SGD Part 1

Dhruv Malik

September 10th 2018

Simple Case

We have a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ which has L-Lipschitz gradients:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \tag{1}$$

satisfies the PL Inequality:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu(f(x) - f(x^*)) \tag{2}$$

and is B-Lipschitz:

$$|f(x) - f(y)| \leq B \|x - y\| \tag{3}$$

The function is defined in such a way that the value $f(x) = \mathbb{E}_\xi[f(x, \xi)]$ where ξ is a random variable. We only have access to the function through noisy function evaluations. So if we query at point x, instead of getting $f(x)$ we receive:

$$f(x, \xi) \text{ where } \xi \text{ is a r.v. s.t. } \mathbb{E}_\xi[f(x, \xi)] = f(x) \tag{4}$$

We also impose the stronger condition on f:

$$|f(x, \xi) - f(y, \xi)| \leq B \|x - y\| \text{ for all instantiations of } \xi \tag{5}$$

Lemma 1. *Let $\mathbb{B}$ be the unit sphere and let v be sampled uniformly at random from $\mathbb{B}$. Let u be a unit norm vector randomly sampled from the shell of $\mathbb{B}$. Let f_r denote the smoothed version of f with radius r as follows:*

$$f_r(x) = \mathbb{E}_v[f(x + rv)] \tag{6}$$

We have the following estimator for $\nabla f_r(x)$:

$$\nabla f_r(x) = d\mathbb{E}_{u,\xi}\left[\frac{f(x + ru, \xi) - f(x, \xi)}{r} u\right] \tag{7}$$

Proof. First note that $\mathbb{E}_u[u] = 0$. By Flaxman we have that:

$$\begin{aligned}
\nabla f_r(x) &= \frac{d}{r}\mathbb{E}_u[f(x + ru)u] && (8)\\
&= \frac{d}{r}\mathbb{E}_u[f(x + ru)u] - \frac{d}{r}\mathbb{E}[f(x)u] && (9)\\
&= \frac{d}{r}\mathbb{E}_u[(f(x + ru) - f(x))u] && (10)
\end{aligned}$$

Now observe that:

$$\mathbb{E}_{u,\xi}[(f(x+ru,\xi)-f(x,\xi))u] = \mathbb{E}_u\left[\mathbb{E}_\xi[(f(x+ru,\xi)-f(x,\xi))u\middle|u]\right] \tag{11}$$

$$= \mathbb{E}_u\left[(f(x+ru)-f(x))u\right] \tag{12}$$

Putting the equations together completes the proof. □

Lemma 2. *The bias between $\nabla f_r(x)$ and $f(x)$ can be driven down arbitrarily low as follows:*

$$\|\nabla f_r(x) - \nabla f(x)\| \leq Lr \tag{13}$$

Proof. We have that:

$$\|\nabla f_r(x) - \nabla f(x)\| = \|\nabla \mathbb{E}_v[f(x+rv)] - \nabla f(x)\| \tag{14}$$

$$= \|\mathbb{E}_v[\nabla f(x+rv) - \nabla f(x)]\| \tag{15}$$

$$\leq \mathbb{E}_v[\|\nabla f(x+rv) - \nabla f(x)\|] \tag{16}$$

$$\leq Lr \tag{17}$$

This completes the proof. □

Lemma 3. *The norm of the estimator can be bounded as follows:*

$$\left\| d\frac{f(x+ru,\xi)-f(x,\xi)}{r}u \right\| \leq dB \tag{18}$$

Proof. We have that:

$$\left\| d\frac{f(x+ru,\xi)-f(x,\xi)}{r}u \right\| = d\left|\frac{f(x+ru,\xi)-f(x,\xi)}{r}\right| \tag{19}$$

$$\leq dB \tag{20}$$

This completes the proof. □

Lemma 4. *Set step size $\eta = \frac{\epsilon\mu}{d^2LB^2}$. Set smoothing radius $r = \eta$. Define g_t to be the gradient estimate obtained by randomly sampling ξ_t and unit direction u_t at time t:*

$$g_t = d\frac{f(x_t+ru_t,\xi_t)-f(x_t,\xi_t)}{r}u_t \tag{21}$$

Take the following step:

$$x_{t+1} = x_t - \eta g_t \tag{22}$$

Then we have the following expected $\mathcal{O}(\frac{d^2}{\epsilon}\log(\frac{1}{\epsilon}))$ convergence:

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \left(1 - \frac{\epsilon\mu^2}{d^2LB^2}\right)^t (f(x_0) - f(x^*)) + \epsilon \tag{23}$$

Proof. Note that:

$$\mathbb{E}_{u,\xi}[g_t] = d\mathbb{E}_{u,\xi}\left[\frac{f(x_t + ru_t, \xi_t) - f(x_t, \xi_t)}{r} u_t\right] \tag{24}$$
$$= \nabla f_r(x_t) \tag{25}$$

Also note that for some y_t such that $\|y_t\| \leq Lr$, we have that:

$$\mathbb{E}_{u,\xi}[g_t] = \nabla f_r(x_t) \tag{26}$$
$$= \nabla f(x_t) + y_t \tag{27}$$

Finally, observe that:

$$\|\nabla f(x_t)\| \leq B \tag{28}$$
$$\|g_t\|^2 \leq d^2 B^2 \tag{29}$$

Conditioning on x_t, consider the following:

$$\mathbb{E}_{u,\xi}[f(x_{t+1}) - f(x_t)] \leq \mathbb{E}_{u,\xi}[\nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2] \tag{30}$$
$$= \mathbb{E}_{u,\xi}[-\eta \nabla f(x_t)^\top (g_t) + \eta^2 \frac{L}{2}\|g_t\|_2^2] \tag{31}$$
$$= -\eta \nabla f(x_t)^\top (\nabla f_r(x_t)) + \eta^2 \frac{L}{2}\mathbb{E}_{u,\xi}[\|g_t\|_2^2] \tag{32}$$
$$= -\eta \nabla f(x_t)^\top (\nabla f(x_t) + y_t) + \eta^2 \frac{L}{2}\mathbb{E}_{u,\xi}[\|g_t\|_2^2] \tag{33}$$
$$\leq -\eta \nabla f(x_t)^\top \nabla f(x_t) + \eta \|\nabla f(x_t)\| \|y_t\| + \eta^2 \frac{L}{2}\mathbb{E}_{u,\xi}[\|g_t\|_2^2] \tag{34}$$
$$\leq -\eta \|\nabla f(x_t)\|_2^2 + \eta BLr + \eta^2 \frac{L}{2} d^2 B^2 \tag{35}$$
$$= -\eta \|\nabla f(x_t)\|_2^2 + \eta^2 BL + \eta^2 \frac{L}{2} d^2 B^2 \tag{36}$$
$$\leq -\eta \|\nabla f(x_t)\|_2^2 + \eta^2 d^2 LB^2 \tag{37}$$
$$\leq -\eta\mu(f(x_t) - f(x^*)) + \eta^2 d^2 LB^2 \tag{38}$$

