Supplementary Material

Proof for Theorem 1

We first give a proof for Theorem 1 using the two lemmas. Proofs for the lemmas then follow.

We arrive at Theorem 1 by bounding separately the two terms R_1 and R_2 in the regret decomposition in (11) of main paper for an arbitrary objective function $f \in \mathcal{F}_{\alpha,\beta}$. Note that the consistency/efficiency level p of an algorithm as defined in (8) and (9) (of main paper) is with respect to the worst-case objective function. This implies that when a p-consistent low-dimensional algorithm is employed for CM, the convergence rates along different coordinates may vary, depending on the reduction of f to the specific coordinate. Let $p_k \geq p$ be the convergence rate in the coordinate i_k chosen in the k-th iteration given $\mathbf{x}^{(k-1)}_{-i_k}$. More specifically, the error with respect to the local minimum $\mathbf{x}^*_{i_k,\mathbf{x}^{(k-1)}}$ in the i_k-th coordinate decays as follows.

$$\left(\mathbb{E}[f(x_{i_k,T}, \mathbf{x}^{(k-1)}_{-i_k})] - f(\mathbf{x}^*_{i_k,\mathbf{x}^{(k-1)}})\right) \sim \Theta(T^{-p_k}), \tag{1}$$

where $x_{i_k,T}$ denotes the one-dimensional query point at time T.

We start by bounding R_2.

$$\begin{aligned}
R_2 &\leq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}\left[F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})},\xi_t) - F(\mathbf{x}^*,\xi_t)\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{s=1}^{\tau(\epsilon_k)}\left[F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})},\xi_s) - F(\mathbf{x}^*,\xi_s)\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}\mathbb{E}\left[\sum_{s=1}^{\tau(\epsilon_k)}\left[F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})},\xi_s) - F(\mathbf{x}^*,\xi_s)\right]\middle|\mathbf{x}^{(k-1)}\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}[f(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})}) - f(\mathbf{x}^*)]\mathbb{E}[\tau(\epsilon_k)]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}[f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)]\mathbb{E}[\tau(\epsilon_k)]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}c_2\gamma^k\epsilon_k^{-1/p_k}\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}c_2'\left(\frac{1}{\gamma}\right)^{k\frac{1-p_k}{p_k}}\right]
\end{aligned} \tag{2}$$

where the fourth line follows from Wald's Identity and $c_2, c_2' > 0$ are constants independent of T. To upper bound the expression obtained in (2), we use that the total number of samples taken would be upper bounded by the length of the horizon.

$$\sum_{k=1}^{K}\mathbb{E}\left[\tau(\epsilon_k)\right] \leq T. \tag{3}$$

Therefore, for some constant $c_3 > 0$ and independent of T, we have,

$$\sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{\frac{k}{p_k}} \leq c_3 T. \tag{4}$$

Now using Jensen's inequality, we can write,

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{k\frac{1-p}{p_k}} &\leq \left(\frac{1}{K}\sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{\frac{k}{p_k}}\right)^{1-p}, \\ \implies \sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{k\frac{1-p}{p_k}} &\leq c_3' T^{1-p} K^p. \end{aligned} \tag{5}$$

for some constant $c_3' > 0$. Note that the expression here is similar to the one obtained in (2) and in fact can be used to upper bound R_2.

$$\begin{aligned} R_2 &\leq c_2' \mathbb{E}\left[\sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{k\frac{1-p_k}{p_k}}\right] \\ &\leq c_2' \mathbb{E}\left[\sum_{k=1}^{K} \left(\frac{1}{\gamma}\right)^{k\frac{1-p}{p_k}}\right] \\ &\leq c_2'' \mathbb{E}\left[T^{1-p}K^p\right] \\ &\leq c_2'' T^{1-p} \mathbb{E}\left[K\right]^p \end{aligned}$$

where $c_2'' > 0$ is a constant independent of T and the second step is obtained by noting $p_k \geq p$. Using the result from Lemma 1 and plugging it in the above equation, we can conclude that R_2 is $O(T^{1-p}\log^p T)$. Note that for $p = 1$ this boils down to $O(\log T)$ as required.

We now consider R_1.

$$\begin{aligned} R_1 &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \left[F(\mathbf{x}_t, \xi_t) - F(\mathbf{x}^*_{(i_k, \mathbf{x}^{(k-1)})}, \xi_t)\right]\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{s=1}^{\tau(\epsilon_k)} \left[F(\mathbf{x}_{s+t_{k-1}}, \xi_{s+t_{k-1}}) - F(\mathbf{x}^*_{(i_k, \mathbf{x}^{(k-1)})}, \xi_{s+t_{k-1}})\right]\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\sum_{s=1}^{\tau(\epsilon_k)} \left[F(\mathbf{x}_{s+t_{k-1}}, \xi_{s+t_{k-1}}) - F(\mathbf{x}^*_{(i_k, \mathbf{x}^{(k-1)})}, \xi_{s+t_{k-1}})\right] \middle| \tau(\epsilon_k)\right]\right] \end{aligned}$$

Next we upper bound the above term separately for p-consistent and efficient routines. For p-consistent ($p < 1$) routines, we have,

$$\begin{aligned} R_1 &\leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\sum_{s=1}^{\tau(\epsilon_k)} \frac{c_1}{s^{p_k}} \middle| \tau(\epsilon_k)\right]\right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K} c_1 \mathbb{E}\left[(\tau(\epsilon_k))^{1-p_k}\right]\right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K} c_1 \mathbb{E}\left[\tau(\epsilon_k)\right]^{1-p_k}\right] \qquad (6) \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K} c_1' \epsilon_k^{\frac{p_k-1}{p_k}}\right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K} c_1'' \left(\frac{1}{\gamma}\right)^{k\frac{1-p_k}{p_k}}\right] \qquad (7) \end{aligned}$$

where $c_1, c_1', c_1'' > 0$ are all constants independent of T and (6) follows from Jensen's inequality. Note that the term obtained in (7) is of the same order as the one obtained in (2). Therefore, using the same analysis as in the case of R_2, we can conclude that R_1 is also $O(T^{1-p}\log^p T)$ for p-consistent ($p < 1$) routines. Now for efficient routines we have $p_k = 1$ for all k. Along with the efficiency in leveraging the favorable initial conditions, we have

$$\begin{align}
R_1 &\leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\left(f(\mathbf{x}^{(k-1)} - f(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})})\right)^{\lambda} \sum_{s=1}^{\tau(\epsilon_k)} \frac{b_2}{s}\middle|\tau(\epsilon_k)\right]\right] \notag\\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} b_2'\gamma^{(k-1)\lambda}\mathbb{E}\left[\log(\tau(\epsilon_k))\right]\right] \notag\\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} b_2''(\epsilon_0\gamma^k)^{\lambda}\log\left(\mathbb{E}[\tau(\epsilon_k)]\right)\right] \tag{8}\\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} b_2''\epsilon_k^{\lambda}\log\left(\frac{b_3}{\epsilon_k}\right)\right] \notag\\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} b_2''\left(\epsilon_k^{\lambda}\log\left(\frac{1}{\epsilon_k}\right) + \log(b_3)\epsilon_k^{\lambda}\right)\right] \notag\\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} b_2''\left(\frac{1}{\lambda e} + \log(b_3)\epsilon_0^{\lambda}\right)\right] \tag{9}\\
&\leq b_4\mathbb{E}[K] \tag{10}
\end{align}$$

where $b_2, b_2', b_2'', b_3, b_4 > 0$ are constants independent of T and (8) and (9) are respectively obtained by Jensen's inequality and the fact that $-x^{\lambda}\log(x)$ is uniformly upper bounded by $(\lambda e)^{-1}$ for all $x > 0$ and for all $\lambda > 0$. Using the upper bound on $\mathbb{E}[K]$ given from Lemma 1 leads to the $O(\log T)$ order of R_1 for efficient algorithms. Combining the above bounds on R_1 and R_2, we arrive at the theorem.

