How to Improve SVRG?

- For SVRG ,
  $\mathbf{v} = \nabla f_{\mathcal{I}_b}(\mathbf{x}) - \nabla F_{\mathcal{I}_b}(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) = \mathbf{g}^{(1)} + \mathbf{g}^{(0)}$,
  $\mathbf{g}^{(1)} = \nabla f_{\mathcal{I}_b}(\mathbf{x}) - \nabla f_{\mathcal{I}_b}(\mathbf{x}_0), \mathbf{g}^{(0)} = \nabla F(\mathbf{x}_0)$.
- Only use two reference points $(\mathbf{x}, \mathbf{x}_0)$ and two reference gradients $(\mathbf{g}^{(1)}, \mathbf{g}^{(0)})$ .....
- Using more than two reference points and reference gradients!
- $\mathbf{v} = \mathbf{g}^{(K)} + \cdots + \mathbf{g}^{(1)} + \mathbf{g}^{(0)}$,
  $\mathbf{g}^{(l)} = \nabla f_{\mathcal{I}_l}(\mathbf{x}^{(l)}) - \nabla f_{\mathcal{I}_l}(\mathbf{x}^{(l-1)}), 1 \leq l \leq K$,
  $\mathbf{g}^{(0)} = \nabla F(\mathbf{x}^{(0)})$.
- $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta \mathbf{v}$.

Stochastic Nested Variance Reduced Gradient Descent(SNVRG)[1]

Algorithm 1 SNVRG (Outer loop)

Input: $\mathbf{z}_0, \eta, B, S, K, \{B_l\}, \{T_l\}$.
for $s = 1, \ldots, S$ **do**
  $[\mathbf{y}_s, \mathbf{z}_s] \leftarrow$ SNVRG-Epoch $(\mathbf{z}_{s-1}, \eta, B, K, \{B_l\}, \{T_l\})$.
end for
Output: Uniformly choose $\mathbf{y}_{\text{out}}$ from $\{\mathbf{y}_s\}$.

Algorithm 2 SNVRG-Epoch

Input: $\mathbf{x}_0, \eta, B, K, \{B_l\}, \{T_l\}$.
Randomly pick $\mathcal{I}_B$ with size B.
$\mathbf{g}_0^{(0)} \leftarrow \nabla f_{\mathcal{I}_B}(\mathbf{x}_0)$, $\mathbf{x}_0^{(0)} \leftarrow \mathbf{x}_0$
$\mathbf{g}_0^{(l)} \leftarrow 0, \mathbf{x}_0^{(l)} \leftarrow \mathbf{x}_0, l \in [K]$
$\mathbf{v}_0 \leftarrow \sum_{l=0}^{K} \mathbf{g}_0^{(l)}, \mathbf{x}_1 \leftarrow \mathbf{x}_0 - \eta \cdot \mathbf{v}_0$
for $t = 1, \ldots, \prod_{l=1}^{K} T_l - 1$ **do**
  Update $\{\mathbf{x}_t^{(l)}\}$ and $\{\mathbf{g}_t^{(l)}\}$
  $\mathbf{v}_t \leftarrow \sum_{l=1}^{K} \mathbf{g}_t^{(l)} + \mathbf{g}_t^{(0)}$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \cdot \mathbf{v}_t$
end for
Output: $[\mathbf{x}_{\text{out}}, \mathbf{x}_{\prod_{l=1}^{K} T_l}]$, $\mathbf{x}_{\text{out}}$ from $\{\mathbf{x}_{0 \le t < \prod_{l=1}^{K} T_l}\}$,

[1] Zhou, Dongruo, et al. "Stochastic nested variance reduced gradient descent for nonconvex optimization." Advances in Neural Information Processing Systems. 2018.