So we have that:

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \left(1 - \eta\mu\right)(f(x_t) - f(x^*)) + \eta^2 d^2 LB^2 \tag{39}$$

Now observe that:

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \left(1 - \eta\mu\right)(f(x_t) - f(x^*)) + \eta^2 d^2 LB^2 \tag{40}$$
$$\leq \left(1 - \eta\mu\right)^{t+1}(f(x_0) - f(x^*)) + \eta^2 d^2 LB^2 \sum_{i=0}^{t}\left(1 - \eta\mu\right)^i \tag{41}$$
$$\leq \left(1 - \eta\mu\right)^{t}(f(x_0) - f(x^*)) + \eta^2 d^2 LB^2 \frac{1}{\eta\mu} \tag{42}$$
$$= \left(1 - \eta\mu\right)^{t}(f(x_0) - f(x^*)) + \eta \frac{d^2 LB^2}{\mu} \tag{43}$$
$$= \left(1 - \frac{\epsilon\mu^2}{d^2 LB^2}\right)^{t}(f(x_0) - f(x^*)) + \epsilon \tag{44}$$

This completes the proof. □

LQR Case

Define the following quantities:

$$\alpha_K = \min\left\{1, \frac{\sigma_{min}(Q)\mu}{4C(K)\left\|B\right\|\left(\left\|A-BK\right\|+1\right)}\right\} \tag{45}$$

$$\omega_K = \frac{\left\|R+B^\top P_K B\right\|\left(C(K)-C(K^*)\right)}{\mu} \tag{46}$$

$$\beta_K = \frac{C(K)}{\sigma_{min}(Q)}\sqrt{\omega} \tag{47}$$

$$\gamma_K = \max\left\{1, \frac{1}{\sigma_{min}(R)}\left(\sqrt{\omega}+\left\|B^\top P_K A\right\|\right)\right\} \tag{48}$$

$$\delta_K = 8\sup\{\|x_0\|_2^2\}\left(\frac{C(K)}{\mu\sigma_{min}(Q)}\right)^2\left(\|Q\|\,\|B\|+\gamma^2\,\|R\|\,\|B\|+\gamma\,\|R\|\right) \tag{49}$$

$$\lambda_K = \|R\|+\|B\|\,\|A\|\frac{\delta}{\mathbb{E}\,\|x_0\|^2}+2\gamma\,\|B\|^2\frac{\delta}{\mathbb{E}\,\|x_0\|^2}+\|B\|^2\,\|P_K\| \tag{50}$$

$$\phi_K = \left(5\frac{C(K)}{\sigma_{min}(Q)}\right)\left(\|R\|+\|B\|^2\frac{C(K)}{\mu}\right)+\sqrt{\omega}\frac{C(K)^2\,\|B\|}{\sigma_{min}(Q)^2\mu} \tag{51}$$

$$\rho = \max\left\{\frac{\|\Sigma_{K^*}\|}{\mu^2\sigma_{min}(R)}, 1\right\} \tag{52}$$

Theorem Rewritten

Lemma A. *If we have that:*

$$\left\|K'-K\right\|_F \le \alpha_K \tag{53}$$

then we have locally Lipschitz gradients as follows:

$$\left\|\nabla C(K')-\nabla C(K)\right\|_F \le \phi_K\left\|K'-K\right\|_F \tag{54}$$

Observe that $\phi_K = \mathcal{O}(C(K)^5)$.

Lemma B. *If we have that:*

$$\left\|K'-K\right\|_F \le \alpha_K \tag{55}$$

then:

$$|C(K')-C(K)| \le \delta_K\left\|K'-K\right\|_F \tag{56}$$

$$|C(K',x_0)-C(K,x_0)| \le \delta_K\left\|K'-K\right\|_F \tag{57}$$

Observe that $\delta_K = \mathcal{O}(C(K)^4)$.

Lemma C. *Let* $\mathbb{B}$ *be the unit sphere sitting in* D *dimensions and let* V *be sampled uniformly at random from* $\mathbb{B}$*. Let* U *be a matrix with unit Frobenius norm randomly sampled from the shell of* $\mathbb{B}$*. Let* $C_r(K)$ *denote the smoothed version of* $C(K)$ *with radius* r *as follows:*

$$C_r(K) = \mathbb{E}_V[C(K+rV)] \tag{58}$$

We have the following estimator for $\nabla C_r(K)$*:*

$$\nabla C_r(K) = \mathrm{D}\,\mathbb{E}_{U,x_0}\left[\frac{C(K+rU,x_0)-C(K,x_0)}{r}U\right] \tag{59}$$

Proof. First note that $\mathbb{E}_U[U] = 0$. By Flaxman we have that:

$$\nabla C_r(K) = \frac{\mathrm{D}}{r}\mathbb{E}_U[C(K+rU)U] \tag{60}$$
$$= \frac{\mathrm{D}}{r}\mathbb{E}_U[C(K+rU)U] - \frac{\mathrm{D}}{r}\mathbb{E}[C(K)U] \tag{61}$$
$$= \frac{\mathrm{D}}{r}\mathbb{E}_U[(C(K+rU) - C(K))U] \tag{62}$$

Now observe that:

$$\mathbb{E}_{U,x_0}[(C(K+rU,x_0) - C(K,x_0))U] = \mathbb{E}_U\left[\mathbb{E}_{x_0}[(C(K+rU,x_0) - C(K,x_0))U\middle|U]\right] \tag{63}$$
$$= \mathbb{E}_U\left[(C(K+rU) - C(K))U\right] \tag{64}$$

