A. Proof of Lemma 3.4

Proof. For simplicity we discuss a gaussian smoothing, (Duchi et al., 2012), instead of the uniform ball smoothing as discussed in Definition 3.1. The proof for the smoothing used in our paper is similar. Note that uniform ball smoothing with parameter δ is equivalent[2] to (zero mean) gaussian smoothing with a covariance matrix of $(\delta/\sqrt{d})I_d$, where I_d is the d dimensional identity matrix. Thus the δ-smoothed version with the gaussian kernel is defined as follows:

$$\hat{f}_\delta(\mathbf{x}) \;=\; \mathbf{E}_{\mathbf{u}}[f(\mathbf{x}+\delta\mathbf{u})], \quad \text{where } \mathbf{u} \sim \mathcal{N}(0, \frac{1}{\sqrt{d}} I_d) \;.$$

here $\mathcal{N}(0, I_d)$ denotes a zero mean normal distribution with the identity covariance matrix. For ease of notation define $\tilde{f}$ as follows:

$$\tilde{f}_\delta(x) \;=\; \hat{f}_{\sqrt{d}\delta}(\mathbf{x}) \;=\; \mathbf{E}_{\mathbf{u}}[f(\mathbf{x}+\delta\mathbf{u})], \quad \text{where } \mathbf{u} \sim \mathcal{N}(0, I_d) \;.$$

Since Lemma 3.4 states that f is $(\sqrt{d}, 0.5)$-nice, we need to prove the following:

1. **Centering property:** For every $\delta > 0$, and every $\mathbf{x}^*_\delta \in \arg\min_{\mathbf{x}\in\mathcal{K}} \tilde{f}_\delta(\mathbf{x})$, there exists $\mathbf{x}^*_{\delta/2} \in \arg\min_{x\in\mathcal{K}} \tilde{f}_{\delta/2}(\mathbf{x})$, such that $\|\mathbf{x}^*_\delta - \mathbf{x}^*_{\delta/2}\| \le \delta/2$.
2. **Local strong convexity of the smoothed versions:** For every $\delta > 0$, let $r_\delta = 3\delta$, and denote $\mathbf{x}^*_\delta = \arg\min_{\mathbf{x}\in\mathcal{K}} \tilde{f}_\delta(\mathbf{x})$, then over $B_{r_\delta}(\mathbf{x}^*_\delta)$ the function $\tilde{f}_\delta(\mathbf{x})$ is σ-strongly-convex.

Remark: We will abuse notation for the rest of the proof and relate to $\tilde{f}_\delta$ as the smoothed version.

In the case of the quadratic function with a valley, we may calculate the smoothed versions explicitly:

$$\begin{aligned}
\tilde{f}_\delta(\mathbf{x}) \;&=\; \mathbf{E}_{\mathbf{u}}[\frac{\|\mathbf{x}+\delta\mathbf{u}\|^2}{2} - \alpha e^{-\frac{(x_1+\delta u_1-1)^2}{2\lambda^2}}] \\
&=\; \frac{1}{2}\mathbf{E}_{\mathbf{u}}[\|\mathbf{x}\|^2 + 2\delta\langle\mathbf{u},\mathbf{x}\rangle + \delta^2\|u\|^2] - \alpha\mathbf{E}_{\mathbf{u}}[e^{-\frac{(x_1+\delta u_1-1)^2}{2\lambda^2}}] \\
&=\; \frac{\|\mathbf{x}\|^2}{2} - \alpha\mathbf{E}_{u_1}[e^{-\frac{(x_1+\delta u_1-1)^2}{2\lambda^2}}] + C \\
&=\; \frac{\|\mathbf{x}\|^2}{2} - \alpha\sqrt{\frac{\lambda^2}{\lambda^2+\delta^2}} e^{-\frac{(x_1-1)^2}{2(\lambda^2+\delta^2)}} + C \;,
\end{aligned}$$

here $C = \mathbf{E}_{\mathbf{u}}[\delta^2\|\mathbf{u}\|^2]$, and we used $\mathbf{E}_{\mathbf{u}}[\mathbf{u}] = 0$, we also used the fact that a convolution between two gaussian kernels is a gaussian kernel with the sum of variances, see (Oppenheim & Willsky, 1997) (note that smoothing with gaussian perturbation is equivalent to convolution with a gaussian kernel). It therefore follows that the smoothed version of a quadratic function with a valley is the same quadratic function with a wider valley, and a smaller amplitude.

