AdaGrad variances for parameter $x \in \mathbb{R}^d$ and gradient $g \in \mathbb{R}^d$ under coordinate noise sparsity β

Coordinate $(k = 1)$

Gradient, State, Update

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ \vdots \\ g_{d-2} \\ g_{d-1} \\ g_d \end{bmatrix} \longrightarrow \begin{bmatrix} v_1 + g_1^2 \\ v_2 + g_2^2 \\ v_3 + g_3^2 \\ v_4 + g_4^2 \\ v_5 + g_5^2 \\ v_6 + g_6^2 \\ \vdots \\ v_{d-2} + g_{d-2}^2 \\ v_{d-1} + g_{d-1}^2 \\ v_d + g_d^2 \end{bmatrix} \in \mathbb{R}^d \qquad \begin{pmatrix} x_1 - \eta \frac{g_1}{\sqrt{v_1}+\epsilon} \\ x_2 - \eta \frac{g_2}{\sqrt{v_2}+\epsilon} \\ x_3 - \eta \frac{g_3}{\sqrt{v_3}+\epsilon} \\ x_4 - \eta \frac{g_4}{\sqrt{v_4}+\epsilon} \\ x_5 - \eta \frac{g_5}{\sqrt{v_5}+\epsilon} \\ x_6 - \eta \frac{g_6}{\sqrt{v_6}+\epsilon} \\ \vdots \\ x_{d-2} - \eta \frac{g_{d-2}}{\sqrt{v_{d-2}}+\epsilon} \\ x_{d-1} - \eta \frac{g_{d-1}}{\sqrt{v_{d-1}}+\epsilon} \\ x_d - \eta \frac{g_d}{\sqrt{v_d}+\epsilon} \end{pmatrix}$$

Complexity

$$\tilde{O}\left(\frac{d^{1.5+\beta}}{\sqrt{T}} + \frac{d^{2.5}}{T}\right)$$

Memory

$$O(d)$$

Subset-Norm $(1 < k < d)$

Gradient, State, Update

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ \vdots \\ g_{d-2} \\ g_{d-1} \\ g_d \end{bmatrix} \longrightarrow \begin{bmatrix} v_1 + \|g_{1:3}\|^2 \\ v_2 + \|g_{4:6}\|^2 \\ \vdots \\ v_{d/3} + \|g_{d-2:d}\|^2 \end{bmatrix} \in \mathbb{R}^{d/k} \qquad \begin{pmatrix} x_1 - \eta \frac{g_1}{\sqrt{v_1}+\epsilon} \\ x_2 - \eta \frac{g_2}{\sqrt{v_1}+\epsilon} \\ x_3 - \eta \frac{g_3}{\sqrt{v_1}+\epsilon} \\ x_4 - \eta \frac{g_4}{\sqrt{v_2}+\epsilon} \\ x_5 - \eta \frac{g_5}{\sqrt{v_2}+\epsilon} \\ x_6 - \eta \frac{g_6}{\sqrt{v_2}+\epsilon} \\ \vdots \\ x_{d-2} - \eta \frac{g_{d-2}}{\sqrt{v_{d/k}}+\epsilon} \\ x_{d-1} - \eta \frac{g_{d-1}}{\sqrt{v_{d/k}}+\epsilon} \\ x_d - \eta \frac{g_d}{\sqrt{v_{d/k}}+\epsilon} \end{pmatrix}$$

Complexity

$$\tilde{O}\left(\frac{d^{0.3+1.8\beta}}{\sqrt{T}} + \frac{d^{\beta+1}}{T}\right) \text{ if } \beta \in [0, 2/3]$$

$$\tilde{O}\left(\frac{d^{0.3+1.8\beta}}{\sqrt{T}} + \frac{d^{1.6\beta+0.6}}{T}\right) \text{ if } \beta \in [2/3, 1]$$

Memory

$$O(\sqrt{d}) \text{ when } k = O(\sqrt{d})$$

Norm $(k = d)$

Gradient, State, Update

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ \vdots \\ g_{d-2} \\ g_{d-1} \\ g_d \end{bmatrix} \longrightarrow v + \|g\|^2 \in \mathbb{R} \qquad \begin{pmatrix} x_1 - \eta \frac{g_1}{\sqrt{v}+\epsilon} \\ x_2 - \eta \frac{g_2}{\sqrt{v}+\epsilon} \\ x_3 - \eta \frac{g_3}{\sqrt{v}+\epsilon} \\ x_4 - \eta \frac{g_4}{\sqrt{v}+\epsilon} \\ x_5 - \eta \frac{g_5}{\sqrt{v}+\epsilon} \\ x_6 - \eta \frac{g_6}{\sqrt{v}+\epsilon} \\ \vdots \\ x_{d-2} - \eta \frac{g_{d-2}}{\sqrt{v}+\epsilon} \\ x_{d-1} - \eta \frac{g_{d-1}}{\sqrt{v}+\epsilon} \\ x_d - \eta \frac{g_d}{\sqrt{v}+\epsilon} \end{pmatrix}$$

Complexity

$$\tilde{O}\left(\frac{d^{2.5\beta}}{\sqrt{T}} + \frac{d^{3\beta}}{T}\right)$$

Memory

$$O(1)$$