Putting the equations together completes the proof. □

Lemma D. *If we pick smoothing radius small enough:*

$$r \leq \alpha_K \tag{65}$$

then the bias between $\nabla C_r(K)$ and $C(K)$ can be driven down arbitrarily low as follows:

$$\|\nabla C_r(K) - \nabla C(K)\|_F \leq \phi_K r \tag{66}$$

Proof. Since $r \leq \alpha_K$, we have locally Lipschitz gradients. So we have that:

$$\|\nabla C_r(K) - \nabla C(K)\|_F = \|\nabla \mathbb{E}_V[C(K+rV)] - \nabla C(K)\|_F \tag{67}$$
$$= \|\mathbb{E}_V[\nabla C(K+rV) - \nabla C(K)]\|_F \tag{68}$$
$$\leq \mathbb{E}_V[\|\nabla C(K+rV) - \nabla C(K)\|_F] \tag{69}$$
$$\leq \phi_K r \tag{70}$$

This completes the proof. □

Lemma E. *If we pick smoothing radius small enough:*

$$r \leq \alpha_K \tag{71}$$

then the norm of the estimator at a point K can be bounded as follows:

$$\left\|\mathrm{D}\,\frac{C(K+rU,x_0) - C(K,x_0)}{r}U\right\|_F \leq \mathrm{D}\,\delta_K \tag{72}$$

Proof. We have that:

$$\left\|\mathrm{D}\,\frac{C(K+rU,x_0) - C(K,x_0)}{r}U\right\|_F = \mathrm{D}\left|\frac{C(K+rU,x_0) - C(K,x_0)}{r}\right| \tag{73}$$
$$\leq \mathrm{D}\,\delta_K \tag{74}$$

This completes the proof. □

Lemma F. *We have the following bound on the gradient:*

$$\|\nabla C(K)\|_F \leq \beta_K \tag{75}$$

We may want to replace this with δ_K***?***

Lemma G. *Assume we have a finite stochastic process X_t, such that for any t, we have that $|X_{t+1} - X_t| \leq c$ for some constant c. Further assume that the entire stochastic process is lower bounded by some constant $X^* \geq 0$. Finally, assume there exists some constant b such that:*

$$\mathbb{E}[X_{t+1}|X_t] \leq X_t \text{ if } X_t \geq X^* + b \tag{76}$$

We start from any X_0, but make the crucial assumption that $\lambda \geq 4b \geq 8c$. Then, if we fix an $n \geq 0$, we have that:

$$P(X_n - X_0 \geq \lambda) \leq \exp\left\{-\frac{\lambda^2}{8nc^2}\right\} \tag{77}$$

Proof. Let's define some events as follows:

- Define event $\mathcal{E}_\phi$ to be the event that for all $t \in \{0 \dots n\}$, $X_t \geq X^* + b$.
- For each $i \in \{0 \dots n\}$, define $\mathcal{E}_i$ to be the event that timestep i is the most recent time (looking backwards from timestep n) that the stochastic process falls below $X^* + b$. More concretely, event $\mathcal{E}_i$ occurs if $X_i < X^* + b$ and for all $t > i$ we have that $X_t \geq X^* + b$.

We observe that these events are disjoint, and one of these events must occur. So we have:

$$P(\mathcal{E}_\phi) + \sum_{i=0}^{n} P(\mathcal{E}_i) = 1 \tag{78}$$

We first condition on event $\mathcal{E}_\phi$. In this case, the entire sequence of iterates $X_0 \dots X_n$ is a super martingale and we directly apply Azuma's Inequality to obtain the following:

$$P(X_n - X_0 \geq \lambda|\mathcal{E}_\phi) \leq \exp\left\{-\frac{\lambda^2}{2nc^2}\right\} \tag{79}$$

We now condition on event $\mathcal{E}_n$. First, assume that $X_0 \geq X^* + b$. In this case we automatically have that $X_n \leq X_0$. In other case, assume that $X_0 \leq X^* + b$. In this case, we use the fact that $\lambda \geq 4b$ so $X_n - X_0 < X^* + b - X^* = b \leq \lambda$. So we conclude:

$$P(X_n - X_0 \geq \lambda|\mathcal{E}_i) = 0 \tag{80}$$

We now condition on event $\mathcal{E}_i$ for any $i \in \{0 \dots n-1\}$. Observe that if event $\mathcal{E}_i$ occurs, then the sequence of iterates $X_{i+1} \dots X_n$ is a super martingale since $X_j \geq X^* + b$ for each $j \in \{i+1 \dots n\}$. Also, note that $X_{i+1} \leq X^* + b + c \leq X^* + 2b$. Finally, use the fact that $\lambda \geq 4b$ to say that $X_0 \geq X^* + 2b$ or $X_{i+1} - X_0 \leq X^* + 2b - X^* \leq \frac{\lambda}{2}$. This implies that $X_0 + \lambda - X_{i+1} \geq \lambda - \frac{\lambda}{2} \geq \frac{\lambda}{2}$. We can thus apply Azuma's Inequality to this sequence of iterates to obtain the following:

$$\begin{aligned}
P(X_n - X_0 \geq \lambda|\mathcal{E}_i) &= P(X_n - X_{i+1} \geq X_0 + \lambda - X_{i+1}|\mathcal{E}_i) && (81)\\
&\leq P(X_n - X_{i+1} \geq \frac{\lambda}{2}|\mathcal{E}_i) && (82)\\
&\leq \exp\left\{-\frac{\lambda^2}{8(n-i-1)c^2}\right\} && (83)
\end{aligned}$$