The last equation implies that it is sufficient to prove that the 1-dim function: $f(x) = \frac{x^2}{2} - \alpha e^{-\frac{(x-1)^2}{2\lambda^2}}$, is σ-nice, where the smoothed versions are:

$$\tilde{f}_\delta(x) = \frac{x^2}{2} - \alpha\sqrt{\frac{\lambda^2}{\tilde{\delta}^2}} e^{-\frac{(x-1)^2}{2\tilde{\delta}^2}} \;, \tag{4}$$

and we denote $\tilde{\delta}^2 = \delta^2 + \lambda^2$. It is possible to validate that the hardest case is when α is the largest possible. We therefore assume from now on that $\alpha = 1/200$.

Step 0: Here we show that if $\lambda \ge 1/10$ then $f(x)$ is $1/2$-strongly-convex and $3/2$-smooth. Deriving $f(x)$ twice we get:

$$f''(x) = 1 + g''(x) = 1 - \frac{1}{200\lambda^2} e^{-\frac{(x-1)^2}{2\lambda^2}} \left(\frac{(x-1)^2}{\lambda^2} - 1\right) \;.$$

[2] Equivalence in the sense that in both cases the bias between the δ-smoothed version and the original function is bounded by $L\delta$

It can be shown that $g''(x)$ has one global maxima at $x=1$ and two global minima at $x=1\pm\sqrt{3}\lambda$, and therefore $\forall x$:

$$-\frac{2e^{-3/2}}{200\lambda^2}=g''(1+\sqrt{3}\lambda)\leq g''(x)\leq g''(1)=\frac{1}{200\lambda^2}\ ,$$

Using $\lambda\geq 1/10$ we conclude that:

$$1/2\leq f''(x)\leq 3/2\ ,$$

which establishes the strong-convexity and smoothness. From now on we assume that $\lambda\leq 1/10$, and establish the "niceness" of f.

Step 1: Here we show the following to hold:

$$0\leq x^*_\delta\leq\frac{1}{5}\max\{\delta,\lambda\} \tag{5}$$

We will require the following lemma (see proof in Section A.1):

Lemma A.1. *Let $\theta\geq 0$, $\alpha\in(0,\frac{1}{25}]$, and $m(x)=\frac{x^2}{2}-\alpha e^{-\frac{(x-1)^2}{2\theta^2}}$. Denoting $x^*=\arg\min_{x\in\mathbb{R}}m(x)$ then:*

$$0\leq x^*\leq 2\sqrt{\alpha}\theta$$

Applying the above lemma to the smoothed version appearing in Equation (4), we conclude that for $x^*_\delta=\arg\min_{x\in\mathbb{R}}\tilde{f}_\delta(x)$ the following applies:

$$0\leq x^*_\delta\leq\frac{2}{\sqrt{200}}\left(\frac{\lambda^2}{\tilde{\delta}^2}\right)^{1/4}\sqrt{\tilde{\delta}^2}=\frac{1}{5\sqrt{2}}\left(\lambda^2(\lambda^2+\delta^2)\right)^{1/4}\ .$$

Note that the above means $x^*_\delta\leq\frac{1}{5}\max\{\delta,\lambda\}$.

Step 2: Here we show that the smoothed versions are 0.5-strongly-convex in a 3δ radius around the global minima. Note that it suffices to show that $\forall x\in x^*_\delta+[-3\delta,3\delta]$, the following holds:

$$-g''_\delta(x)=\frac{1}{200\tilde{\delta}^2}\sqrt{\frac{\lambda^2}{\tilde{\delta}^2}}\left(\frac{(x-1)^2}{\tilde{\delta}^2}-1\right)e^{-\frac{(x-1)^2}{2\tilde{\delta}^2}}\leq 0.5\ .$$