Proof of Lemma 1

Note that the first $K-1$ iterations are complete by the end of the horizon of length T. We thus have, for some constants $b_1, b_1' > 0$,

$$\begin{align}
T &\geq \mathbb{E}\left[\sum_{k=1}^{K-1} \mathbb{E}[\tau(\epsilon_k)]\right] \notag\\
&\geq \mathbb{E}\left[\sum_{k=1}^{K-1} b_1\epsilon_k^{-1/p_k}\right] \notag\\
&\geq \mathbb{E}\left[\sum_{k=1}^{K-1} b_1\epsilon_k^{-1}\right] \notag\\
&\geq \mathbb{E}\left[\sum_{k=1}^{K-1} b_1'\gamma^{-k}\right] \notag\\
&\geq \mathbb{E}\left[b_1'\frac{\gamma^{-K} - \gamma^{-1}}{\gamma^{-1} - 1}\right] \tag{11}\\
& \tag{12}
\end{align}$$

Therefore, we have that $\mathbb{E}\left[\left(\frac{1}{\gamma}\right)^K\right] \leq \frac{T(1-\gamma^{-1})}{b_1'} + \gamma^{-1}$. Taking logarithms on both sides and then applying Jensen's inequality, we obtain $\mathbb{E}[K] \leq \log_{\gamma^{-1}}\left(\frac{T(1-\gamma^{-1})}{b_1'} + \gamma^{-1}\right)$ as required.

Proof of Lemma 2

The main idea of the proof revolves around the use of proximal operators which is similar to the convergence analysis in [1]. Specifically, for $f = \psi + \phi$, define

$$\mathcal{D}_\phi(\mathbf{x}, \rho) := -2\rho \min_{\mathbf{y}\in\mathcal{X}} \left[\langle \nabla\psi(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\rho}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \phi(\mathbf{y}) - \phi(\mathbf{x})\right]. \tag{13}$$

Let us assume that we take a step of length z_{i_k} along a fixed coordinate i_k. Therefore, using smoothness of $\nabla\psi$ and separability of ϕ, we can write,

$$f(\mathbf{x}^{(k-1)} + z_{i_k}\mathbf{e}_{i_k}) \le f(\mathbf{x}) + z_{i_k}[\nabla\psi(\mathbf{x})]_{i_k} + \frac{\beta}{2}z_{i_k}^2 + \phi_{i_k}(x_{i_k} + z_{i_k}) - \phi_{i_k}(x_{i_k}). \tag{14}$$

Let z_{i_k} be such that

$$z_{i_k} = \arg\min_t \left[t[\nabla\psi(\mathbf{x})]_{i_k} + \frac{\beta}{2}t^2 + \phi_{i_k}(x_{i_k} + t) - \phi_{i_k}(x_{i_k}).\right] \tag{15}$$

Using the precision guaranteed by the termination rule and conditioning on $\mathbf{x}^{(k-1)}$, we have

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^k)|\mathbf{x}^{(k-1)}] &\le f(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})}) + \epsilon_k \\ &\le f(\mathbf{x}^{(k-1)} + z_{i_k}\mathbf{e}_{i_k}) + \epsilon_k \end{aligned} \tag{16}$$

Taking expectation over the coordinate index i_k, which is uniformly distributed over the set $\{1, 2, \dots, d\}$, we can write,

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^k)|\mathbf{x}^{(k-1)}] &\le \mathbb{E}_{i_k}[f(\mathbf{x}^{(k-1)} + z_{i_k}\mathbf{e}_{i_k})] + \epsilon_k, \\ &\le \mathbb{E}_{i_k}\left[f(\mathbf{x}^{(k-1)}) + z_{i_k}[\nabla\psi(\mathbf{x}^{(k-1)})]_{i_k} + \frac{\beta}{2}z_{i_k}^2 2 + \phi_{i_k}(x_{i_k}^{(k-1)} + z_{i_k}) - \phi_{i_k}(x_{i_k}^{(k-1)})\right] + \epsilon_k, \\ &\le f(\mathbf{x}^{(k-1)}) + \frac{1}{d}\sum_{i=1}^{d}\min_{t_i}\left[t_i[\nabla\psi(\mathbf{x}^{(k-1)})]_i + \frac{\beta}{2}t_i^2 2 + \phi_i(x_i^{(k-1)} + t_i) - \phi_i(x_i^{(k-1)})\right] + \epsilon_k, \\ &\le f(\mathbf{x}^{(k-1)}) + \frac{1}{d}\min_{t_1,t_2,\dots,t_d}\left[\sum_{i=1}^{d} t_i[\nabla\psi(\mathbf{x}^{(k-1)})]_i + \frac{\beta}{2}t_i^2 2 + \phi_i(x_i^{(k-1)} + t_i) - \phi_i(x_i^{(k-1)})\right] + \epsilon_k, \\ &\le f(\mathbf{x}^{(k-1)}) + \frac{1}{d}\min_{\mathbf{y}}\left[\langle F_1(\mathbf{x}^{(k-1)}), \mathbf{y} - \mathbf{x}^{(k-1)}\rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}^{(k-1)}\| + \phi(\mathbf{y}) - \phi(x^{(k-1)})\right] + \epsilon_k, \\ &\le f(\mathbf{x}^{(k-1)}) - \frac{1}{2d\beta}\mathcal{D}_g(\mathbf{x}^{(k-1)}, \beta) + \epsilon_k, \\ &\le f(\mathbf{x}^{(k-1)}) - \frac{\alpha}{d\beta}(f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)) + \epsilon_k \end{aligned} \tag{17}$$

where the step uses the proximal PL inequality for strongly convex functions described in [1].

Taking expectation over $\mathbf{x}^{(k-1)}$, we obtain,

$$\mathbb{E}[F(\mathbf{x}^k)] - f(\mathbf{x}^*) \le \left(\mathbb{E}[F(\mathbf{x}^{(k-1)})] - f(\mathbf{x}^*)\right)\left(1 - \frac{\alpha}{d\beta}\right) + \epsilon_0\gamma^k \tag{18}$$

Let $\phi_k = \mathbb{E}[F(\mathbf{x}^k)] - f(\mathbf{x}^*)$. We claim that $\phi_k \le F_0\gamma^k$ where $F_0 = \max\left\{f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*), \frac{\epsilon_0}{(1-\gamma)}\right\}$. This can be proved using induction. For the base case, we have $\phi_0 = f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*) \le F_0$ by definition. Assume it is true for $k - 1$, then we have,

$$\begin{aligned}\phi_k &\le \left(1 - \frac{\alpha}{d\beta}\right)\phi_{k-1} + \epsilon_0\gamma^k \\ &\le \gamma^2(F_0\gamma^{k-1}) + \epsilon_0\gamma^k \\ &\le F_0\gamma^{k+1} + \epsilon_0\gamma^k \\ &\le \gamma^k\left(F_0\gamma + \epsilon_0\right) \\ &\le F_0\gamma^k. \end{aligned} \tag{19}$$

The last step follows from the choice of F_0. This completes the proof.

