Improved Performance of Stochastic Gradients with Gaussian Smoothing

A. Starnes* C. Webster*

Appendix

Table of Contents

*Lirio AI Research, Lirio, Inc., Knoxville, TN 37923 (astarnes@lirio.com, cwebster@lirio.com).

A Proofs of Background Results

We begin this section by stating the definitions of convexity and L-smoothness as well as mentioning some key identities that we use throughout our proofs, then we provide the proofs of Lemmas 2.

Definition A.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$.*

(a) We say f is convex, if for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}).$$

(b) We say f is γ-strongly convex if

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{\gamma}{2}t(1-t)\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

(c) We say f is L-smooth, if for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle| \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

As in [5], it is often convenient to represent the gradient of the smoothed function as an integral difference[1]

$$\nabla f_\sigma(\boldsymbol{x}) = \frac{2}{\pi^{\frac{d}{2}}\sigma}\int_{\mathbb{R}^d}\Big(f(\boldsymbol{x} + \sigma\boldsymbol{u}) - f(\boldsymbol{x})\Big)\boldsymbol{u}e^{-\|\boldsymbol{u}\|^2}\, d\boldsymbol{u}.$$

Another convenience from [5], is rewriting the gradient of the original function in an integral

$$\nabla f(\boldsymbol{x}) = \frac{1}{\pi^{\frac{d}{2}}}\int_{\mathbb{R}^d}\langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle\boldsymbol{u}e^{-\|\boldsymbol{u}\|^2}\, d\boldsymbol{u}.$$

We state Lemma 4 from [5], which we use to bound the norm of the original function by the norm of the smoothed function.

Lemma A.1 (Lemma 4 [5])**.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\sigma \geq 0$*

$$\|\nabla f(\boldsymbol{x})\|^2 \leq 2\|\nabla f_\sigma(\boldsymbol{x})\|^2 + \frac{L^2\sigma^2}{4}(6+d)^3.$$

Now, we prove Lemma 2.

Proof of Lemma 2. The first result can be shown just by a modification of the proof of Lemma A.1.

[1] This comes from the fact that $\int_{\mathbb{R}^d} f(\boldsymbol{x})\boldsymbol{u}e^{-\|\boldsymbol{u}\|^2}\, d\boldsymbol{u} = 0$.

Observe

$$\begin{aligned}
\|\nabla f_\sigma(\boldsymbol{x})\|^2 &= \left\| \frac{2}{\pi^{\frac{d}{2}}\sigma} \int_{\mathbb{R}^d} \big(f(\boldsymbol{x}+\sigma\boldsymbol{u}) - f(\boldsymbol{x})\big)\boldsymbol{u} e^{-\|\boldsymbol{u}\|^2} du \right\|^2 \\
&= \left\| \frac{2}{\pi^{\frac{d}{2}}\sigma} \int_{\mathbb{R}^d} \Big[\big(f(\boldsymbol{x}+\sigma\boldsymbol{u}) - f(\boldsymbol{x}) - \sigma\langle\nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle\big) + \sigma\langle\nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle\Big]\boldsymbol{u} e^{-\|\boldsymbol{u}\|^2} d\boldsymbol{u} \right\|^2 \\
&\leq \frac{8}{\pi^d\sigma^2} \int_{\mathbb{R}^d} \big(f(\boldsymbol{x}+\sigma\boldsymbol{u}) - f(\boldsymbol{x}) - \sigma\langle\nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle\big)^2 \|\boldsymbol{u}\|^2 e^{-\|\boldsymbol{u}\|^2} d\boldsymbol{u} + 2\|\nabla f(\boldsymbol{x})\|^2 \\
&\leq \frac{2L^2\sigma^2}{\pi^d} \int_{\mathbb{R}^d} \|\boldsymbol{u}\|^6 e^{-\|\boldsymbol{u}\|^2} d\boldsymbol{u} + 2\|\nabla f(\boldsymbol{x})\|^2 \\
&\leq \frac{L^2\sigma^2}{4}(6+d)^3 + 2\|\nabla f(\boldsymbol{x})\|^2,
\end{aligned}$$

where the last inequality comes from Lemma 1 of [5] which states that for $p \geq 2$

$$\frac{1}{\pi^{p/2}} \int_{\mathbb{R}^d} \|\boldsymbol{u}\|^p e^{-\|\boldsymbol{u}\|^2}\, d\boldsymbol{u} \leq \left(\frac{p+d}{2}\right)^{p/2}.$$

For the second part of the lemma, we will assume, without loss of generality, $\sigma \leq \tau$. By Lemma 3 from [5], for any $\eta > 0$,

$$\|\nabla f_\eta(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| \leq L\eta \left(\frac{3+d}{2}\right)^{3/2}.$$

