta}\bigg|_{\theta=\theta_t} &= \hat{g}_t^{\text{Mom}} &(4)\\
\hat{L}^{\text{Mom}} &= \int \sum_{t=1}^{T} (1-\beta)\beta^{T-t} g_t d\theta &(5)\\
&= (1-\beta)\sum_{t=1}^{T} \beta^{T-t} \int g_t d\theta &(6)\\
&= (1-\beta)\sum_{t=1}^{T} \beta^{T-t} \mathbb{E}[L(\theta_t)] &(7)
\end{aligned}$$

SGD with momentum is minimizing an exponential average of the loss of the model. So it is stabilizing learning.