Proof of Lemma 3

We first prove the efficiency of SGD followed by the order optimality of the termination rule.

Efficiency of the SGD Routine

Consider a one-dimensional stochastic function $F(x)$ with stochastic gradient given by $G(x)$. Let x^* be the minimizer of the function, i.e., $x^* = \arg\min_{x \in \mathcal{X}} f(x)$ where $f(x) = \mathbb{E}[F(x)]$ and $\mathcal{X}$ is the domain of the function. The iterates generated by SGD with initial point x_0 satisfy the following relation,

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] &= \mathbb{E}\left[\|\mathrm{proj}_{\mathcal{X}}(x_t - \eta_t G(x_t) - x^*)\|^2\right] \\
&\leq \mathbb{E}\left[\|x_t - \eta_t G(x_t) - x^*\|^2\right] \\
&\leq \mathbb{E}\left[\|x_t - x^*\|^2 - 2\eta_t \langle G(x_t), x_t - x^* \rangle + \eta_t^2 \|G(x_t)\|^2\right] \\
&\leq \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta_t \mathbb{E}\left[\alpha \|x_t - x^*\|^2\right] + \eta_t^2 \mathbb{E}\left[\|G(x_t)\|^2\right] \\
&\leq (1 - 2\eta_t \alpha)\mathbb{E}\left[\|x_t - x^*\|^2\right] + \eta_t^2 g_{\max}^2 \qquad (20)
\end{aligned}$$

Next we show that the iterates satisfy $\mathbb{E}\left[\|x_t - x^*\|^2\right] \leq \dfrac{\mu_0}{1 + \nu t}$ for all $t \geq 0$ based on an inductive argument. The base case is ensured by choosing μ_0 satisfying $\mu_0 \geq \mathbb{E}[|x_0 - x^*|^2]$. For the induction step, note that the stepsizes are chosen as $\eta_t = \dfrac{\mu}{1 + \nu t}$ with $\mu = \dfrac{\mu_0 \alpha}{2 g_{\max}^2}$ and $\nu = \dfrac{\mu_0 \alpha^2}{4 g_{\max}^2}$. We continue with (20) as follows.

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] &\leq (1 - 2\eta_t \alpha)\mathbb{E}\left[\|x_t - x^*\|^2\right] + \eta_t^2 g_{\max}^2 \\
&\leq \left(1 - 2\frac{\mu\alpha}{1 + \nu t}\right)\frac{\mu_0}{1 + \nu t} + \frac{\mu^2}{(1 + \nu t)^2} g_{\max}^2 \\
&\leq \frac{\mu_0}{1 + \nu(t+1)} + \left(\frac{\mu_0}{1 + \nu t} - \frac{\mu_0}{1 + \nu(t+1)}\right) + \frac{\mu}{(1 + \nu t)^2}(\mu g_{\max}^2 - 2\mu_0 \alpha) \\
&\leq \frac{\mu_0}{1 + \nu(t+1)} + \frac{\mu_0^2}{(1 + \nu t)^2}(\mu^2 g_{\max}^2 - 2\mu\mu_0\alpha + \mu_0 \nu) \qquad (21) \\
&\leq \frac{\mu_0}{1 + \nu(t+1)} + \frac{\mu_0^2}{(1 + \nu t)^2}\left(\frac{\mu_0^2 \alpha^2}{4 g_{\max}^4} g_{\max}^2 - \frac{\mu_0^2 \alpha^2}{g_{\max}^2} + \frac{\mu_0^2 \alpha^2}{4 g_{\max}^2}\right) \qquad (22) \\
&\leq \frac{\mu_0}{1 + \nu(t+1)} \qquad (23)
\end{aligned}$$

Therefore, the iterates generated by SGD satisfy $\mathbb{E}\left[\|x_t - x^*\|^2\right] \leq \dfrac{\mu_0}{1 + \nu t}$ for all $t \geq 0$.
To ensure efficiency with respect to the initial point x_0, μ_0 should be of the order $\mu_0 \leq C\mathbb{E}[|x_0 - x^*|^2]$ as x_0 goes to x^* for some $C > 0$ (see below how this can be ensured within the PCM framework). Based on the strong convexity and smoothness of the function, the condition on the iterates can be translated to a condition on the function values as given below

$$\mathbb{E}[F(x_t) - f(x^*)] \leq \frac{\beta C}{\alpha} \frac{\mathbb{E}[f(x_0) - f(x^*)]}{1 + \nu t}, \qquad (24)$$

which implies that SGD is an efficient policy with $\lambda = 1$.

For implementation in PCM, the choice of μ_0 can be simplified using the relation on the CM iterates outlined in Lemma 2. In iteration k, $\mathbf{x}^{(k-1)}$ is the initial point, therefore, we can write, $\mathbb{E}\left[\|\mathbf{x}^{(k-1)} - \mathbf{x}^*_{(i_k, x^{(k-1)})}\|^2\right] \leq \dfrac{2}{\alpha}\mathbb{E}\left[f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*_{(i_k, x^{(k-1)})})\right] \leq \mathbb{E}\left[f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)\right] \leq F_0 \gamma^{k-1}$. Thus, for an appropriate choice of μ_0 for the first iteration, its value for consequent iterations can be obtained by the relation $\mu_0(k) = \gamma \mu_0(k-1)$, where $\mu_0(k)$ is the value of μ_0 used in iteration k.

Order Optimality of the Termination Rule

The correctness of the termination rule follows in a straightforward manner from the relation obtained on the iterates in the previous part. Using smoothness of the function and the relation obtained in (23), we have, $\mathbb{E}[F(x_t) - f(x^*)] \leq \frac{\mu_0 \beta}{2(1+\nu t)}$. Let t_0 be such that, $\frac{\mu_0 \beta}{2(1+\nu t_0)} \leq \epsilon$. On rearranging this equation, we obtain $t_0 \geq \frac{\mu_0 \beta}{2\epsilon\nu} - \frac{1}{\nu}$. Therefore, for all $t \geq t_0$, we have $\mathbb{E}[F(x_t) - f(x^*)] \leq \epsilon$. Since our choice of termination rule satisfies the above condition, we can conclude that our termination rule ensures the required precision. The order optimality of the termination also follows directly from the expression.

Analysis of PCM-RWT

In this section, we analyze the performance of the Random Walk on a Tree (RWT) under the PCM setup. We begin with briefly outlining the RWT algorithm for PCM setup followed by the termination rule and then conclude the section with the performance analysis of PCM-RWT.

Let $F(x,\xi)$ be the one dimensional stochastic function to be minimized and $G(x,\xi)$ denote its stochastic gradient while $f(x)$ and $g(x)$ respectively denote their expected values. Also we assume that $|g(x)| \leq g_{\max}$ for all $x \in \mathcal{X}'$, where $\mathcal{X}'$ is the domain of the function.