By Lemma 2.8 of [6], if $\eta = \sqrt{\tau^2 - \sigma^2}$, then $(f_\sigma)_\eta = f_\tau$. So,

$$\|\nabla f_\tau(\boldsymbol{x}) - \nabla f_\sigma(\boldsymbol{x})\| = \|\nabla (f_\sigma)_\eta - \nabla f_\sigma(\boldsymbol{x})\| \leq L\eta \left(\frac{3+d}{2}\right)^{3/2} = L\sqrt{\tau^2 - \sigma^2} \left(\frac{3+d}{2}\right)^{3/2}.$$

□

B Proof of GSmoothSGD Convergence

The proof of convergence of GSmoothSGD adapts the standard proof of SGD (see, e.g., [1]). Based on the background results, we have two ways to change between f_σ and f_τ (where τ could be 0): switch at the function level (Lemma 1) or switch at the gradient level (Lemma 2). Our proof of SGD carefully chooses where to change between smoothing values in order to keep the added smoothing bound as small as possible.

Proof of Theorem 1. Since $E(\|\nabla f_k\|^2) \leq \lambda$, there exists λ_t so that $E(\|\nabla f_{k_t,\sigma_{t+1}}\|^2) \leq \lambda_{t+1}$ (see Lemma 2). We begin by repeating typical analysis done in the SGD proof:

$$\begin{aligned} f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1}) &\overset{L-\text{smooth}}{\leq} f_{\sigma_{t+1}}(\boldsymbol{x}_t) + \langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t \rangle + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 \\ &= f_{\sigma_{t+1}}(\boldsymbol{x}_t) + \langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), -\eta \nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t) \rangle + \frac{L\eta^2}{2}\|\nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t)\|^2 \\ &= f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \eta \langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t) \rangle + \frac{L\eta^2}{2}\|\nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t)\|^2 \end{aligned}$$

Taking the expectation and using the gradient bound gives

$$E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) \leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - \eta E(\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t) \rangle) + \frac{L\eta^2}{2}\lambda_{t+1}.$$

Repeating the regular SGD proof but for $f_{\sigma_{t+1}}$, we have

$$E\big(\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t) \rangle\big) = E\left(\big\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), E(\nabla f_{k,\sigma_{t+1}}(\boldsymbol{x}_t)|k_t = k) \big\rangle\right) = E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2).$$

This means that

$$E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) \leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - \eta E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L\eta^2}{2}\lambda_{t+1}.$$

Rearranging gives

$$\begin{aligned} \eta E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) &\leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t) - f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) + \frac{L\eta^2}{2}\lambda_{t+1} \\ &\leq E(f_{\sigma_t}(\boldsymbol{x}_t) - f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) + \frac{L\eta^2}{2}\lambda_{t+1} + \frac{Ld}{4}|\sigma_{t+1}^2 - \sigma_t^2|. \end{aligned}$$

Summing over the steps shows (where $\sigma_0 = 0$ for notational convenience)

$$\begin{aligned} \eta \sum_{t=0}^{T-1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) &\leq \sum_{t=0}^{T-1} E(f_{\sigma_t}(\boldsymbol{x}_t) - f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) + \frac{L\eta^2}{2}\sum_{t=0}^{T-1}\lambda_{t+1} + \frac{Ld}{4}\sum_{t=0}^{T-1}|\sigma_{t+1}^2 - \sigma_t^2| \\ &= E(f_{\sigma_0}(\boldsymbol{x}_0) - f_{\sigma_T}(\boldsymbol{x}_T)) + \frac{L\eta^2}{2}\sum_{t=0}^{T-1}\lambda_{t+1} + \frac{Ld}{4}\sum_{t=0}^{T-1}|\sigma_{t+1}^2 - \sigma_t^2| \end{aligned}$$

We know that $f^* \leq f_{\sigma_T}(\boldsymbol{x}_T)$, so

$$\begin{aligned}\eta \sum_{t=0}^{T-1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) &\leq E(f(\boldsymbol{x}_0) - f^*) + \frac{L\eta^2}{2}\sum_{t=1}^{T}\lambda_t + \frac{Ld}{4}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2| \\ &= f(\boldsymbol{x}_0) - f^* + \frac{L\eta^2}{2}\sum_{t=1}^{T}\lambda_t + \frac{Ld}{4}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2|.\end{aligned}$$