We now use the fact that the events are disjoint and Bayes Rule to make the following argument:

$$
\begin{aligned}
P(X_n - X_0 \geq \lambda) &= P((X_n - X_0 \geq \lambda) \bigcap \mathcal{E}_\phi) + \sum_{i=0}^{n} P((X_n - X_0 \geq \lambda) \bigcap \mathcal{E}_i) && (84) \\
&= P(X_n - X_0 \geq \lambda | \mathcal{E}_\phi) P(\mathcal{E}_\phi) + \sum_{i=0}^{n} P(X_n - X_0 \geq \lambda | \mathcal{E}_i) P(\mathcal{E}_i) && (85) \\
&= P(X_n - X_0 \geq \lambda | \mathcal{E}_\phi) P(\mathcal{E}_\phi) + \sum_{i=0}^{n-1} P(X_n - X_0 \geq \lambda | \mathcal{E}_i) P(\mathcal{E}_i) && (86) \\
&\leq \exp\left\{ - \frac{\lambda^2}{2nc^2} \right\} P(\mathcal{E}_\phi) + \sum_{i=0}^{n-1} \exp\left\{ - \frac{\lambda^2}{8(n-i-1)c^2} \right\} P(\mathcal{E}_i) && (87) \\
&\leq \exp\left\{ - \frac{\lambda^2}{8nc^2} \right\} P(\mathcal{E}_\phi) + \sum_{i=0}^{n-1} \exp\left\{ - \frac{\lambda^2}{8nc^2} \right\} P(\mathcal{E}_i) && (88) \\
&= \exp\left\{ - \frac{\lambda^2}{8nc^2} \right\} \left[P(\mathcal{E}_\phi) + \sum_{i=0}^{n-1} P(\mathcal{E}_i) \right] && (89) \\
&\leq \exp\left\{ - \frac{\lambda^2}{8nc^2} \right\} && (90)
\end{aligned}
$$

This completes the proof. $\square$

Given a $C(K_0)$, define the following quantities by upper bounding the $C(K)$ terms in all the earlier polynomials by $C(K_0)+2$:

$$\alpha = \min\left\{1, \frac{\sigma_{min}(Q)\mu}{8(C(K_0)+2)\left\|B\right\|}\right\} \tag{91}$$

$$\omega = \max\left\{\frac{(\left\|R\right\| + \left\|B\right\|^2 \frac{C(K_0)+2}{\mu})(C(K_0)+2-C(K^*))}{\mu}, 1\right\} \tag{92}$$

$$\beta = \max\left\{\frac{C(K_0)+2}{\sigma_{min}(Q)}\sqrt{\omega}, 1\right\} \tag{93}$$

$$\gamma = \max\left\{1, \frac{1}{\sigma_{min}(R)}\left(\sqrt{\omega} + \left\|A\right\|\left\|B\right\|\frac{C(K_0)+2}{\mu}\right)\right\} \tag{94}$$

$$\delta = \max\left\{8\sup\{\|x_0\|_2^2\}\left(\frac{C(K)}{\mu\sigma_{min}(Q)}\right)^2(\left\|Q\right\|\left\|B\right\| + \gamma^2\left\|R\right\|\left\|B\right\| + \gamma\left\|R\right\|), 1\right\} \tag{95}$$

$$\lambda = \left\|R\right\| + \left\|B\right\|\left\|A\right\|\frac{\delta}{\mathbb{E}\left\|x_0\right\|^2} + 2\gamma\left\|B\right\|^2\frac{\delta}{\mathbb{E}\left\|x_0\right\|^2} + \left\|B\right\|^2\frac{C(K_0)+2}{\mu} \tag{96}$$

$$\phi = \max\left\{\left(5\frac{C(K)}{\sigma_{min}(Q)}\right)\left(\left\|R\right\| + \left\|B\right\|^2\frac{C(K)}{\mu}\right) + \sqrt{\omega}\frac{C(K)^2\left\|B\right\|}{\sigma_{min}(Q)^2\mu}, \frac{1}{\alpha}, 1\right\} \tag{97}$$

$$\rho = \max\left\{\frac{\left\|\Sigma_{K^*}\right\|}{\mu^2\sigma_{min}(R)}, 1\right\} \tag{98}$$

Observe that this choice ensures that for any K such that $C(K) \leq C(K_0)+2$, we have that $\alpha \leq \alpha_K$ and $\omega \geq \omega_K$, $\beta \geq \beta_K$, $\gamma \geq \gamma_K$, $\delta \geq \delta_K$, $\lambda \geq \lambda_K$, $\phi \geq \phi_K$.

Theorem 1. *For a value of $p \in (0,1)$, pick a k which satisfies:*

$$k = \max\left\{0, \frac{\log\left\{\frac{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2\rho}{\phi\delta}\log\left(\frac{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(\mathrm{D})}\right\} \tag{99}$$

$$\approx \mathcal{O}\left(\frac{\log\left\{\epsilon\left(\log(\frac{1}{\epsilon})\right)^2\log\left(\frac{\mathrm{D}\log(\frac{1}{\epsilon})}{p}\right)\right\}}{\log(\mathrm{D})}\right) \tag{100}$$

Define $\eta = \frac{\epsilon}{\mathrm{D}^{3+k}}\frac{1}{\rho\phi\delta^7}$, this satisfies $\eta \leq \alpha$. Set smoothing radius $r = \frac{\eta}{2}$ at each timestep sample a random unit (in Frobenius norm) matrix U as well as an initial state x_0. Use this to obtain a gradient estimate as follows:

$$\mathrm{G}_t = \mathrm{D}\,\frac{C(K+rU, x_0) - C(K, x_0)}{r}U \tag{101}$$

At any time t, take the following step:

$$K_{t+1} = K_t - \eta\,\mathrm{G}_t \tag{102}$$

Define $T = \frac{\mathrm{D}^{3+k}}{\epsilon}\rho^2\phi\delta^7\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$. Define $\mathcal{E}$ to be the event that for all $t \in \{1\ldots T\}$, we have that $C(K_t) \leq C(K_0)+2$. Then we have:

$$P(\mathcal{E}) \geq 1-p \tag{103}$$

Conditioned on $\mathcal{E}$, we have the following $\mathcal{O}(\frac{\mathrm{D}^{3+k}}{\epsilon}\log(\frac{1}{\epsilon}))$ convergence:

$$\mathbb{E}[C(K_T) - C(K^*)|\mathcal{E}] \leq \epsilon \tag{104}$$

where we use $\mathcal{O}(1)$ function evaluations at each iteration. For any one of arbitrarily high values of dimension, arbitrarily high values of $C(K_0)$, or arbitrarily low values of ϵ, this is an algorithm which has $k \approx 0$ and so uses a total of $\approx \mathcal{O}(\frac{\mathrm{D}^3}{\epsilon}\log(\frac{1}{\epsilon}))$ zero order information. This is a dimension factor away from the PL+Smooth dependence on dimension (that we obtained), and a logarithmic factor away from optimal dependence on ϵ.