RWT Algorithm for PCM

In the k^{th} iteration of PCM, optimization is carried out in the along the direction i_k, chosen in that iteration. Therefore, the one dimensional domain is the interval given by $\{x : (x, \mathbf{x}^{(k-1)}_{-i_k}) \in \mathcal{X}\}$, that is, all the points in the domain whose all but the i_k^{th} coordinates are same as that of $\mathbf{x}^{(k-1)}$. The length of this interval depends upon the diameter of the domain along the i_k^{th} direction. Without loss of generality, we assume that the one-dimensional domain is the closed interval $[0,1]$ (as the extension to any interval $[a,b]$ is straightforward).

The basic idea of RWT is to construct an infinite-depth binary tree based on successive partitions of the interval. Each node of the tree represents a sub-interval with nodes at the same level giving an equal-length partition of $[0,1]$. The query point at each time is then generated based on a biased random walk on the interval tree that initiates at the root and is biased toward the node containing the minimizer x^* (equivalently, the node/interval that sees a sign change in the gradient). When the random walk reaches a node, the two end points along with the middle point of the corresponding interval are queried in serial to determine, with a required confidence level $\breve{p}$, the sign of $g(x)$ at those points. The test on the sign of $g(x)$ at any given x is done through a confidence-bound based local sequential test using random gradient observations. The outcomes of the sign tests at the three points of the interval determines the next move of the random walk: to the child that contains a sign change or back to the parent under inconsistent test outcomes.

A crucial aspect of the above algorithm is the local sequential test. Let the sample mean of s samples of the stochastic gradient at a point $x \in \mathcal{X}'$ be denoted as $\bar{G}_s(x) = \frac{1}{s}\sum_{t=1}^{s} G(x,\xi_t)$. The sequential test in RWT for sub-Gaussian noise is given below. For heavy-tailed noise, the only required change to RWT is in the confidence bounds used in the sequential test (see [2]).

▷ If $\overline{G}_s(x) > \sqrt{\frac{5\sigma_0^2}{s}\log\left(\frac{6\log s}{\sqrt{\breve{p}}}\right)}$, terminate; output 1.

▷ If $\overline{G}_s(x) < -\sqrt{\frac{5\sigma_0^2}{s}\log\left(\frac{6\log s}{\sqrt{\breve{p}}}\right)}$, terminate; output -1.

▷ Otherwise, take another sample of $G(x,\xi)$ and repeat.

Figure 1: The sequential test at a sampling point x under sub-Gaussian noise.

where $\breve{p}$ is the confidence parameter for the sequential test. To ensure the bias in the random walk, $\breve{p}$ is set to a value in $(0, 1 - 2^{-1/3})$.

The RWT algorithm as described in [2, 3] initializes the random walk at the root of the tree as there is no prior information about the location of the minimizer. However, if we have some prior information about the location of the minimizer, we can initialize the random walk at a lower level in the tree. This enables us to give higher preference to the region where the minimizer is likely to be located, thereby reducing the expected time to convergence. Consequently, such an initialization allows RWT to leverage favorable initial conditions. Therefore, for PCM, we initialize RWT at the node which contains the initial point and is at a level where the interval length is lesser than $\sqrt{\log_2\left(\frac{\beta\sqrt{2}}{\sqrt{\alpha\epsilon}}\right)\frac{2\mu_0}{\alpha}}$, where μ_0 is a carefully chosen hyperparameter and ϵ is the required precision. If the threshold exceeds 1, then we begin at the root. The significance of this choice of initialization and the allowed values of μ_0 are discussed in a later section which outlines an upper bound on $\mathbb{E}[\tau(\epsilon)]$.

Termination Rule

We begin with a lemma that states the correctness of the termination rule and also relate the expected second moment of the gradient of the final point to the required precision ϵ. Recall that the termination rule specified that if at a certain point the number of samples taken in a sequential test exceeds $N_0(\epsilon) = \frac{40\sigma_0^2}{\alpha\epsilon}\log\left(\frac{2}{\breve{p}}\log\left(\frac{80\sigma_0^2}{\alpha\breve{p}\epsilon}\right)\right)$ the algorithm must terminate, returning the current point being probed.

Lemma 1. *Let $x_{\tau(\epsilon)}$ denote the final point obtained under the termination rule. Then we have the following relations $\mathbb{E}[f(x_{\tau(\epsilon)}) - f(x^*)] \leq \epsilon$ and $\mathbb{E}[g^2(x_{\tau(\epsilon)})] \leq 2\alpha\epsilon$.*

Proof. The first part of the lemma directly follows from the second part using the strong convexity of the function. Since f is strongly convex, we have $\mathbb{E}[f(x_{\tau(\epsilon)}) - f(x^*)] \leq \dfrac{1}{2\alpha}\mathbb{E}[g^2(x_{\tau(\epsilon)})] \leq \epsilon$ as required. Hence, we just focus on the proving the bound on the gradient.

To obtain the bound on the gradient, we leverage the primary idea underlying the design of the threshold in the termination rule. The threshold is designed to ensure that the gradient at the point at which the algorithm terminates is sufficiently small with high probability. We use this high probability bound to obtain the required bound on the second moment of the gradient.

Define $\rho := \dfrac{\alpha\epsilon}{2}$. We claim that under the given termination rule, $|g(x_{\tau(\epsilon)}| \leq \rho$ holds with high probability. To prove the claim, we consider the probability that the random number of samples taken in a sequential test, denoted by $\hat{T}$, exceed any number n. For any point with $g(x) > 0$, we have,

$$\begin{aligned}
\mathbb{P}[\hat{T} > n] &\leq \mathbb{P}\left[\forall s \leq n : \overline{G}_s(x) + \sqrt{\frac{5\sigma_0^2}{s}\log\left(\frac{6\log s}{\sqrt{\breve{p}}}\right)} > 0, \text{ and } \overline{G}_s - \sqrt{\frac{5\sigma_0^2}{s}\log\left(\frac{6\log s}{\sqrt{\breve{p}}}\right)} < 0\right], \\
&\leq \mathbb{P}\left[\forall s \leq n : \overline{G}_s - \sqrt{\frac{5\sigma_0^2}{s}\log\left(\frac{6\log s}{\sqrt{\breve{p}}}\right)} < 0\right], \\
&\leq \mathbb{P}\left[\overline{G}_n - \sqrt{\frac{5\sigma_0^2}{n}\log\left(\frac{6\log n}{\sqrt{\breve{p}}}\right)} < 0\right], \\
&\leq \mathbb{P}\left[\overline{G}_n - \mathbb{E}_\xi[G(x,\xi)] < \sqrt{\frac{5\sigma_0^2}{n}\log\left(\frac{6\log n}{\sqrt{\breve{p}}}\right)} - g(x)]\right], \\
&\leq \exp\left(-\frac{n}{2\sigma_0^2}\left(\sqrt{\frac{5\sigma_0^2}{n}\log\left(\frac{6\log n}{\sqrt{\breve{p}}}\right)} - g(x)\right)^2\right).
\end{aligned} \tag{25}$$

The threshold $N_0(\epsilon)$ can be equivalently written in terms of ρ as $s_0(\rho) = \dfrac{20\sigma_0^2}{\rho^2}\log\left(\dfrac{2}{\breve{p}}\log\left(\dfrac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right)$. For all $n > s_0(\rho)$, we have,

$$\mathbb{P}[\hat{T} > n] \leq \exp\left(-\frac{n}{2\sigma_0^2}\left(\frac{\rho}{2} - g(x)\right)^2\right) \tag{26}$$

This can be obtained by plugging $n = s_0(\rho)$ in the upper bound in (25). A more detailed analysis of this step can be found in Appendix B in [3]. A similar analysis can be carried out for any point with $g(x) < 0$. Using (26), we can conclude that if the number of samples in a local test at a point x exceed $s_0(\rho)$, then $|g(x)| \leq \rho$ with probability at least $1 - \delta_0$ where $\delta_0 = \exp\left(-\dfrac{s_0(\rho)\rho^2}{8\sigma_0^2}\right)$.