Taking the average gives

$$\begin{aligned}\sum_{t=0}^{T-1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) &\leq \frac{f(\boldsymbol{x}_0) - f^*}{T\eta} + \frac{L\eta}{2T}\sum_{t=1}^{T}\lambda_t + \frac{Ld}{4T\eta}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2| \\ &\overset{\eta<\frac{1}{L}}{\leq} \frac{f(\boldsymbol{x}_0) - f^*}{T\eta} + \frac{1}{2T}\sum_{t=1}^{T}\lambda_t + \frac{d}{4T\eta^2}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2|.\end{aligned}$$

From Lemma A.1 and using $L < \frac{1}{\eta}$, we have

$$\begin{aligned}\sum_{t=0}^{T-1} E(\|\nabla f(\boldsymbol{x}_t)\|^2) &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left(2E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L^2\sigma_{t+1}^2}{4}(6+d)^3\right) \\ &\leq \frac{2(f(\boldsymbol{x}_0) - f^*)}{T\eta} + \frac{1}{T}\sum_{t=1}^{T}\lambda_t + \frac{d}{2T\eta^2}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2| + \frac{(6+d)^3}{4T\eta^2}\sum_{t=1}^{T}\sigma_t^2.\end{aligned}$$

Finally, from Lemma 2, since $E(\|\nabla f_k\|^2) \leq \lambda$, we have that

$$\begin{aligned}E(\|\nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) &\leq 2E(\|\nabla f_{k_t}(\boldsymbol{x}_t)\|^2) + \frac{L^2(6+d)^3}{4}\sigma_{t+1}^2 \\ &\leq 2\lambda + \frac{L^2(6+d)^3}{4}\sigma_{t+1}^2 \\ &\leq 2\lambda + \frac{(6+d)^3}{4\eta^2}\sigma_{t+1}^2.\end{aligned}$$

Combining the previous two equations yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} E(\|\nabla f(\boldsymbol{x}_t)\|^2) &\leq \frac{2(f(\boldsymbol{x}_0) - f^*)}{T\eta} + \frac{1}{T}\sum_{t=1}^{T}\lambda_t + \frac{d}{2T\eta^2}\sum_{t=1}^{T}|\sigma_t^2 - \sigma_{t-1}^2| + \frac{(6+d)^3}{4T\eta^2}\sum_{t=1}^{T}\sigma_t^2 \\ &= \frac{2(f(\boldsymbol{x}_0) - f^*)}{T\eta} + 2\lambda + \frac{1}{2T\eta^2}\sum_{t=1}^{T}\left(|\sigma_t^2 - \sigma_{t-1}^2|d + \sigma_t^2(6+d)^3\right).\end{aligned}$$

□

We now turn towards proving Proposition 1. We begin with a formal definition of the basin of attraction.

Definition B.1. *Let $\sigma \geq 0$ and suppose f_σ has a unique minimizer $\boldsymbol{x}_\sigma^\star$. Let $(\sigma_n)_{n=1}^\infty \subseteq [\sigma, \infty)$ with $\sigma_t \to \sigma$, define*

$$\mathcal{B}\big(\sigma, (\sigma_t)_{t=1}^\infty; (k_t)_{t=1}^\infty\big) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}_t \to \boldsymbol{x}_\sigma^\star \text{ where } \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f_{k_t, \sigma_{t+1}}(\boldsymbol{x}_t) \text{ and } \boldsymbol{x}_0 = \boldsymbol{x}\}$$
$$\mathcal{B}\big(\sigma, (\sigma_t)_{t=1}^\infty\big) = \bigcap \big\{\mathcal{B}\big(\sigma, (\sigma_t)_{t=1}^\infty; (k_t)_{t=1}^\infty\big) : k_t \sim \textit{Unif}([K])\big\}$$
$$\mathcal{B}(\sigma) = \bigcup \big\{\mathcal{B}\big(\sigma, (\sigma_t)_{t=1}^\infty\big) : (\sigma_n)_{n=1}^\infty \subseteq [\sigma, \infty) \text{ with } \sigma_t \to \sigma\big\}$$

We call $\mathcal{B}(\sigma)$ the basin of attraction of $\boldsymbol{x}_\sigma^\star$.

While we assume that there are unique minimizers, the proof can be generalized to the setting where the minimizers form a connected, bounded set. Regardless, for simplicity, we assume uniqueness of the minimizers.

The proof of Proposition 1chains together the iterates from traveling though $\sigma_1, \sigma_2, \ldots, \sigma_N$ by running GSmoothSGD until the iterate is close enough to $\boldsymbol{x}_{\sigma_n}^\star$ so that it is also in the basin of attraction of $\boldsymbol{x}_{\sigma_{n+1}}^\star$. The following proposition provides the proof of this chaining idea. In particular, Proposition 1is actually a corollary of the following result, which provides how we transition from $\mathcal{B}(\sigma_n)$ to $\mathcal{B}(\sigma_{n+1})$.