Proof. First, observe that if at timestep t we have $C(K_t) \leq C(K_0) + 2$, then since $r \leq \alpha \leq \alpha_{K_t}$, we have that:

$$\mathbb{E}[\mathrm{G}_t] = \nabla C_r(K_t) \tag{105}$$

$$= \nabla C(K_t) + P \tag{106}$$

where P is a matrix such that $\|P\|_F \leq \phi_{K_t} r$. Additionally, for any t such that $C(K_t) \leq C(K_0) + 2$, we have that $\|\mathrm{G}_t\|_F \leq \delta_{K_t}\,\mathrm{D} \leq \delta\,\mathrm{D}$.

Assume that at timestep t we have $C(K_t) \leq C(K_0) + 2$. Then since $\eta \leq \alpha \leq \alpha_{K_t}$ we have that:

$$\mathbb{E}[C(K_{t+1}) - C(K_t)] \leq \mathbb{E}[Tr(\nabla C(K_t)^\top (K_{t+1} - K_t)) + \frac{\phi_{K_t}}{2} Tr((K_{t+1} - K_t)^\top (K_{t+1} - K_t))] \tag{107}$$

$$= \mathbb{E}[-\eta Tr(\nabla C(K_t)^\top \mathrm{G}_t) + \eta^2 \frac{\phi_{K_t}}{2} Tr(\mathrm{G}_t^\top \mathrm{G}_t)] \tag{108}$$

$$= -\eta Tr(\nabla C(K_t)^\top [\nabla C(K_t) + P]) + \mathbb{E}[\eta^2 \frac{\phi_{K_t}}{2} \|\mathrm{G}_t\|_F^2] \tag{109}$$

$$= -\eta \|\nabla C(K_t)\|_F^2 - \eta Tr(\nabla C(K_t)^\top P]) + \mathbb{E}[\eta^2 \frac{\phi_{K_t}}{2} \|\mathrm{G}_t\|_F^2] \tag{110}$$

$$\leq -\eta \|\nabla C(K_t)\|_F^2 + \eta \|\nabla C(K_t)\|_F \|P\|_F + \mathbb{E}[\eta^2 \frac{\phi_{K_t}}{2} \|\mathrm{G}_t\|_F^2] \tag{111}$$

$$\leq -\eta \|\nabla C(K_t)\|_F^2 + \frac{\eta^2}{2}\beta_{K_t}\phi_{K_t} + \mathbb{E}[\eta^2 \frac{\phi_{K_t}}{2} \|\mathrm{G}_t\|_F^2] \tag{112}$$

$$\leq -\eta \|\nabla C(K_t)\|_F^2 + \frac{\eta^2}{2}\beta_{K_t}\phi_{K_t} + \eta^2 \frac{\phi_{K_t}}{2}\,\mathrm{D}^2\,\delta_{K_t}^2 \tag{113}$$

$$\leq -\eta \|\nabla C(K_t)\|_F^2 + \eta^2 \phi_{K_t}\,\mathrm{D}^2\,\delta_{K_t}^2 \tag{114}$$

$$\leq -\eta \frac{1}{\rho}\Big(C(K_t) - C(K^*)\Big) + \eta^2 \phi_{K_t}\,\mathrm{D}^2\,\delta_{K_t}^2 \tag{115}$$

$$\leq -\eta \frac{1}{\rho}\Big(C(K_t) - C(K^*)\Big) + \eta^2\,\mathrm{D}^2\,\phi\delta^2 \tag{116}$$

We observe that this is expectation is only guaranteed to be negative when $C(K_t) - C(K^*) \geq \frac{\epsilon}{\mathrm{D}^{1+k}}\frac{1}{\delta^5}$.

Now, conditioning on $\mathcal{E}$ means that our cost iterates $C(K_t)$ are uniformly upper bounded by

$C(K_0)+2$. So, we have that:

$$\mathbb{E}[C(K_{t+1})-C(K^*)|\mathcal{E}] \le \left(1-\eta\frac{1}{\rho}\right)(C(K_t)-C(K^*))+\eta^2\,\mathrm{D}^2\,\phi\delta^2 \tag{117}$$

$$\le \left(1-\eta\frac{1}{\rho}\right)^t(C(K_0)-C(K^*))+\sum_{i=0}^{\infty}\left(1-\eta\frac{1}{\rho}\right)^i\eta^2\,\mathrm{D}^2\,\phi\delta^2 \tag{118}$$

$$\le \left(1-\eta\frac{1}{\rho}\right)^t(C(K_0)-C(K^*))+\frac{1}{\eta\frac{1}{\rho}}\eta^2\,\mathrm{D}^2\,\phi\delta^2 \tag{119}$$

$$= \left(1-\eta\frac{1}{\rho}\right)^t(C(K_0)-C(K^*))+\eta\rho\,\mathrm{D}^2\,\phi\delta^2 \tag{120}$$

$$= \left(1-\eta\frac{1}{\rho}\right)^t(C(K_0)-C(K^*))+\frac{\epsilon}{\mathrm{D}^{3+k}}\frac{1}{\rho\phi\delta^7}\rho\,\mathrm{D}^2\,\phi\delta^2 \tag{121}$$

$$\le \left(1-\eta\frac{1}{\rho}\right)^t(C(K_0)-C(K^*))+\frac{\epsilon}{2} \tag{122}$$

$$= \left(1-\frac{\epsilon}{\mathrm{D}^{3+k}}\frac{1}{\rho^2\phi\delta^7}\right)^t(C(K_0)-C(K^*))+\frac{\epsilon}{2} \tag{123}$$

Therefore, after $T=\frac{\mathrm{D}^{3+k}}{\epsilon}\rho^2\phi\delta^7\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ iterations, we have that $\mathbb{E}[C(K_T)-C(K^*)|\mathcal{E}]\le\epsilon$. The remainder of the proof is dedicated to showing that $\mathcal{E}$ occurs with high probability. We will use martingale arguments.

In order to do this, we first make a key observation: a priori the stochastic process $C(K_t)$ is not a super martingale. There is the relatively minor issue we saw above that $C(K_t)-C(K^*)\ge\frac{\epsilon}{\mathrm{D}^{1+k}}\frac{1}{\delta^5}$ is necessary for $\mathbb{E}[C(K_{t+1})-C(K_t)]\le 0$. However, a larger issue is that the cost function is only finite over a bounded region of stability. But for $C(K_t)$ to be a super martingale, we need the fact that $\mathbb{E}[C(K_t)]<\infty$ for all t. An even larger issue is that even if $C(K_t)$ was always finite, in order to write out Equations 107 through 116 to get expected decrease we need to ensure that the cost function is not blowing up arbitrarily high otherwise our choice of step size will fail.