We can now bound the second moment of the gradient as follows by noting that $\delta_0 \leq 1/2$

$$\begin{aligned}
\mathbb{E}[g^2(x_{\tau(\epsilon)})] &\leq \rho^2 \mathbb{P}[g(x_{\tau(\epsilon)}) \leq \rho] + \mathbb{E}[g^2(x_{\tau(\epsilon)}) \mathbb{1}_{\{g(x_{\tau(\epsilon)}) > \rho\}}], \\
&\leq \rho^2 + \sum_{r=1}^{\infty} (r+1)^2 \rho^2 \mathbb{P}(r\rho < g(x_{\tau(\epsilon)}) \leq (r+1)\rho), \\
&\leq \rho^2 + \sum_{r=1}^{\infty} (r+1)^2 \rho^2 \mathbb{P}(g(x_{\tau(\epsilon)}) > r\rho), \\
&\leq \rho^2 + \sum_{r=1}^{\infty} (r+1)^2 \rho^2 \exp\left(-\frac{s_0(\rho)}{2\sigma_0^2} \left(\frac{\rho}{2} - r\rho \right)^2 \right), \\
&\leq \rho^2 + \sum_{r=1}^{\infty} (r+1)^2 \rho^2 \exp\left(-\frac{s_0(\rho)\rho^2}{8\sigma_0^2} (2r-1)^2 \right), \\
&\leq \rho^2 + \sum_{r=1}^{\infty} (r+1)^2 \rho^2 \delta_0^{(2r-1)^2}, \\
&\leq \rho^2 + 2.01\rho^2, \\
&\leq 4\rho^2. \qquad (27)
\end{aligned}$$

By plugging in $\rho = \sqrt{\frac{\alpha\epsilon}{2}}$, we arrive at the required result. □

As it is easier to analyze expressions in terms of the gradient and not the function values, we will use the expressions in terms of ρ for the rest of the section keeping in mind its relation with the required precision of ϵ.

Upper Bound on $\mathbb{E}[\tau(\epsilon)]$

To bound the expected number of samples taken in one iteration of the PCM-RWT algorithm with precision ϵ, $\mathbb{E}[(\tau(\epsilon)]$, we need to obtain a bound on the number of steps taken by the random walk before termination. The bound on $\mathbb{E}[\tau(\epsilon)]$ follows by noting that the number of samples taken in each sequential test before termination is bounded by the threshold specified in the termination rule. For the bound on the number of steps in the random walk, we note that as the random walk gets to a deeper level in the tree, the magnitude of the gradient reduces. Consequently, the probability that the number of samples taken in the sequential test will cross the threshold increases as the walk goes to a deeper level in the tree. These decreasing tail probabilities can then be used to obtain a bound on the expected number of steps in the random walk.

Assume the minima to be $x^* = 0$. Such an assumption leads to no loss of generality as the analysis can easily be modified for any point in the interval and for any interval of any length. We begin the analysis for the case when the random walk is initialized at the root node. This analysis can be easily modified to accommodate the initialization at a deeper level.

We divide the tree into a sequence of subtrees given by $\mathcal{T}_1, \mathcal{T}_2, \dots$ where for all $i = 1, 2, \dots,$ the subtree $\mathcal{T}_i$ contains the node corresponding to the interval $[0, 2^{-(i-1)}]$ and its right child along with all its children. Thus, $\mathcal{T}_i$'s are half trees rooted at level $i-1$, along with their root. This construction is similar to the one outlined in [4]. Since the random walk is biased towards the minimizer, therefore given the construction of $\mathcal{T}_i$, the probability that random walk is still in one of such subtrees would decrease with time. To formalize this idea, we consider the last passage times of any subtree $\mathcal{T}_i$. Let τ_1 denote the last passage time to $\mathcal{T}_1$.

The analysis of the last passage time of $\mathcal{T}_1$ can be mapped to the problem of a random walk on the set $S = \{-1, 0, 1, 2, \dots\}$. The underlying idea is that each non-negative integer can be mapped to the corresponding level in subtree. Our random walk on the tree can between different levels is then equivalent to a random walk on these integers. The equivalence follows by noting that the specific intervals on any level are all identical as they do not contain the minimizer and thus can be abstracted into a single entity. Hence, we map the root node to 0, and set of nodes at level j in subtree $\mathcal{T}_1$ to integer j for $j > 0$. Lastly, we map the left subtree containing all nodes in the interval $[0, 0.5]$ to -1, which corresponds to an exit from the subtree $\mathcal{T}_1$.

The random walk can be modelled as a Markov chain on the set S, where $\mathbb{P}(j \to j+1) = 1-p$ for all $j \in S$, $\mathbb{P}(j \to j-1) = p$ for all $j \geq 0$ and $\mathbb{P}(-1 \to -1) = p$. The probability $p = \breve{p}^3 > 0.5$ is the probability of moving in correct direction where $\breve{p}$ is the confidence level in the sequential test. The initial state is 0.

Since -1 denotes the state corresponding to exiting the subtree $\mathcal{T}_1$, therefore our random walk still being in $\mathcal{T}_1$ after n steps is the same as the Markov Chain being in a state j for $j \geq 0$ after n steps. Furthermore, since the Markov Chain was initialized at 0, therefore being in state $j \geq 0$ implies that the number of steps taken in the positive direction are at least as many as those taken in the negative direction. Combining all these ideas along with noting the specific structure of the transition matrix, we can conclude that

$$\mathbb{P}(\tau_1 > n) = \mathbb{P}(Z \leq n/2), \tag{28}$$

where $Z \sim \text{Bin}(n,p)$. Writing expectation as the sum of tail probabilities,

$$\begin{aligned}
\mathbb{E}[\tau_1] &= \sum_{n=0}^{\infty} \mathbb{P}(\tau_1 > n), \\
&= \sum_{n=0}^{\infty} \mathbb{P}(Z \leq n/2), \\
&= \sum_{n=0}^{\infty} \exp(-2(p-1/2)^2 n), \\
&= \frac{1}{1-\exp(-2(p-1/2)^2)}.
\end{aligned} \tag{29}$$

The third step is obtained using Hoeffding's inequality. We can leverage the symmetry of the random walk and the binary tree to obtain the expected last passage time for any other subtree $\mathcal{T}_i$.