In the previous paragraph we have shown that $0\leq x^*_\delta\leq\frac{1}{5}\max\{\delta,\lambda\}$, since $\max\{\delta,\lambda\}\leq\tilde{\delta}$ it suffices to prove that the above holds $\forall x\in(-\infty,\frac{\tilde{\delta}}{2}+3\tilde{\delta}]$. Now suppose that there exists $x\in(-\infty,\frac{\tilde{\delta}}{2}+3\tilde{\delta}]$ such that $(x-1)^2\leq\frac{1}{9}$, then it follows that $\tilde{\delta}\geq 4/21$. In this case:

$$\begin{aligned}-g''_\delta(x)&=\frac{1}{200\tilde{\delta}^2}\sqrt{\frac{\lambda^2}{\tilde{\delta}^2}}\left(\frac{(x-1)^2}{\tilde{\delta}^2}-1\right)e^{-\frac{(x-1)^2}{2\tilde{\delta}^2}}\\&\leq\frac{1}{200\tilde{\delta}^2}\max_{z\in\mathbb{R}}(z-1)\,e^{-z/2}\\&\leq\frac{1}{200\tilde{\delta}^2}2e^{-3/2}<0.5\ .\end{aligned}$$

where in the first inequality we used $\lambda^2\leq\tilde{\delta}^2$, in the second inequality we used $\max_{z\in\mathbb{R}}(z-1)e^{-z/2}=2e^{-3/2}$, later we used used $\tilde{\delta}\geq 4/21$.

Consider the other case, in which $\forall x\in(-\infty,\frac{\tilde{\delta}}{2}+3\tilde{\delta}]$ it holds that $(x-1)^2\geq 1/9$, then:

$$\begin{aligned}-g''_\delta(x)&=\frac{1}{200\tilde{\delta}^2}\sqrt{\frac{\lambda^2}{\tilde{\delta}^2}}\left(\frac{(x-1)^2}{\tilde{\delta}^2}-1\right)e^{-\frac{(x-1)^2}{2\tilde{\delta}^2}}\\&=\frac{1}{200}\sqrt{\frac{\lambda^2}{\tilde{\delta}^2}}\frac{1}{(x-1)^2}\frac{(x-1)^2}{\tilde{\delta}^2}\left(\frac{(x-1)^2}{\tilde{\delta}^2}-1\right)e^{-\frac{(x-1)^2}{2\tilde{\delta}^2}}\\&\leq\frac{9}{200}\max_{y\geq 0}y(y-1)e^{-y/2}<0.5\ .\end{aligned}$$

where we used $\lambda^2\leq\tilde{\delta}^2$, also $(x-1)^2\geq 1/9$, and finally we used $\max_{y\geq 0}y(y-1)e^{-y/2}\leq 1.665$.

Step 3: Letting $x^* = \arg\min_x f(x)$, we show here that $\forall \delta \leq \lambda,\ |x^*_\delta - x^*| \leq \frac{\delta}{3}$.

First note that we have already shown that $\tilde{f}_\delta$ is 0.5-strongly convex in the section $(-\infty, 3.5\tilde{\delta})$ (see Step 2), this section contains x^*_δ and x^*, since $\forall \delta < \lambda$ then $x^*_\delta \leq \lambda/5 \leq \tilde{\delta}/5$ (see Step 1). The strong convexity implies:

$$\begin{aligned}
|x^* - x^*_\delta| &\leq \sqrt{\frac{2}{0.5}}\sqrt{\tilde{f}_\delta(x^*) - \tilde{f}_\delta(x^*_\delta)} \\
&\leq 2\sqrt{f(x^*) - f(x^*_\delta) + \left(\tilde{f}_\delta(x^*) - f(x^*)\right) - \left(\tilde{f}_\delta(x^*_\delta) - f(x^*_\delta)\right)} \\
&\leq 2\sqrt{\max_{x\in[0,\lambda/5]} |\tilde{f}_\delta(x) - f(x)|}\,. \qquad (6)
\end{aligned}$$

in the second inequality we used $x^* = \arg\min_{x\in\mathbb{R}} f(x)$. Now fix $x \in [0, \lambda/5]$ and lets us denote $b = \frac{(x-1)^2}{\lambda^2}$, also denote $z = \frac{\delta^2}{\lambda^2} \leq 1$, using this notation we can write:

$$\begin{aligned}
200(f(x) - \tilde{f}_\delta(x)) &= \sqrt{\frac{\lambda^2}{\lambda^2+\delta^2}} e^{-\frac{(x-1)^2}{2(\lambda^2+\delta^2)}} - e^{-\frac{(x-1)^2}{2\lambda^2}} \\
&= \sqrt{\frac{1}{1+z}} e^{-\frac{b}{2(1+z)}} - e^{-b/2} = h'(z_0)z\,. \qquad (7)
\end{aligned}$$

for some $z_0 \in [0, z] \subseteq [0, 1]$, where we denote $h(z) = \sqrt{\frac{1}{1+z}} e^{-\frac{b}{2(1+z)}}$, deriving $h(z)$ we get:

$$\begin{aligned}
|h'(z)| &= |\frac{1}{2}(1+z)^{-5/2} e^{-\frac{b}{2(1+z)}}(b - z)| \\
&\leq \frac{1}{2} e^{-\frac{b}{2(1+z)}} b \\
&\leq \frac{1}{2} e^{-\frac{(49/100)^2}{\lambda^2}} \frac{1}{\lambda^2}\,.
\end{aligned}$$

where in the first inequality we $1 + z \geq 1$, and also $b - (1 + z) = \frac{(x-1)^2}{\lambda^2} - (1 + z) > 0;\ \forall z \in [0, 1], x \leq \frac{\lambda}{5} \leq \frac{1}{50}$. This is since $\frac{(x-1)^2}{\lambda^2} > \frac{(2\lambda/5-1)^2}{\lambda^2} \geq \frac{(49/50)^2}{0.5^2} > 2 \geq 1 + z$. In the second inequality we used $(49/50)^2 \leq (x - 1)^2 \leq 1$, and $\frac{1}{1+z} \geq \frac{1}{2}$. Plugging the above bound on $|h'(z)|$ into Equation (7) and substituting $z = \delta^2/\sigma^2$, we get:

$$|f(x) - \tilde{f}_\delta(x)| \leq \frac{1}{400} e^{-\frac{(49/100)^2}{\lambda^2}} \frac{1}{\lambda^4}\delta^2 \leq \frac{1}{40}\delta^2\,.$$

where we used $\max_\lambda e^{-\frac{(49/100)^2}{\lambda^2}} \frac{1}{\lambda^4} < 10$. Plugging the above into Equation (6) we conclude:

$$|x^* - x^*_\delta| \leq \frac{\delta}{3}, \qquad \forall \delta \leq \lambda \qquad (8)$$

Conclusion: In Step 2 we have shown that $\tilde{f}_\delta$ is 0.5-strongly-convex a radius of 3δ around x^*_δ. We are left to show that $\forall \delta > 0,\ |x^*_\delta - x^*_{\delta/2}| \leq \delta/2$. According to Equation (5), $\forall \delta \geq \lambda$ we have:

$$|x^*_\delta - x^*_{\delta/2}| \leq \delta/5\,.$$

Equation (8) implies that $\forall \delta \leq \lambda$:

$$|x^*_\delta - x^*_{\delta/2}| \leq |x^* - x^*_\delta| + |x^* - x^*_{\delta/2}| \leq \delta/3 + (\delta/2)/3 = \delta/2\,.$$

Thus $f(x)$ is $(\sqrt{d}, 0.5)$-nice. □

A.1. Proof of Lemma A.1

Proof. It is can be noticed that x^* must be positive (using the symmetry of the quadratic function around 0, and the "valley" function around 1). Now, note that the optimality of x^* means:

$$\frac{(x^*)^2}{2} - \alpha \leq \frac{(x^*)^2}{2} - \alpha e^{-\frac{(x^*-1)^2}{2\theta^2}} = m(x) \leq m(0) \leq 0 \,.$$

and therefore, we always have:

$$x^* \leq 2\sqrt{\alpha}$$

this establishes the lemma for $\theta \geq 1$.