Proposition B.1. *Assume f_σ has a unique minimizer $\boldsymbol{x}_\sigma^\star$ for $0 \leq \sigma \leq \Sigma$. Let $0 \leq \tau < \sigma \leq \Sigma$ be such that*

$$\boldsymbol{x}_\sigma^\star \in \mathcal{B}(\tau).$$

Suppose further that there is an open set around $\boldsymbol{x}_\sigma^\star$ contained in $\mathcal{B}(\tau)$. Then $\mathcal{B}(\sigma) \subseteq \mathcal{B}(\tau)$.

Proof. Let $\boldsymbol{x} \in \mathcal{B}(\sigma)$, so there exists $(\alpha_t)_{t=1}^\infty \subseteq [\sigma, \infty)$ such that

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta \nabla f_{k_t, \alpha_{t+1}}(\boldsymbol{x}_{t-1})$$

with $\boldsymbol{x}_0 = \boldsymbol{x}$ and $k_t \sim \text{Unif}([K])$ satisfies $\boldsymbol{x}_t \to \boldsymbol{x}_\sigma^\star$. Let $\boldsymbol{k} = (k_1, k_2, ...)$ be a realization of samples where $k_t \sim \text{Unif}([K])$. By assumption, there exists $R > 0$ such that

$$B(\boldsymbol{x}_\sigma^\star, R) \subseteq \mathcal{B}(\tau).$$

Since $\boldsymbol{x} \in \mathcal{B}(\sigma_n)$, there exists $(\alpha_t)_{t=1}^\infty \subseteq [\sigma, \infty)$ so that if

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta \nabla f_{k_t, \alpha_{t+1}}(\boldsymbol{x}_{t-1})$$

with $\boldsymbol{x}_0 = \boldsymbol{x}$ then $\boldsymbol{x}_t \to \boldsymbol{x}_\sigma^\star$. As such, there exists $N \in \mathbb{N}$ so that $\|\boldsymbol{x}_N - \boldsymbol{x}_\sigma^\star\| < R$. Hence,

$$\boldsymbol{x}_N \in B(\boldsymbol{x}_\sigma^\star, R) \subseteq \mathcal{B}(\tau).$$

This means that there exists $(\beta_t)_{t=1}^\infty \subseteq [\tau, \infty)$ with $\beta_t \to \tau$ such that if

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \eta \nabla f_{k_{t+N}, \beta_{t+1}}(\boldsymbol{y}_{t-1})$$

with $\boldsymbol{y}_0 = \boldsymbol{x}_N$ then $\boldsymbol{y}_t \to \boldsymbol{x}_\tau^\star$. Define

$$\gamma_t = \begin{cases} \alpha_t & \text{for } t = 1, ..., N-1 \\ \beta_t & \text{for } t \geq N \end{cases}$$

and

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta \nabla f_{k_t, \gamma_{t+1}}(\boldsymbol{x}_{t-1})$$

with $\boldsymbol{x}_0 = \boldsymbol{x}$. Then $\boldsymbol{x}_t \to \boldsymbol{x}_\tau^\star$ and $(\gamma_t)_{t=1}^\infty \subseteq [\tau, \infty)$ with $\gamma_t \to \tau$. Therefore,

$$x \in \mathcal{B}(\tau, (\gamma_t)_{t=1}^\infty; \boldsymbol{k}) \subseteq \mathcal{B}(\tau).$$

Since this can be done for any $\boldsymbol{k}$, the proof is complete. □

Practically, this does not provide a sequence of smoothing parameters because $(\beta_t)_{t=1}^\infty$ (and hence $(\gamma_t)_{t=1}^\infty$) depends on $\boldsymbol{k}$. Next, we state a more general corollary of the previous proposition compared to Proposition 1. In fact, Proposition 1is just a particular example of the following corollary that leverages the convexity of f_Σ to increase the initial basin of attraction.

Corollary B.1. *Suppose there exists an increasing sequence, $(\sigma_n)_{n=1}^N$, starting at 0 such that for $n = 2, \ldots, N$, f_{σ_n} has a unique minimizer $\boldsymbol{x}_{\sigma_n}^\star$ and there exists $R_n > 0$ so that*

$$B(\boldsymbol{x}_{\sigma_n}^\star, R_n) \subseteq \mathcal{B}(\sigma_{n-1}).$$

Then $\mathcal{B}(\Sigma) \subseteq \mathcal{B}(0)$.