Let for all $i \geq 1$, $N_{\mathcal{T}_i}$ denote the random number of steps taken in subtree $\mathcal{T}_i$ before exiting that subtree and E_i denote the event that the random walk does not terminate in tree $\mathcal{T}_i$. If N_{RW} denotes the random number of steps taken by the random walk before termination then

$$\mathbb{E}[N_{RW}] = \mathbb{E}[N_{\mathcal{T}_1}] + \sum_{i=2}^{\infty} \mathbb{P}\left(\bigcap_{j=1}^{i-1} E_j\right) \mathbb{E}\left[N_{\mathcal{T}_i} \middle| \bigcap_{j=1}^{i-1} E_j\right]. \tag{30}$$

By definition we have $\mathbb{E}[N_{\mathcal{T}_1}] = \mathbb{E}[\tau_1]$. Furthermore, one can note that due to symmetry in the structure of the binary tree, $\mathbb{E}\left[N_{\mathcal{T}_i} \middle| \bigcap_{j=1}^{i-1} E_j\right] = \mathbb{E}[\tau_1]$. Hence, to evaluate (30) we need to find a bound on $\mathbb{P}\left(\bigcap_{j=1}^{i-1} E_j\right)$, the probability that the random walk does not terminate in $\mathcal{T}_j$ for $j = 1, 2, \ldots, i-1$ and $i \geq 2$. To bound this probability, consider the event that local sequential test takes less than $s_0(\rho)$ samples before termination when the magnitude of the gradient of the point being sampled is less than ρ. Let the event be denoted by $E_f(\rho)$ and let $\mathbb{P}(E_f(\rho)) \leq \eta_\rho$ for some $\eta_\rho < 1$.

Note that for any level $i > \log_2(\beta/\rho)$,the length of the interval at this level would be lesser than ρ/β. Using the smoothness of the function, it follows that the magnitude of gradient of any point probed in $\mathcal{T}_i$ for $i > \log_2(\beta/\rho)$ would be lesser than ρ. Therefore, for every $i > \log_2(\beta/\rho)$, if E_i occurs then $E_f(\rho)$ would definitely have occurred. Consequently, for all $i > i_0$,

$$\mathbb{P}\left(\bigcap_{j=1}^{i-1} E_j\right) \leq \eta_\rho^{i-i_0}, \tag{31}$$

where $i_0 = \lceil \log_2(\beta/\rho) \rceil$. For $i \leq i_0$, we can crudely upper bound this probability with 1. Plugging these relations into (30), we obtain,

$$
\begin{aligned}
\mathbb{E}[N_{RW}] &= \mathbb{E}[N_{\mathcal{T}_1}] + \sum_{i=2}^{\infty} \mathbb{P}\left(\bigcap_{j=1}^{i-1} E_j\right) \mathbb{E}\left[N_{\mathcal{T}_i}\Big|\bigcap_{j=1}^{i-1} E_j\right], \\
&\leq \mathbb{E}[\tau_1]\left(i_0 + 1 + \sum_{i=i_0+1}^{\infty} \eta_\rho^{i-i_0}\right), \\
&\leq \frac{1}{1-\exp(-2(p-1/2)^2)}\left(\lceil \log_2(\beta/\rho)\rceil + 1 + \sum_{i=1}^{\infty} \eta_\rho^i\right), \\
&\leq \frac{1}{1-\exp(-2(p-1/2)^2)}\left(\log_2(\beta/\rho) + 2 + \frac{\eta_\rho}{1-\eta_\rho}\right). \qquad (32)
\end{aligned}
$$

The above analysis provides an upper bound for the number of steps taken by the random walk when it it initialized at the root node. However, as mentioned previously, we would want the RWT to be initialized at a deeper level in the tree to leverage the favorable initial conditions. We can perform a similar analysis for number of steps taken by random walk in the case when RWT is initialized at a deeper level to leverage the favorable initial conditions.

As given in the description of PCM-RWT, we initialize the algorithm the node which contains the initial point and at a level where the interval length is lesser than $\sqrt{\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}}$, where μ_0 is a carefully chosen hyperparameter. If the threshold exceeds 1, then we begin at the root. To analyze the number of steps taken by the random walk, we consider the event that $|x_0 - x^*|^2 \leq \log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}$, where x_0 is randomly chosen. We denote the event by E_{x_0}. Under this event, we can carry out a similar analysis as before, with a minor change that instead of i_0, the maximum depth would be $i_1 \leq \log_2\left(\sqrt{\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}}\frac{\beta}{\rho}\right) + 1$. Under the case the above event does not occur, the random walk would have to take no more than an additional $i_2 \leq \log_2\left(\sqrt{\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}}\right) + 1$ steps before the previous analysis is again applicable. If $N_{RW-\text{new}}$ denotes the random number of steps taken by the random walk under this initialization scheme, then on combining the above results, we can write,

$$
\begin{aligned}
\mathbb{E}[N_{RW-\text{new}}] \leq \mathbb{P}(E_{x_0})&\left(\zeta_p\left(\log_2\left(\sqrt{\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}}\frac{\beta}{\rho}\right) + \hat{\eta}_\rho\right)\right), \\
&+ \mathbb{P}(E_{x_0}^c)\left(\zeta_p\left(\log_2\left(\sqrt{\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}}\right) + \log_2\left(\frac{\beta}{\rho}\right) + 1 + \hat{\eta}_\rho\right)\right), \qquad (33)
\end{aligned}
$$

where E^c denotes the complement of an event E, $\zeta_p = \frac{1}{1-\exp(-2(p-1/2)^2)}$ and $\hat{\eta}_\rho = 2 + \frac{\eta_\rho}{1-\eta_\rho}$. We can bound $\mathbb{P}(E_{x_0}^c)$ using Markov's inequality as follows,

$$
\begin{aligned}
\Pr\left(|x_0 - x^*|^2 > \log_2\left(\frac{\beta}{\rho}\right)\frac{2\mu_0}{\alpha}\right) &\leq \Pr\left(f(x_0) - f(x^*) > \log_2\left(\frac{\beta}{\rho}\right)\mu_0\right), \\
&\leq \mathbb{E}\left[f(x_0) - f(x^*)\right]\left(\log_2\left(\frac{\beta}{\rho}\right)\mu_0\right)^{-1}. \qquad (34)
\end{aligned}
$$

Setting $\mu_0 = \mathbb{E}\left[f(x_0) - f(x^*)\right]$ and plugging (34) in (33), we obtain,

$$\begin{aligned}
\mathbb{E}[N_{RW-\text{new}}] &\leq \left(\zeta_p\left(\frac{1}{2}\log_2\left(\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mathbb{E}\left[f(x_0)-f(x^*)\right]}{\alpha}\frac{\beta^2}{\rho^2}\right)+\hat{\eta}_\rho\right)\right) \\
&\quad + \left(\log_2\left(\frac{\beta}{\rho}\right)\right)^{-1}\left(\zeta_p\left(\frac{1}{2}\log_2\left(\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mathbb{E}\left[f(x_0)-f(x^*)\right]}{\alpha}\right)+\log_2\left(\frac{\beta}{\rho}\right)+1+\hat{\eta}_\rho\right)\right), \\
&\leq \left(\zeta_p\left(\frac{1}{2}\log_2\left(\log_2\left(\frac{\beta}{\rho}\right)\frac{2\mathbb{E}\left[f(x_0)-f(x^*)\right]}{\alpha}\frac{\beta^2}{\rho^2}\right)+\hat{\eta}_\rho\right)\right) \\
&\qquad + \zeta_p\left(\frac{\log_2(\beta/g_{\max}) + 0.5\log_2(2g_{\max}/\alpha) + \hat{\eta}_\rho + 1.5}{\log_2(\beta/g_{\max})}\right). \qquad (35)
\end{aligned}$$

As in the case of SGD, a similar analysis can be carried out for any $\mu_0 \sim \Theta(E\left[f(x_0) - f(x^*)\right])$. Furthermore, for PCM-RWT, μ_0 can be tuned for each iteration in the same manner as described for PCM-SGD, that is, by decreasing it by a factor of γ after every iteration. This proof can be readily extended to interval of any length l, by changing the value of i_0 to $\log_2(\beta l/\rho)$ and also appropriately changing the bound on $\mathbb{E}\left[f(x_0) - f(x^*)\right]$ in (35). For a different minimizer, the sequence of subtrees $\mathcal{T}_i$'s can be appropriately modified as described in [4] to obtain the same result.