Update rules

- Update parameters: batch size parameters $\{B_l\}$, loop length parameters $\{T_l\}$.
- Let r be the smallest number where t can be divided by $\prod_{l=r+1}^{K} T_l$.
- Update rules for reference points $\{\mathbf{x}_t^{(l)}\}$:
  - $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(r-1)}$ remain the same as $\mathbf{x}_{t-1}^{(1)}, \ldots, \mathbf{x}_{t-1}^{(r-1)}$
  - Set $\mathbf{x}_t^{(r)}, \ldots, \mathbf{x}_t^{(K)} \leftarrow \mathbf{x}_t$.
- Update rules for reference gradients $\{\mathbf{g}_t^{(l)}\}$:
  - We do not need to upgrade reference gradients unless they have changed!
  - $\mathbf{g}_t^{(1)}, \ldots, \mathbf{g}_t^{(r-1)}$ remain the same as $\mathbf{g}_{t-1}^{(1)}, \ldots, \mathbf{g}_{t-1}^{(r-1)}$
  - For $r \leq l \leq K$,
    randomly pick up $\mathcal{I}$ with size B_l, set $\mathbf{g}_t^{(l)} \leftarrow \nabla f_{\mathcal{I}}(\mathbf{x}_t^{(l)}) - \nabla f_{\mathcal{I}}(\mathbf{x}_t^{(l-1)})$.

Step 0

$K = 2, T_1 = 2, T_2 = 3$ as an example.

Reference points:

$$\mathbf{x}_0^{(0)} \leftarrow \mathbf{x}_0, \mathbf{x}_0^{(1)} \leftarrow \mathbf{x}_0, \mathbf{x}_0^{(2)} \leftarrow \mathbf{x}_0,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_0^{(0)} &\leftarrow \nabla f_{\mathcal{I}_0}(\mathbf{x}_0^{(0)}), \\
\mathbf{g}_0^{(1)} &\leftarrow \nabla f_{\mathcal{I}_1}(\mathbf{x}_0^{(1)}) - \nabla f_{\mathcal{I}_1}(\mathbf{x}_0^{(0)}), \\
\mathbf{g}_0^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_0^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_0^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_1 \leftarrow \mathbf{x}_0 - \eta(\mathbf{g}_0^{(0)} + \mathbf{g}_0^{(1)} + \mathbf{g}_0^{(2)}).$$

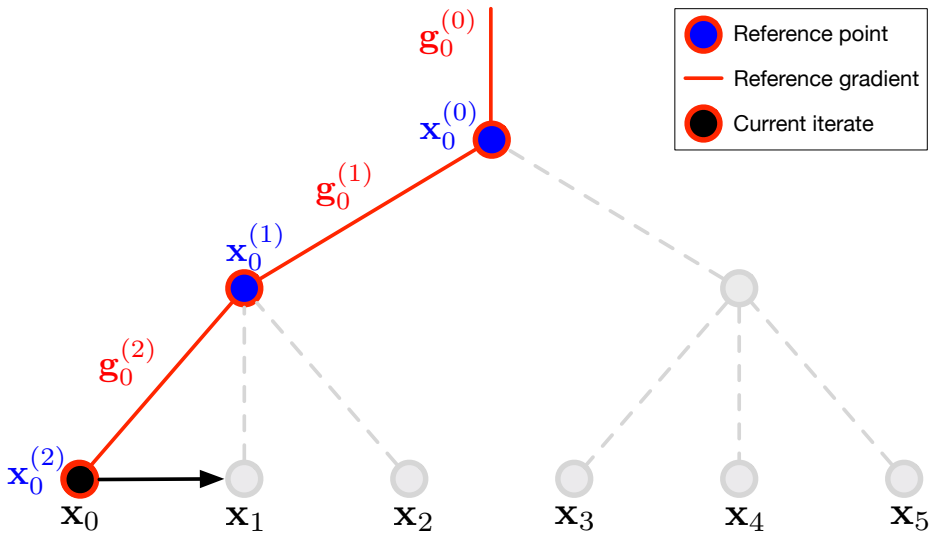

Figure: Iterate $t = 0$.

Step 1

Reference points:

$$\mathbf{x}_1^{(0)} \leftarrow \mathbf{x}_0^{(0)}, \mathbf{x}_1^{(1)} \leftarrow \mathbf{x}_0^{(1)}, \mathbf{x}_1^{(2)} \leftarrow \mathbf{x}_1,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_1^{(0)} &\leftarrow \mathbf{g}_0^{(0)}, \\
\mathbf{g}_1^{(1)} &\leftarrow \mathbf{g}_0^{(1)}, \\
\mathbf{g}_1^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_1^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_1^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_2 \leftarrow \mathbf{x}_1 - \eta(\mathbf{g}_1^{(0)} + \mathbf{g}_1^{(1)} + \mathbf{g}_1^{(2)}).$$