Now let $\theta \leq 1$, suppose by contradiction that $|x^* - 1| \leq \theta\sqrt{2\max\{0, \log 1/2\theta^2\}}$, it therefore follows that:

$$m(x^*) = \frac{(x^*)^2}{2} - \alpha e^{-\frac{(x^*-1)^2}{2\theta^2}} \geq \frac{(1 - \theta\sqrt{2\max\{0, \log 1/2\theta^2\}})^2}{2} - \alpha \geq (1 - 0.65)^2 - \frac{1}{25} \geq 0 > m(0) \,.$$

which is a contradiction since x^* is the global optimum. Note that we used $\alpha \leq 1/25$, $\max_{\theta\in[0,1]} \theta\sqrt{2\max\{0, \log 1/2\theta^2\}} \leq 0.65$, and $m(0) < 0$. It therefore follows that for $\theta \leq 1$, we always have $|x^* - 1| \geq \theta\sqrt{2\max\{0, \log 1/2\theta^2\}}$, thus:

$$0 > m(0) \geq m(x^*) = \frac{(x^*)^2}{2} - \alpha e^{-\frac{(x^*-1)^2}{2\theta^2}} \geq \frac{(x^*)^2}{2} - \alpha 2\theta^2 \,,$$

and therefore $x^* \leq 2\sqrt{\alpha}\theta$ for $\theta \in [0, 1]$, which establishes the lemma. □

B. Proof of Theorem 5.1

Notice that at each epoch m of GradOpt$_V$, it initiates Suffix-SGD with a gradient oracle SGO$_V(\cdot, \delta_m)$. According to Lemma 3.3, SGO$_V(\cdot, \delta_m)$ produces an unbiased and dC/δ_m-bounded estimates for the gradients of $\hat{f}_{\delta_m}$, thus in the analysis of each epoch we can use Corollary 4.1 for $\hat{f}_{\delta_m}$, taking $G = dC/\delta_m$.

Following is our key Lemma:

Lemma B.1. *Consider M, $\mathcal{K}_m$ and $\bar{\mathbf{x}}_{m+1}$ as defined in Algorithm 3. Also denote by $\mathbf{x}^*_m$ the minimizer of $\hat{f}_{\delta_m}$ in $\mathcal{K}$. Then the following holds for all $1 \leq m \leq M$ w.p.$\geq 1 - p$:*

1. *The smoothed version $\hat{f}_{\delta_m}$ is σ-strongly convex over $\mathcal{K}_m$, and $\mathbf{x}^*_m \in \mathcal{K}_m$.*
2. *Also, $\hat{f}_{\delta_m}(\bar{\mathbf{x}}_{m+1}) - \hat{f}_{\delta_m}(\mathbf{x}^*_m) \leq \sigma\delta^2_{m+1}/8$*

The proof of Lemma B.1 is similar to the proof of Lemma 4.1 given in Section 4.1, we therefore omit the details.

We are now ready to prove Theorem 5.1:

Proof of Theorem 5.1. Let $\bar{\mathbf{x}}_{M+1}$ be the output of Algorithm 3. Similarly to the proof of Theorem 4.1, we can show that for every $\mathbf{x} \in \mathcal{K}$:

$$f(\bar{\mathbf{x}}_{M+1}) - f(\mathbf{x}) \leq \varepsilon$$

Let T_{total}, be the total number of queries made by Algorithm 3, then we have:

$$\begin{aligned}
T_{\text{total}} &\leq \sum_{m=1}^{M} \frac{12480 d^2 C^2}{\sigma \varepsilon_m \delta_m^2} \log \Gamma \\
&\leq \sum_{m=1}^{M} \frac{12480 d^2 C^2}{\sigma (\sigma \delta_m^2 / 32) \delta_m^2} \log \Gamma \\
&\leq \frac{4 \cdot 10^5 d^2 C^2 \log \Gamma}{\sigma^2} \sum_{i=1}^{M} \frac{8^{i-1}}{\delta_1^4} \\
&\leq \frac{6 \cdot 10^4 d^2 C^2 \log \Gamma}{\sigma^2} \frac{8^M}{\delta_1^4} \\
&\leq \frac{6 \cdot 10^4 d^2 C^2 \log \Gamma}{\sigma^2} \max\{256 L^4, \sigma^2/4\} \frac{1}{\varepsilon^4}
\end{aligned}$$

here we used the notation:

$$\begin{aligned}
\Gamma &:= \frac{2M}{p} + 2 \log(12480 d^2 C^2 / \sigma \varepsilon_M \delta_M^2) \\
&\leq \frac{2M}{p} + 2 \log(4 \cdot 10^5 d^2 C^2 \max\{256 L^4, \frac{\sigma^2}{4}\} / \sigma^2 \varepsilon^4)
\end{aligned}$$

□