Proof. The previous proposition shows that $\mathcal{B}(\sigma_n) \subseteq \mathcal{B}(\sigma_{n-1})$ for each $n = 2, ..., N$. Therefore,

$$\mathcal{B}(\Sigma) = \mathcal{B}(\sigma_1) \subseteq \mathcal{B}(\sigma_2) \subseteq \cdots \subseteq \mathcal{B}(\sigma_N) = \mathcal{B}(0).$$

□

Finally, we prove Proposition 1.

Proof of Proposition 1. Since each $f_{k,\Sigma}$ is convex and they share a minimizer, f_Σ is also convex with the same minimizer. As such, $\mathcal{B}(\Sigma) = \mathbb{R}^d$. By the previous corollary, we have

$$\mathbb{R}^d = \mathcal{B}(\Sigma) \subseteq \mathcal{B}(0) \subseteq \mathbb{R}^d.$$

□

C Proof of Convergence of GSmoothAdam

The proof that GSmoothAdam converges almost surely adapts the proof that Adam does from [2]. In order to prove Theorem 2, we needed to replicate almost the entirety of two of their other key results (Theorems 4 and 9). For this section, we adopt $\mathcal{F}_t$ as the notation for the sigma algebra generated by $\boldsymbol{x}_0, ..., \boldsymbol{x}_t$.

As the first step to proving Theorem 2, we provide the following smoothed version of Lemma 19 from [2]. This proof relies on Lemmas 16, 17 and 18 from [2], which do not need to be modified for smoothing. Lemma 16 provides a uniform bound on $\boldsymbol{m}_t$, $\boldsymbol{v}_t$, and the Adam update to $\boldsymbol{x}_t$, where the subtle difference between the smoothed and unsmoothed results are in this bound. In particular, the M is the original statement, which bounds $\|\nabla f(\boldsymbol{x})\|^2$, becomes

$$M := 2\lambda + \frac{L^2(6+d)^3}{4} \max_t \sigma_t^2,$$

using Lemma 2to bound $\|\nabla f_\sigma(\boldsymbol{x})\|^2$. Since we assume that $\sigma_t \to 0$, the sequence (σ_t) is bounded and hence $M \in \mathbb{R}$. We use this definition of M throughout this section. Lemmas 17 and 18 provide technical bounds needed in the proof of Lemma 19 that do not need to be adapted because σ is fixed in both of them.

Lemma C.1 (Lemma 19 [2])**.** *Let $(\boldsymbol{x}_t)_{t\geq 1}$, $(\boldsymbol{m}_t)_{t\geq 1}$, and $(\boldsymbol{v}_t)_{t\geq 1}$ be the sequences generated by GSmoothAdam. Let f be L-smooth and f^* denote the minimum of f. Assume $E(f_k(\boldsymbol{x})) = f(\boldsymbol{x})$ and $E(\nabla f_k(\boldsymbol{x})) = \nabla f(\boldsymbol{x})$. Suppose that $E(\|\nabla f_k\|^2) \leq \lambda$ for any $k \in [K]$. Then for all $t \geq 1$, we have*

$$E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}} \right\rangle\right) \geq \sum_{i=1}^{t+1} \prod_{j=i+1}^{t+1} \beta_j D_i + \sum_{i=1}^{t+1} \prod_{j=i+1}^{t+1} \beta_j \widetilde{D}_i \sqrt{|\sigma_i^2 - \sigma_{i+1}^2|},$$

where

$$D_i = -L\beta\eta_i E\left(\left\| \frac{\boldsymbol{m}_i}{\sqrt{\boldsymbol{v}_i+\epsilon}} \right\|^2\right) + \frac{1-\beta_{i+1}}{\sqrt{M^2+\epsilon}} E(\|\nabla f_{\sigma_{i+1}}(\boldsymbol{x}_i)\|^2) - \frac{\sqrt{d}M^4}{\epsilon^{3/2}}(1-\theta_{i+1})$$

and

$$\widetilde{D}_i = -\frac{\beta_i LM}{\sqrt{\epsilon}} \left(\frac{3+d}{2}\right)^{3/2}.$$

Proof. We repeat the proof from [2], but include the necessary changes for smoothing. Let

$$\begin{aligned} \Theta_{t+1} &= E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}} \right\rangle\right) \\ &= \underbrace{E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_t+\epsilon}} \right\rangle\right)}_{I} + \underbrace{E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}} - \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_t+\epsilon}} \right\rangle\right)}_{II}. \end{aligned} \tag{1}$$