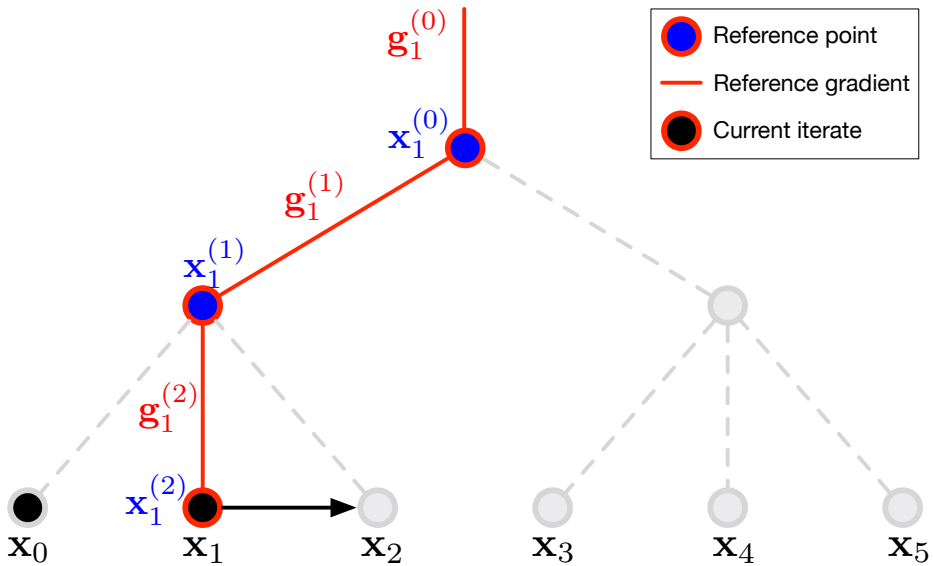

Figure: Iterate $t = 1$.

Step 2

Reference points:

$$\mathbf{x}_2^{(0)} \leftarrow \mathbf{x}_1^{(0)}, \mathbf{x}_2^{(1)} \leftarrow \mathbf{x}_1^{(1)}, \mathbf{x}_2^{(2)} \leftarrow \mathbf{x}_2,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_2^{(0)} &\leftarrow \mathbf{g}_1^{(0)}, \\
\mathbf{g}_2^{(1)} &\leftarrow \mathbf{g}_1^{(1)}, \\
\mathbf{g}_2^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_2^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_2^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_3 \leftarrow \mathbf{x}_2 - \eta(\mathbf{g}_2^{(0)} + \mathbf{g}_2^{(1)} + \mathbf{g}_2^{(2)}).$$

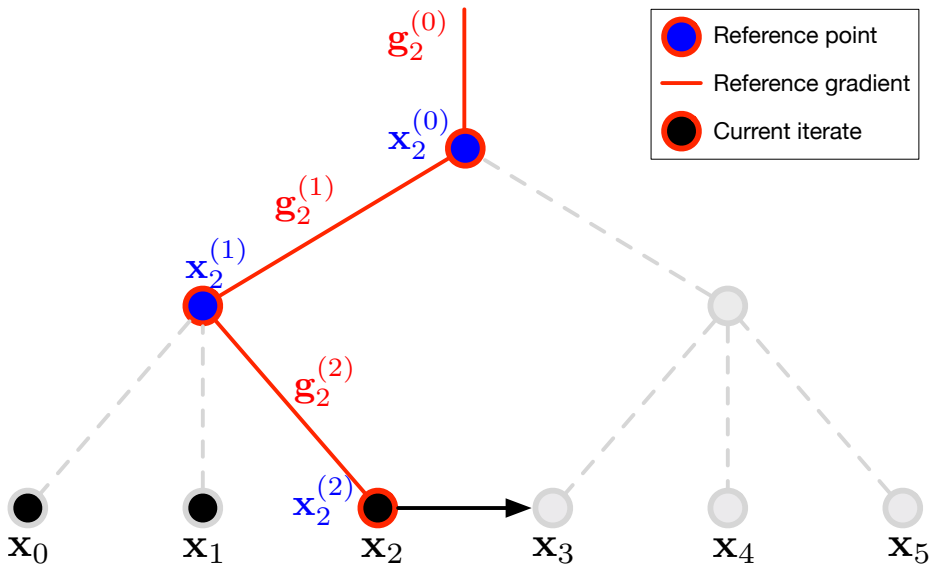

Figure: Iterate $t = 2$.