We address the minor issue using the new concentration inequality we give in Lemma G, which is basically a modified version of Azuma's Inequality designed to handle this edge case. We address the larger issues by only considering the following interval of time:

$$T_{total} \coloneqq \{0,1\ldots T\} \tag{124}$$

$$= \left\{0,1\ldots\frac{\mathrm{D}^{3+k}}{\epsilon}\rho^2\phi\delta^7\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right\} \tag{125}$$

All our results hold only for this interval. Even in this interval, we are not immediately guaranteed that $\mathbb{E}[C(K_t)]<\infty$ for all $t\in T_{total}$ nor the ability to write Equations 107 through 116. The way we show this is to use the Lipschitzness of the cost function shown in Lemma H. We make the crucial observation that starting at $C(K_0)$, if we take $\mathcal{O}(\frac{1}{\eta})$ steps where each step is of size $\mathcal{O}(\eta)$, then the cost function has increased by at most 1. This means that $C(K_{\mathcal{O}(\frac{1}{\eta})})\le C(K_0)+1$. Clearly then, we have that $C(K_t)$ is finite and satisfies our upper bound of $C(K_0)+2$ for all $t\in\{0,1\ldots\mathcal{O}(\frac{1}{\eta})\}$. Within this interval (which we refer to as a chunk), we can write down Equations 107 through 116 and thus have that $C(K_t)$ is a super martingale (ignoring the minor issue).

We can now apply our modified Azuma's Inequality to $C(K_t)$ within this interval to show that with some probability $C(K_t) \leq \mathcal{O}(\frac{1}{T\eta})$ for all t in the chunk. If we condition on this event, then we can make the same argument for the next chunk $\{\mathcal{O}(\frac{1}{\eta}) \dots \mathcal{O}(\frac{2}{\eta})\}$ to show that $C(K_t) \leq \frac{2}{T\eta}$ for all t in this next chunk. Iteratively condition on these events happening for the past chunks and show this for each new chunk. Since the chunks are of size $\mathcal{O}(\frac{1}{\eta})$, there are only $\mathcal{O}(T\eta)$ total chunks. So by showing that in each chunk we increase by at most $\mathcal{O}(\frac{1}{T\eta})$, the final amount that we increase after $\mathcal{O}(T\eta)$ chunks is about 1. This gives us a uniform upper bound and concludes our proof sketch.

We now begin the rigorous proof, in the same spirit as our sketch. Observe that for any t such that $C(K_t) \leq C(K_0) + 2$, our choice of step size $\eta \leq \alpha \leq \alpha_{K_t}$ allows us to use Lemma H to say that:

$$\begin{aligned}
\left|C(K_{t+1}) - C(K_t)\right| &\leq \delta_{K_t} \left\|K_{t+1} - K_t\right\|_F && (126)\\
&\leq \delta\eta \left\|\mathrm{G}_t\right\|_F && (127)\\
&\leq \eta\delta^2 \,\mathrm{D} && (128)\\
&= \frac{\epsilon}{\mathrm{D}^{2+k}} \frac{1}{\rho\delta^5\phi} && (129)
\end{aligned}$$

We define $c \coloneqq \frac{\epsilon}{\mathrm{D}^{2+k}} \frac{1}{\rho\delta^5\phi}$. We now define the following chunks (time intervals):

$$\begin{aligned}
T_i &\coloneqq \left\{\frac{i-1}{c} + 1 \dots \frac{i}{c}\right\} && (130)\\
&= \left\{\frac{\mathrm{D}^{2+k}}{\epsilon}\rho\delta^5\phi(i-1) + 1 \dots \frac{\mathrm{D}^{2+k}}{\epsilon}\rho\delta^5\phi i\right\} && (131)
\end{aligned}$$

Since we are considering a total of $T = \frac{\mathrm{D}^{3+k}}{\epsilon}\rho^2\phi\delta^7 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ timesteps, and each chunk T_i has size $\frac{1}{c} = \frac{\mathrm{D}^{2+k}}{\epsilon}\rho\delta^5\phi$, we get that there are a total of $Tc = \mathrm{D}\,\rho\delta^2 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ chunks.

We now define the following events for $i \in \{1 \dots Tc\}$:

$$\mathcal{E}_i \coloneqq \text{ The event that for all } t \in T_i, \text{ we have that } C(K_t) \leq C(K_0) + 2 \qquad (132)$$

We have a total of $Tc = \mathrm{D}\,\rho\delta^2 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ such events.

Now, define the last timestep in chunk T_i to be m_i. So, $m_i \coloneqq \frac{\mathrm{D}^{2+k}}{\epsilon}\rho\delta^5\phi i$. Now define the following events:

$$\begin{aligned}
\mathcal{X}_i \coloneqq &\text{ The event that } C(K_{m_i}) \leq C(K_0) + \frac{i}{Tc} && (133)\\
= &\text{ The event that } C(K_{m_i}) \leq C(K_0) + \frac{i}{\mathrm{D}\,\rho\delta^2 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} && (134)
\end{aligned}$$

Again, we have a total of $Tc = \mathrm{D}\,\rho\delta^2 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ such events. We additionally define $m_0 = 0$, so $C(K_{m_0}) = C(K_0)$ and of course the corresponding new event $\mathcal{X}_0$ trivially occurs almost surely.

We now show that $\mathcal{X}_i$ implies $\mathcal{E}_{i+1}$. In order to do so, we use an inductive argument to prove that conditioning on $\mathcal{X}_i$ for any $i \geq 0$, we have that $C(K_t) \leq C(K_{m_i}) + (t - m_i)c$ for all $t \in T_{i+1}$.

- **Base Case:** If we condition on $\mathcal{X}_i$, then $C(K_{m_i}) \leq C(K_0) + \frac{i}{Tc} \leq C(K_0) + 1$. Since we have $\eta \leq \alpha \leq \alpha_{K_t}$, we apply Lemma H to say that $C(K_{m_i+1}) \leq C(K_{m_i}) + c$. Since $m_i + 1$ is the first timestep in T_{i+1}, this shows the base case.