Focusing on I, we have

$$
\begin{aligned}
&E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \middle| \mathcal{F}_t\right) \\
&= E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\beta_{t+1}\boldsymbol{m}_t + (1-\beta_{t+1})\nabla f_{k_t,\sigma_{t+1}}(\boldsymbol{x}_t)}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \middle| \mathcal{F}_t\right) \\
&= \beta_{t+1}\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle + (1-\beta_{t+1})\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \\
&= \beta_{t+1}\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle + (1-\beta_{t+1})\left\| \frac{(\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t))^2}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|_1 \\
&= \beta_{t+1}\left\langle \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle + (1-\beta_{t+1})\left\| \frac{(\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t))^2}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|_1 \\
&\qquad - \beta_{t+1}\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle,
\end{aligned} \tag{2}
$$

where $(\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t))^2$ is done coordinate-wise. Focusing on the last term of the previous equation, we have

$$
-\beta_{t+1}\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \geq -\beta\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1})\|\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|.
$$

Note that

$$
\begin{aligned}
\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1})\| &\leq \|\nabla f_{\sigma_t}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1})\| + \|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_t)\| \\
&\leq L\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\| + L\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2 - \sigma_{t+1}^2|}
\end{aligned}
$$

using Lemma 2and the fact that f_{σ_t} is L-smooth. So,

$$
\begin{aligned}
&-\beta_{t+1}\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \\
&\geq -\beta L\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\| - \beta L\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2 - \sigma_{t+1}^2|}\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\| \\
&= -\beta L\eta_t\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2 - \beta L\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2 - \sigma_{t+1}^2|}\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|.
\end{aligned}
$$

Combining this with where we were in (2), we have

$$
\begin{aligned}
&E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \middle| \mathcal{F}_t\right) \\
&\geq \beta_{t+1}\left\langle \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle + (1-\beta_{t+1})\left\| \frac{(\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t))^2}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|_1 \\
&\qquad - \beta L\eta_t\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2 - \beta_{t+1}L\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2 - \sigma_{t+1}^2|}\left\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|.
\end{aligned}
$$

So,

$$
\begin{aligned}
I &= E\left(E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle \middle| \mathcal{F}_t\right)\right) \\
&\geq -\beta_{t+1}E\left(\left\langle \nabla f_{\sigma_t}(\boldsymbol{x}_{t-1}), \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle\right) + (1-\beta_{t+1})E\left(\left\| \frac{(\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t))^2}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|_1\right) \\
&\quad - \beta L\eta_t E\left(\left\| \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2\right) \\
&\quad - \beta_{t+1}L\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2-\sigma_{t+1}^2|}E\left(\left\| \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|\right) \\
&\geq \beta_{t+1}\Theta_t + \frac{1-\beta_{t+1}}{\sqrt{M^2+\epsilon}}E\left(\left\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\right\|^2\right) - \beta L\eta_t E\left(\left\| \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2\right) \\
&\quad - \frac{\beta_{t+1}LM}{\sqrt{\epsilon}}\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2-\sigma_{t+1}^2|}
\end{aligned}
$$

where the last inequality used Lemmas 16 and 17 from [2].

Now focusing on II in (1),

$$
\begin{aligned}
II &= E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}} - \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_t+\epsilon}}\right\rangle\right) \\
&= -E\left(\left\langle \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_t+\epsilon}} - \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle\right) \\
&\geq E\left(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|\left\| \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_t+\epsilon}} - \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|\right) \\
&\geq -\frac{\sqrt{d}M^4}{\epsilon^{\frac{3}{2}}}(1-\theta_t)
\end{aligned}
$$

by Lemma 18 from [2].

Therefore, using the definitions of D_t and $\widetilde{D}_t$ from the statement of the lemma,

$$
\begin{aligned}
\Theta_{t+1} &\geq \beta_{t+1}\Theta_t + \frac{1-\beta_{t+1}}{\sqrt{M^2+\epsilon}}E\left(\left\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\right\|^2\right) - \beta L\eta_t E\left(\left\| \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2\right) \\
&\quad - \frac{\sqrt{d}M^4}{\epsilon^{\frac{3}{2}}}(1-\theta_{t+1}) - \frac{\beta_{t+1}LM}{\sqrt{\epsilon}}\left(\frac{3+d}{2}\right)^{3/2}\sqrt{|\sigma_t^2-\sigma_{t+1}^2|} \\
&= \beta_{t+1}\Theta_t + D_t + \widetilde{D}_t\sqrt{|\sigma_t^2-\sigma_{t+1}^2|}
\end{aligned}
$$

Recursively applying this inequality, we see that

$$\begin{aligned}\Theta_{t+1} &\geq \prod_{i=1}^{t+1}\beta_i\Theta_0 + \sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta_j D_i + \sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta_j\widetilde{D}_i\sqrt{|\sigma_i^2-\sigma_{i+1}^2|}\\ &= \sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta_j D_i + \sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta_j\widetilde{D}_i\sqrt{|\sigma_i^2-\sigma_{i+1}^2|}\end{aligned}$$

since $\Theta_0 = 0$ and we used $\prod_{j=t+2}^{t+1}\beta_t = 1$ for notational convenience. □

The next result, which is the GSmoothAdam analogue of Theorem 1for GmoothSGD, shows that the minimum gradient of the iterates converges to 0.