Step 3

Reference points:

$$\mathbf{x}_3^{(0)} \leftarrow \mathbf{x}_2^{(0)}, \mathbf{x}_3^{(1)} \leftarrow \mathbf{x}_3, \mathbf{x}_3^{(2)} \leftarrow \mathbf{x}_3,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_3^{(0)} &\leftarrow \mathbf{g}_2^{(0)}, \\
\mathbf{g}_3^{(1)} &\leftarrow \nabla f_{\mathcal{I}_1}(\mathbf{x}_3^{(1)}) - \nabla f_{\mathcal{I}_1}(\mathbf{x}_3^{(0)}), \\
\mathbf{g}_3^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_3^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_3^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_4 \leftarrow \mathbf{x}_3 - \eta(\mathbf{g}_3^{(0)} + \mathbf{g}_3^{(1)} + \mathbf{g}_3^{(2)}).$$

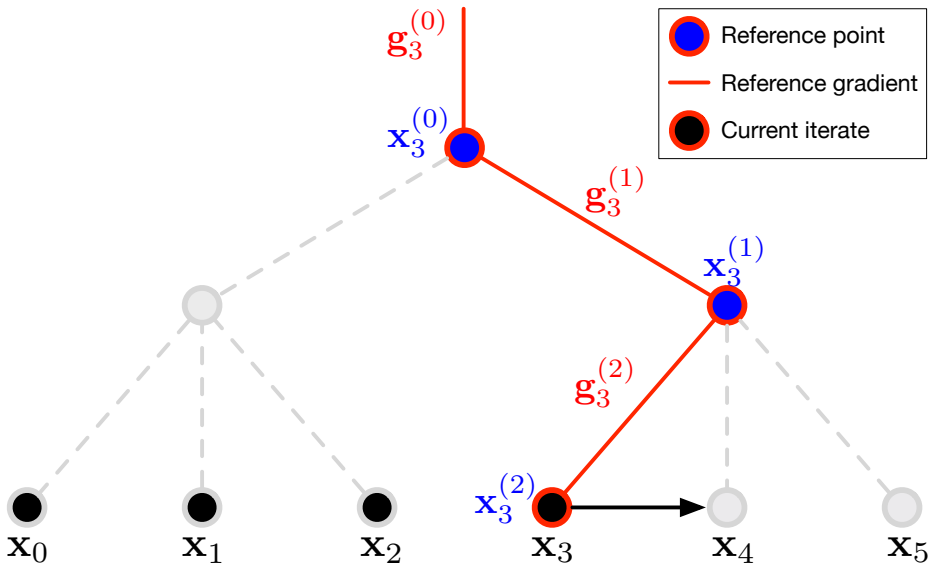

Figure: Iterate $t = 3$.

Step 4

Reference points:

$$\mathbf{x}_4^{(0)} \leftarrow \mathbf{x}_2^{(0)}, \mathbf{x}_3^{(1)} \leftarrow \mathbf{x}_3, \mathbf{x}_3^{(2)} \leftarrow \mathbf{x}_3,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_3^{(0)} &\leftarrow \mathbf{g}_2^{(0)}, \\
\mathbf{g}_3^{(1)} &\leftarrow \nabla f_{\mathcal{I}_1}(\mathbf{x}_3^{(1)}) - \nabla f_{\mathcal{I}_1}(\mathbf{x}_3^{(0)}), \\
\mathbf{g}_3^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_3^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_3^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_4 \leftarrow \mathbf{x}_3 - \eta(\mathbf{g}_3^{(0)} + \mathbf{g}_3^{(1)} + \mathbf{g}_3^{(2)}).$$