- **Inductive Step:** Now assume that for some $t \neq m_{i+1} \in T_{i+1}$, we have that $C(K_t) \leq C(K_{m_i}) + (t - m_i)c$. Observe that since $t \neq m_{i+1} \in T_{i+1}$ we have $t - m_i < \frac{1}{c}$, so $C(K_t) \leq C(K_{m_i}) + (t - m_i)c < C(K_{m_i}) + 1 \leq C(K_0) + 2$. Again, we have $\eta \leq \alpha \leq \alpha_{K_t}$ so we apply Lemma H and say that $C(K_{t+1}) \leq C(K_t) + c \leq C(K_{m_i}) + (t - m_i)c + c = C(K_{m_i}) + (t + 1 - m_i)c$.

This completes the inductive argument to prove that conditioning on $\mathcal{X}_i$, we have that $C(K_t) \leq C(K_{m_i}) + (t - m_i)c$ for any $t \in T_{i+1}$. This statement implies that given $\mathcal{X}_i$ we have that $C(K_t) \leq C(K_{m_i}) + (t - m_i)c \leq C(K_{m_i}) + 1 \leq C(K_0) + 2$ for any $t \in T_{i+1}$ since such a t must satisfy $t - m_i \leq \frac{1}{c}$. This exactly shows that $\mathcal{X}_i$ implies $\mathcal{E}_{i+1}$.

Thus, we have that $\bigcap_{i=0}^{Tc} \mathcal{X}_i$ implies $\bigcap_{i=1}^{Tc} \mathcal{E}_i$. And of course, $\bigcap_{i=1}^{Tc} \mathcal{E}_i$ implies $\mathcal{E}$. Thus, to show that $\mathcal{E}$ holds with high probability it is sufficient to show that $\bigcap_{i=0}^{Tc} \mathcal{X}_i$ holds with high probability.

We now show that $\bigcap_{i=0}^{Tc} \mathcal{X}_i$ holds with high probability. Let's start with showing $\mathcal{X}_1$. Observe that since $\mathcal{X}_0$ is trivially satisfied, the above argument shows it implies $\mathcal{E}_1$. So the necessary boundedness condition of $C(K_t) \leq C(K_0) + 2$ is satisfied for all $t \in T_1$. This implies that for $t \in T_1$, we have $\eta \leq \alpha \leq \alpha_{K_t}$ and can write down Equations 107 through 116. So, $C(K_t)$ either has expected decrease or we have that $C(K_t) \leq C(K^*) + \frac{\epsilon}{\mathrm{D}^{1+k}} \frac{1}{\delta^5}$. Crucially observe that our choice of c and T satisfy the conditions imposed on $\lambda \geq 4b \geq 8c$ in Lemma G. So, for $t \in T_1$ the stochastic process $C(K_t)$ exactly satisfies the conditions given in Lemma G. We thus apply the modified Azuma's Inequality to the stochastic process in the chunk T_1 of size $\frac{1}{c}$ as follows:

$$
\begin{aligned}
P(\mathcal{X}_1^{\mathcal{C}} | \mathcal{X}_0) &= P\bigg(C(K_{m_1}) \geq C(K_0) + \frac{1}{Tc}\bigg) && (135)\\
&\leq \exp\bigg\{ - \frac{(\frac{1}{Tc})^2}{2\frac{1}{c}c^2} \bigg\} && (136)\\
&= \exp\bigg\{ - \frac{\frac{1}{T^2c^2}}{2c} \bigg\} && (137)\\
&= \exp\bigg\{ - \frac{1}{2T^2c^3} \bigg\} && (138)\\
&= \exp\Bigg\{ - \frac{1}{2\bigg(\frac{\mathrm{D}^{3+k}}{\epsilon}\rho^2\phi\delta^7 \log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\bigg)^2 \bigg(\frac{\epsilon}{\mathrm{D}^{2+k}}\frac{1}{\rho\delta^5\phi}\bigg)^3} \Bigg\} && (139)\\
&= \exp\bigg\{ - \frac{1}{2\frac{\mathrm{D}^{6+2k}}{\epsilon^2}(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}))^2 \frac{\epsilon^3}{\mathrm{D}^{6+3k}}\frac{\rho}{\delta\phi}} \bigg\} && (140)\\
&= \exp\bigg\{ - \frac{\phi\delta}{2(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}))^2 \frac{\epsilon}{\mathrm{D}^k}\rho} \bigg\} && (141)\\
&= \exp\bigg\{ - \frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho} \bigg\} && (142)\\
& && (143)
\end{aligned}
$$

Hence, we have that:

$$P(\mathcal{X}_1|\mathcal{X}_0) \geq \left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right) \tag{144}$$

Now, we observe that if we condition on $\bigcap_{j=0}^{i-1}\mathcal{X}_j$, then $\mathcal{E}_i$ is satisfied. So we have the necessary boundedness condition that $C(K_t) \leq C(K_0) + 2$ for all $t \in T_i$, and $\eta \leq \alpha \leq \alpha_{K_t}$ so we can write down Equations 107 through 116. This implies that can make the exact same argument to show the following statement:

$$\begin{aligned}
P(\mathcal{X}_i\Big|\bigcap_{j=0}^{i-1}\mathcal{X}_j) &= P(C(K_{m_i}) \leq C(K_0) + \frac{i}{Tc}\Big|\bigcap_{j=0}^{i-1}\mathcal{X}_j) && (145)\\
&\geq P(C(K_{m_i}) \leq C(K_{m_{i-1}}) + \frac{1}{Tc}\Big|\bigcap_{j=0}^{i-1}\mathcal{X}_j) && (146)\\
&\geq \left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right) && (147)
\end{aligned}$$

So, we apply Bayes Rule to show the following probability bound:

$$\begin{aligned}
P(\mathcal{E}) &\geq P(\bigcap_{i=1}^{Tc}\mathcal{E}_i) && (148)\\
&\geq P(\bigcap_{i=0}^{Tc}\mathcal{X}_i) && (149)\\
&= \prod_{i=0}^{Tc} P(\mathcal{X}_i\Big|\bigcap_{j=0}^{i-1}\mathcal{X}_j) && (150)\\
&\geq \prod_{i=1}^{Tc}\left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right) && (151)\\
&= \left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right)^{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} && (152)
\end{aligned}$$