Finally, using (35), we can obtain the bound on $\mathbb{E}[\tau(\epsilon)]$. If M_{RW} denotes the random number of local tests carried out before termination then $\mathbb{E}[M_{RW}] \leq \mathbb{E}[3N_{RW-\text{new}} + 3]$. Moreover, since the number of samples in each test can be at most $s_0(\rho)$, therefore, the expected number of samples can be no more than $\mathbb{E}[M_{RW}]s_0(\rho)$. Substituting the different bounds and the relation between ρ and ϵ, we obtain that for some constant $\tau_0 > 0$, independent of ϵ

$$\mathbb{E}[\tau(\epsilon)] \leq \frac{\tau_0}{\epsilon}\log\left(\frac{\mathbb{E}\left[f(x_0)-f(x^*)\right]}{\epsilon}\right)\log^2\left(\log\left(\frac{1}{\epsilon}\right)\right). \qquad (36)$$

Regret in One CM Iteration

In this section, we perform a brief analysis of the regret incurred in one CM iteration. This corresponds to the inner sum in the term R_1 in the regret decomposition of PCM, capturing the regret incurred by the routine υ in the local one dimensional minimization. Let $x_{(m)}$ denote the sampling point at the m^{th} time the local test is called by the random walk module and $\hat{T}_m$ denote the random number of samples taken at this point. Therefore, if $R_{RWT}(\epsilon)$ denotes the regret incurred by RWT in one CM iteration to get to precision of ϵ, then we have,

$$\begin{aligned}
R_{RWT}(\epsilon) &= \mathbb{E}\left[\sum_{m=1}^{M_{RW}}\sum_{t=1}^{\hat{T}_m} F(x_{(m)};\xi_t) - F(x^*,\xi_t)\right], \\
&\leq \mathbb{E}\left[\sum_{m=1}^{M_{RW}}\sum_{t=1}^{\hat{T}_m}\frac{1}{2\alpha}[g(x_{(m)})]^2\right], \\
&\leq \mathbb{E}\left[\sum_{m=1}^{M_{RW}}\mathbb{E}[\hat{T}_m]\frac{1}{2\alpha}[g(x_{(m)})]^2\right]. \qquad (37)
\end{aligned}$$

Note that $\hat{T}_m$ is the random number of samples taken at sampling point $x_{(m)}$ with the termination rule. If $\tilde{T}$ denotes the random number of samples taken without the termination rule then, $\hat{T}_m = \tilde{T}\mathbb{1}\{\tilde{T} \leq s_0(\rho)\}$ where $\rho = \sqrt{\alpha\epsilon/2}$. To bound $\mathbb{E}[\hat{T}_m]$, we use different methods depending on the gradient of the sampling point. If $|g(x_{(m)}| \leq \rho$, then we use the trivial bound $\mathbb{E}[\hat{T}_m] = \mathbb{E}[\tilde{T}\mathbb{1}\{\tilde{T} \leq s_0(\rho)\}] \leq s_0(\rho)$. For the other case of $|g(x_{(m)}| > \rho$, we note that $\mathbb{E}[\hat{T}_m] \leq \mathbb{E}[\tilde{T}] \leq \dfrac{40\sigma_0^2}{g(x_{(m)})^2}\log\left(\dfrac{2}{\check{p}}\log\left(\dfrac{40\sigma_0^2}{\check{p}g(x_{(m)})^2}\right)\right) + 2 \leq$

$\dfrac{40\sigma_0^2}{g(x_{(m)})^2} \log\left(\dfrac{2}{\breve{p}} \log\left(\dfrac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) + 2$. Plugging these bounds in (37), we obtain,

$$\begin{aligned}
R_{RWT}(\epsilon) \leq \mathbb{E}&\left[\sum_{m=1}^{M_{RW}} \frac{g(x_{(m)})^2}{2\alpha} \left(\mathbb{E}[\hat{T}_m]\mathbb{1}\{|g(x_{(m)})| > \rho\} + \mathbb{E}[\hat{T}_m]\mathbb{1}\{|g(x_{(m)})| \leq \rho\}\right)\right], \\
&\leq \mathbb{E}\Bigg[\sum_{m=1}^{M_{RW}} \left\{\frac{40\sigma_0^2}{[g(x_{(m)})]^2} \log\left(\frac{2}{\breve{p}} \log\left(\frac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) + 2\right\} \frac{1}{2\alpha}[g(x_{(m)})]^2 \mathbb{1}\{|\mathbb{E}_\xi[G(x_{(m)};\xi)]| > \rho\} \\
&\quad + \frac{20\sigma_0^2}{\rho^2} \log\left(\frac{2}{\breve{p}} \log\left(\frac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) \frac{\rho^2}{2\alpha} \mathbb{1}\{|\mathbb{E}_\xi[G(x_{(m)};\xi)]| \leq \rho\}\Bigg], \\
&\leq \mathbb{E}\Bigg[\sum_{m=1}^{M_{RW}} \left\{\frac{20\sigma_0^2}{\alpha} \log\left(\frac{2}{\breve{p}} \log\left(\frac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) + \frac{g_{\max}^2}{\alpha}\right\} \mathbb{1}\{|\mathbb{E}_\xi[G(x_{(m)};\xi)]| > \rho\} \\
&\quad + \frac{20\sigma_0^2}{\alpha} \log\left(\frac{2}{\breve{p}} \log\left(\frac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) \mathbb{1}\{|\mathbb{E}_\xi[G(x_{(m)};\xi)]| \leq \rho\}\Bigg], \\
&\leq \left(\frac{20\sigma_0^2}{\alpha} \log\left(\frac{2}{\breve{p}} \log\left(\frac{40\sigma_0^2}{\breve{p}\rho^2}\right)\right) + \frac{g_{\max}^2}{\alpha}\right) \mathbb{E}[M_{RW}], \\
&\leq \left(\frac{20\sigma_0^2}{\alpha} \log\left(\frac{2}{\breve{p}} \log\left(\frac{80\sigma_0^2}{\breve{p}\alpha\epsilon}\right)\right) + \frac{g_{\max}^2}{\alpha}\right) \mathbb{E}[3N_{RW-\text{new}} + 3]. \qquad (38)
\end{aligned}$$

Substituting the bound from (35) in the above equation, we can show that for some constant $\bar{R} > 0$, independent of ϵ,

$$R_{RWT}(\epsilon) \leq \bar{R} \log\left(\frac{\mathbb{E}\left[f(x_0) - f(x^*)\right]}{\epsilon}\right) \log^2\left(\log\left(\frac{1}{\epsilon}\right)\right). \qquad (39)$$

Regret Analysis of PCM-RWT

We can now combine all the results obtained about performance of RWT to analyze the performance of PCM-RWT. Using the decomposition of regret in R_1 and R_2, we bound each of these terms individually to obtain the bound on the overall regret.