Theorem C.1 (Theorem 4 from [2])**.** *Let f be L-smooth and f^* denote the minimum of f. Assume $E(f_k(\boldsymbol{x})) = f(\boldsymbol{x})$ and $E(\nabla f_k(\boldsymbol{x})) = \nabla f(\boldsymbol{x})$. Suppose that $E(\|\nabla f_k\|^2) \leq \lambda$ for any $k \in [K]$ and $(\alpha_t)_{t\geq 1}$ is a non-increasing real sequence. Let $(\boldsymbol{x}_t)_{t\geq 1}$ be generated by GSmoothAdam with $\eta_t = \Theta(\alpha_t)$ (i.e., there exist $C_0, \widetilde{C_0} > 0$ such that $C_0\alpha_t \leq \eta_t \leq \widetilde{C_0}\alpha_t$). Then for $T \geq 1$, we have*

$$\begin{aligned}\min_{1\leq t\leq T}\left(E(\|\nabla f(\boldsymbol{x}_t)\|^2)\right)\sum_{t=1}^{T}\eta_t &\leq 2C_1(f_{\sigma_1}(\boldsymbol{x}_1)-f^*) + 2C_2\sum_{t=1}^{T}\eta_t(1-\theta_t) + 2C_3\sum_{t=1}^{T}\eta_t^2\\ &\quad + \frac{Ld\sqrt{M^2+\epsilon}}{2(1-\beta)}\sum_{t=1}^{T}|\sigma_t^2-\sigma_{t+1}^2| + \frac{L^2(6+d)^3}{4}\sum_{t=1}^{T}\eta_t\sigma_{t+1}^2,\end{aligned}$$

where

$$C_1 = \frac{\sqrt{M^2+\epsilon}}{1-\beta},\ C_2 = \frac{\widetilde{C_0}\sqrt{d}M^4\sqrt{M^2+\epsilon}}{\epsilon^{3/2}C_0(1-\beta)^2},\ C_3 = \frac{2\widetilde{C_0}^2LM^2\sqrt{M^2+\epsilon}}{\epsilon C_0^2(1-\beta)^2}.$$

Proof. Again, we will repeat the proof from [2], but with the necessary changes for smoothing. Repeating what was done at the beginning of the GSmoothSGD proof, we have

$$\begin{aligned}f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1}) &\leq f_{\sigma_{t+1}}(\boldsymbol{x}_t) + \langle\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{L}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2\\ &= f_{\sigma_{t+1}}(\boldsymbol{x}_t) - \eta_{t+1}\left\langle\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle + \frac{L\eta_{t+1}^2}{2}\left\|\frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2.\end{aligned}$$

This means

$$\begin{aligned}E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) &\leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - \eta_{t+1}E\left(\left\langle\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t), \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\rangle\right)\\ &\quad + \frac{L\eta_{t+1}^2}{2}E\left(\left\|\frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2\right)\\ &\leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - \eta_{t+1}\sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta D_i - \eta_{t+1}\sum_{i=1}^{t+1}\prod_{j=i+1}^{t+1}\beta\widetilde{D}_i\sqrt{|\sigma_i^2-\sigma_{i+1}^2|}\\ &\quad + \frac{L\eta_{t+1}^2}{2}E\left(\left\|\frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1}+\epsilon}}\right\|^2\right).\end{aligned}$$

Now, repeating the analysis that was done in [2], the smoothed version of their equation (23) is

$$E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) \leq E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - \frac{(1-\beta)\eta_{t+1}}{\sqrt{M^2+\epsilon}} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{\beta L \eta_{t+1} M^2}{\epsilon} \sum_{i=1}^{t+1} \beta^{t-i} \eta_i$$
$$+ \frac{L\eta_{t+1}^2 M^2}{2\epsilon} + \frac{\sqrt{d} M^4 \eta_{t+1}}{\epsilon^{3/2}} \sum_{i=1}^{t+1} \beta^{t-i}(1-\theta_i) - \eta_{t+1} \sum_{i=1}^{t+1} \prod_{j=i+1}^{t+1} \beta_j \widetilde{D}_i \sqrt{|\sigma_i^2 - \sigma_{i+1}^2|}.$$