Observe that this is a strong bound for any fixed $k > 0$ (strict inequality), in the following sense where we hold any two parameters constant and take the limit with respect to the third:

$$\lim_{d\to\infty,\epsilon,C(K_0)}\left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right)^{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} = 1 \tag{153}$$

$$\lim_{d,\epsilon\to 0,C(K_0)}\left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right)^{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} = 1 \tag{154}$$

$$\lim_{d,\epsilon,C(K_0)\to\infty}\left(1 - \exp\left\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\big)^2\rho}\right\}\right)^{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} = 1 \tag{155}$$

We now cite Bernoulli's Inequality that $(1+y)^r \geq 1+ry$ if $y \geq -1$ and $r \geq 1$. By a change of variable $y=-s$ this is equivalent to $(1-s)^r \geq 1-sr$ if $s \leq 1$ and $r \geq 1$. Setting $s = \exp\Bigg\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}\Bigg\}$, we see that $s \leq 1$, and setting $r = \mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ we have that $r \geq 1$. Hence:

$$
\begin{aligned}
P(\mathcal{E}) &\geq (1-s)^r && (156)\\
&\geq 1-sr && (157)\\
&= 1-\exp\Bigg\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}\Bigg\}\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) && (158)
\end{aligned}
$$

So, in order to ensure that $P(\mathcal{E}) \geq 1-p$, we need $\exp\Bigg\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}\Bigg\}\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) \leq p$. We show the following simplification:

$$
\begin{aligned}
&\exp\Bigg\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}\Bigg\}\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) \leq p && (159)\\
&\exp\Bigg\{-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}\Bigg\} \leq \frac{p}{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} && (160)\\
&-\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho} \leq \log\left(\frac{p}{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}\right) && (161)\\
&\frac{\mathrm{D}^k\,\phi\delta}{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho} \geq \log\left(\frac{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right) && (162)\\
&\mathrm{D}^k \geq \frac{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}{\phi\delta}\log\left(\frac{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right) && (163)\\
&k \geq \frac{\log\left\{\frac{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}{\phi\delta}\log\left(\frac{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(\mathrm{D})} && (164)
\end{aligned}
$$

So, we need:

$$
\begin{aligned}
k &\geq \frac{\log\left\{\frac{2\epsilon\Big(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\Big)^2\rho}{\phi\delta}\log\left(\frac{\mathrm{D}\,\rho\delta^2\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(\mathrm{D})} && (165)\\
&\approx \mathcal{O}\left(\frac{\log\left\{\epsilon\Big(\log(\frac{1}{\epsilon})\Big)^2\log\big(\frac{\mathrm{D}\log(\frac{1}{\epsilon})}{p}\big)\right\}}{\log(\mathrm{D})}\right) && (166)
\end{aligned}
$$

to ensure that $P(\mathcal{E}) \geq 1-p$.

This completes the proof. $\square$

Discussion: The dependence on d is a bit worse here than in the other writeup because here I'm making my step size super small instead of using zero order information. I don't have a complete proof of that argument since I haven't been able to prove the right version of Azuma's Inequality. But this right here should be a complete proof. Additionally, observe that the dependence on ρ, β, δ is pretty bad, but can be significantly improved by having just a single ϕ or something. Comparing this dependence on these quantities to the one in the previous writeup looks fairly similar.

We have that if $\eta = \frac{\epsilon}{d^{2+k}} \frac{1}{\rho^3 \beta^3} \frac{1}{\phi} \frac{1}{\delta^4}$ then:

$$P(\mathcal{E}) \geq \left(1 - \exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\}\right)^{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} \tag{167}$$

We now cite Bernoulli's Inequality that $(1+y)^r \geq 1 + ry$ if $y \geq -1$ and $r \geq 1$. By a change of variable $y = -s$ this is equivalent to $(1-s)^r \geq 1 - sr$ if $s \leq 1$ and $r \geq 1$. Setting $s = \exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\}$, we see that $s \leq 1$, and setting $r = d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})$ we have that $r \geq 1$. Hence:

$$\begin{aligned} P(\mathcal{E}) &\geq (1-s)^r && (168)\\ &\geq 1 - sr && (169)\\ &= 1 - \exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\} d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) && (170)\end{aligned}$$

So, in order to ensure that $P(\mathcal{E}) \geq 1-p$, we need $\exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\} d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) \leq p$. We show the following simplification:

$$\exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\} d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon}) \leq p \tag{171}$$

$$\exp\left\{ - \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}\right\} \leq \frac{p}{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})} \tag{172}$$

$$- \frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2} \leq \log\left(\frac{p}{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}\right) \tag{173}$$

$$\frac{d^k \rho\beta\phi\delta}{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2} \geq \log\left(\frac{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right) \tag{174}$$

$$d^k \geq \frac{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}{\rho\beta\phi\delta}\log\left(\frac{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right) \tag{175}$$

$$k \geq \frac{\log\left\{\frac{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}{\rho\beta\phi\delta}\log\left(\frac{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(d)} \tag{176}$$

So, we need:

$$k \geq \frac{\log\left\{\frac{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}{\rho\beta\phi\delta}\log\left(\frac{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(d)} \tag{177}$$

$$\approx \mathcal{O}\left(\frac{\log\left\{\epsilon\left(\log(\frac{1}{\epsilon})\right)^2\log\left(\frac{d\log(\frac{1}{\epsilon})}{p}\right)\right\}}{\log(d)}\right) \tag{178}$$

to ensure that $P(\mathcal{E}) \geq 1-p$. First observe that increasing d and decreasing ϵ lead to a decrease in the required value of k for a fixed p when the RHS is positive. Also, this calculation shows that the bound is definitely stronger than what I thought it was, because when $\frac{1}{\epsilon} \geq \log(\frac{d}{p})$ this suggests that we can get away with a (slightly) negative number for k instead of requiring $k > 0$. But of course then p shows up in our number of iterations. Also, I need $k \geq 0$ as a technical condition in the proof, to apply Azuma's Inequality, although it might be possible to do away with this with a different proof. Either way, we can always write $k \geq \max\{0, \frac{\log\left\{\frac{2\epsilon\left(\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})\right)^2}{\rho\beta\phi\delta}\log\left(\frac{d\frac{\rho\beta\delta}{2}\log(\frac{2(C(K_0)-C(K^*))}{\epsilon})}{p}\right)\right\}}{\log(d)}\}$ to ensure that $P(\mathcal{E}) \geq 1-p$.