We begin with bounding R_1. Note that we can now rewrite R_1 as,

$$\begin{aligned}
R_1 &\leq \mathbb{E}\left[\sum_{k=1}^{K} R_{RWT}(\epsilon_k)\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \bar{R} \log\left(\frac{\mathbb{E}[f(\mathbf{x}^{(k-1)}) - f(\mathbf{x}^*)]}{\epsilon_k}\right) \log^2\left(\log\left(\frac{1}{\epsilon_k}\right)\right)\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \bar{R} \log\left(\frac{F_0\gamma^k}{\epsilon_0\gamma^k}\right) \log^2\left(\log\left(\frac{1}{\epsilon_0}\right) + k\log\left(\frac{1}{\gamma}\right)\right)\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \bar{R}' \log^2\left(\log\left(\frac{1}{\epsilon_0}\right) + k\log\left(\frac{1}{\gamma}\right)\right)\right]. \qquad (40)
\end{aligned}$$

Using the result from Lemma 1 along with Jensen's inequality, we conclude that R_1 is of the order

$O(\log T \log^2(\log T))$. Similarly, we now consider R_2.

$$\begin{aligned}
R_2 &\leq \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}\left[F(\mathbf{x}^*_{(i_k,\mathbf{x}^{(k-1)})},\xi_t) - F(\mathbf{x}^*,\xi_t)\right]\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}[f(\mathbf{x}^{(k-1)} - f(\mathbf{x}^*)]\mathbb{E}[\tau(\epsilon_k)]\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}(F_0\gamma^{k-1})\frac{\tau_0}{\epsilon_k}\log\left(\frac{\mathbb{E}[f(\mathbf{x}^{k-1}) - f(x^*)]}{\epsilon_k}\right)\log^2\left(\log\left(\frac{1}{\epsilon_k}\right)\right)\right], \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K}\tau_0'\log^2\left(\log\left(\frac{1}{\epsilon_0}\right) + k\log\left(\frac{1}{\gamma}\right)\right)\right]. \qquad (41)
\end{aligned}$$

This is similar to the term we obtained in R_1 implying that R_2 is also of the order $O(\log T \log^2(\log T))$. Combining the two, we arrive at our required result.

Effect of Parallelization

In this section, we briefly describe the advantages obtained using parallelization. Consider the setup of m cores, connected in parallel to the main server. To make it similar to our original setup, we assume that each processor has the access to the oracle independently of others. It is assumed that $m \leq d$. The algorithm for implementing PCM using parallel updates is described as follows

1. Read the current iterate $\mathbf{x}$ and pass it to all cores.
2. Select m different indices from $\{1, 2, \ldots, d\}$ uniformly at random and allocate them to the cores.
3. On each core, run the one dimensional optimization routine along the dimension whose was index assigned to that core. The initial point for all the cores will be the same point $\mathbf{x}$. Let the points returned by the cores to the server be denoted as $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_m$.
4. Generate the next iterate $\mathbf{x}_1 = \dfrac{1}{m}\sum_{k=1}^{m} \mathbf{y}_k$.

The last step is the update or the synchronization step which ensures that the function value at the new iterate is lesser than that at the one previous one. Note that in the second step m different indices are chosen uniformly at random, that is, one of the $\binom{d}{m}$ sets is chosen.

The analysis of the above mentioned parallel implementation scheme is very similar to that of the sequential case. Let $\mathbb{1}_{(i,j)}$ denote the indicator variable for the i^{th} direction and j^{th} core. It is 1 if the i^{th} direction was chosen for optimization on the j^{th} core where $i = 1, 2, \ldots d$ and $j = 1, 2, \ldots m$. Thus, from equation (17), we have that for each $j = 1, 2, \ldots m$, $\mathbb{E}[f(\mathbf{y}_k)|\mathbf{x}] \leq f(\mathbf{x}) - \dfrac{1}{2\beta}\sum_{i=1}^{d} \mathbb{1}_{(i,j)}[g_i(\mathbf{x})]^2 + \epsilon$, where ϵ is the required accuracy. Using the update scheme in the synchronization step, we have,

$$\begin{aligned}
m\mathbb{E}[f(\mathbf{x}_1)|\mathbf{x}] &= m\mathbb{E}\left[f\left(\frac{1}{m}\sum_{j=1}^{m} \mathbf{y}_j \right) \Bigg| \mathbf{x} \right] \\
&\leq m\left(\frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[f(\mathbf{y}_j)|\mathbf{x}] \right) \\
&\leq \sum_{j=1}^{m}\left(f(\mathbf{x}) - \frac{1}{2\beta}\sum_{i=1}^{d} \mathbb{1}_{(i,j)}[g_i(\mathbf{x})]^2 + \epsilon \right) \\
\implies \mathbb{E}[f(\mathbf{x}_1)|\mathbf{x}] &\leq \frac{1}{m}\left(\sum_{j=1}^{m}\left(f(\mathbf{x}) - \frac{1}{2\beta}\sum_{i=1}^{d} \mathbb{1}_{(i,j)}[g_i(\mathbf{x})]^2 + \epsilon \right) \right) \\
&\leq f(\mathbf{x}) - \frac{1}{2m\beta}\sum_{j=1}^{m}\sum_{i=1}^{d} \mathbb{1}_{(i,j)}[g_i(\mathbf{x})]^2 + \epsilon
\end{aligned}$$

where the second step follows from the convexity of the function. Now taking expectation over the random choice of coordinates, we get,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_1)|\mathbf{x}] &\leq f(\mathbf{x}) - \mathbb{E}\left[\frac{1}{2m\beta}\sum_{j=1}^{m}\sum_{i=1}^{d}\mathbb{1}_{(i,j)}[g_i(\mathbf{x})]^2\right] + \epsilon \\
&\leq f(\mathbf{x}) - \frac{1}{2m\beta}\sum_{j=1}^{m}\sum_{i=1}^{d}\frac{m}{d}[g_i(\mathbf{x})]^2 + \epsilon \\
&\leq f(\mathbf{x}) - \frac{m}{2d\beta}\sum_{i=1}^{d}[g_i(\mathbf{x})]^2 + \epsilon \\
&\leq f(\mathbf{x}) - \frac{m}{2d\beta}\|g(\mathbf{x})\|^2 + \epsilon \\
&\leq f(\mathbf{x}) - \frac{m\alpha}{d\beta}\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \epsilon
\end{aligned}$$

Note that this expression is similar to one obtained in equation (17). Using an analysis similar to the one in Appendix A, we can obtain convergence rates for the case of parallel updates. The reduction in dimensionality dependence is evident through the factor d being replaced by d/m.

References

[1] H. Karimi, J. Nutini, and M. Schmidt, “Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9851 LNAI, 2016, pp. 795–811.

[2] S. Vakili, S. Salgia, and Q. Zhao, “Stochastic Gradient Descent on a Tree: An Adaptive and Robust Approach to Stochastic Convex Optimization,” in *2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019*, 2019, pp. 432–438.

[3] S. Vakili and Q. Zhao, “A random walk approach to first-order stochastic convex optimization,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 395–399.

[4] C. Wang, K. Cohen, and Q. Zhao, “Information-Directed Random Walk for Rare Event Detection in Hierarchical Processes,” 2018.