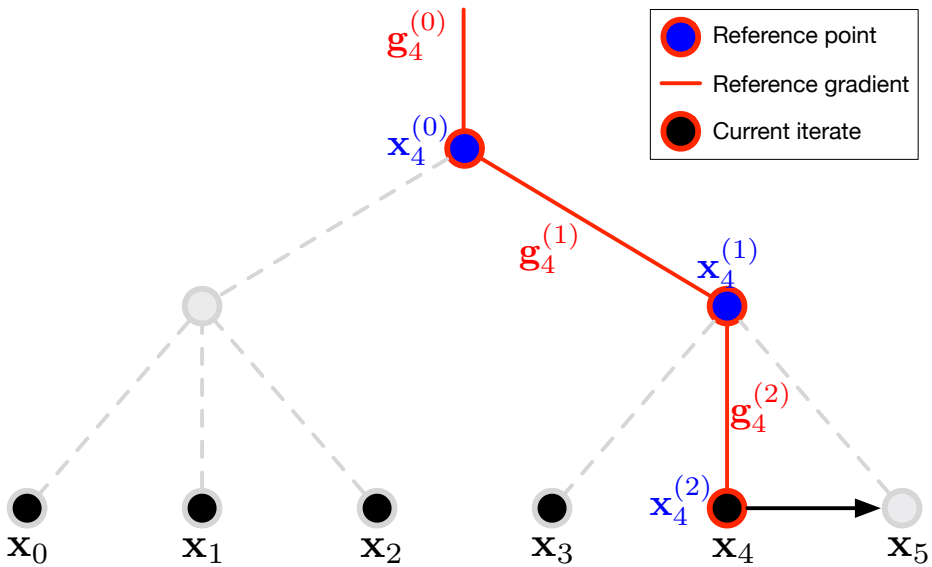

Figure: Iterate $t = 4$.

Step 5

Reference points:

$$\mathbf{x}_5^{(0)} \leftarrow \mathbf{x}_4^{(0)}, \mathbf{x}_5^{(1)} \leftarrow \mathbf{x}_4^{(1)}, \mathbf{x}_5^{(2)} \leftarrow \mathbf{x}_5,$$

Reference gradients:

$$\begin{aligned}
\mathbf{g}_5^{(0)} &\leftarrow \mathbf{g}_4^{(0)}, \\
\mathbf{g}_5^{(1)} &\leftarrow \mathbf{g}_4^{(1)}, \\
\mathbf{g}_5^{(2)} &\leftarrow \nabla f_{\mathcal{I}_2}(\mathbf{x}_5^{(2)}) - \nabla f_{\mathcal{I}_2}(\mathbf{x}_5^{(1)}),
\end{aligned}$$

Updating rule:

$$\mathbf{x}_6 \leftarrow \mathbf{x}_5 - \eta(\mathbf{g}_5^{(0)} + \mathbf{g}_5^{(1)} + \mathbf{g}_5^{(2)}).$$

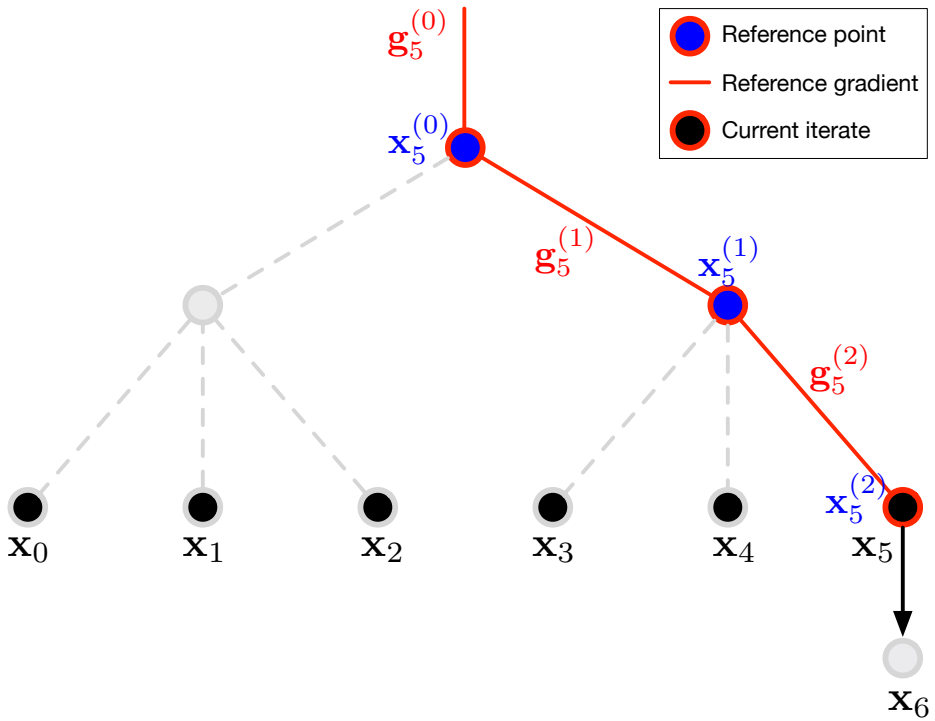

Figure: Iterate $t = 5$.