As such, rearranging and summing (as in [2]),

$$\frac{(1-\beta)}{\sqrt{M^2+\epsilon}} \sum_{t=1}^{T} \eta_{t+1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) \leq \sum_{t=1}^{T} \Big(E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) \Big)$$
$$+ \frac{L\eta_{t+1}^2 M^2}{2\epsilon} + \frac{\beta L \eta_{t+1} M^2}{\epsilon} \sum_{i=1}^{t+1} \beta^{t-i} \eta_i$$
$$+ \frac{\sqrt{d} M^4 \eta_{t+1}}{\epsilon^{3/2}} \sum_{i=1}^{t} \beta^{t-i}(1-\theta_i) - \eta_{t+1} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \beta_j \widetilde{D}_i \sqrt{|\sigma_i^2 - \sigma_{i+1}^2|}.$$

Now,

$$E(f_{\sigma_{t+1}}(\boldsymbol{x}_t)) - E(f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) \leq E(f_{\sigma_t}(\boldsymbol{x}_t) - f_{\sigma_{t+1}}(\boldsymbol{x}_{t+1})) + \frac{Ld}{4} |\sigma_t^2 - \sigma_{t+1}^2|,$$

which means

$$\frac{(1-\beta)}{\sqrt{M^2+\epsilon}} \sum_{t=1}^{T} \eta_{t+1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) \leq f_{\sigma_1}(\boldsymbol{x}_1) - f^* + \frac{L\eta_k^2 M^2}{2\epsilon}$$
$$+ \frac{\beta L \eta_{t+1} M^2}{\epsilon} \sum_{i=1}^{t} \beta^{t-i} \eta_{i-1} + \frac{\sqrt{d} M^4 \eta_{t+1}}{\epsilon^{3/2}} \sum_{i=1}^{t+1} \beta^{t-i}(1-\theta_i)$$
$$- \eta_{t+1} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \beta_j \widetilde{D}_i \sqrt{|\sigma_i^2 - \sigma_{i+1}^2|} + \frac{Ld}{4} \sum_{t=1}^{T} |\sigma_t^2 - \sigma_{t+1}^2|.$$

Again repeating the analysis in [2] (and carrying along the additional additive σ terms), our version of their equation (30) is

$$\sum_{t=1}^{T} \eta_{t+1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) \leq C_1 (f_{\sigma_1}(\boldsymbol{x}_1) - f^*) + C_2 \sum_{t=1}^{T} \eta_{t+1}(1-\theta_{t+1}) + C_3 \sum_{t=1}^{T} \eta_{t+1}^2$$
$$- \frac{\sqrt{M^2+\epsilon}}{1-\beta} \sum_{t=1}^{T} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \eta_{t+1} \beta_j \widetilde{D}_i \sqrt{|\sigma_i^2 - \sigma_{i+1}^2|} + \frac{Ld\sqrt{M^2+\epsilon}}{4(1-\beta)} \sum_{t=1}^{T} |\sigma_t^2 - \sigma_{t+1}^2|.$$

Although it will improve the estimate, for simplicity since all of the terms of the fourth term of the

previous equation are positive, we have

$$\sum_{t=1}^{T} \eta_{t+1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) \leq C_1(f_{\sigma_1}(\boldsymbol{x}_1) - f^*) + C_2 \sum_{t=1}^{T} \eta_{t+1}(1-\theta_{t+1}) + C_3 \sum_{t=1}^{T} \eta_{t+1}^2 + \frac{Ld\sqrt{M^2+\epsilon}}{4(1-\beta)} \sum_{t=1}^{T} |\sigma_t^2 - \sigma_{t+1}^2|. \quad (3)$$

Now, applying Lemma 4 from [5] and finishing the proof from [2], we have

$$\begin{aligned}
&\min_{1\leq t\leq T} \left(E(\|\nabla f(\boldsymbol{x}_t)\|^2)\right) \sum_{t=1}^{T} \eta_{t+1} \quad (4)\\
&\leq \min_{1\leq t\leq T} \left(2E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L^2\sigma_{t+1}^2(6+d)^3}{4}\right) \sum_{t=1}^{T} \eta_{t+1}\\
&= \sum_{t=1}^{T} \eta_{t+1} \min_{1\leq t\leq T} \left(2E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L^2\sigma_{t+1}^2(6+d)^3}{4}\right)\\
&\leq \sum_{t=1}^{T} \eta_{t+1} \left(2E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L^2\sigma_{t+1}^2(6+d)^3}{4}\right)\\
&= 2\sum_{t=1}^{T} \eta_{t+1} E(\|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2) + \frac{L^2(6+d)^3}{4} \sum_{t=1}^{T} \eta_{t+1}\sigma_{t+1}^2.
\end{aligned}$$

Combining (3) and (4) gives the result. □

The next result provides almost sure convergence of GSmoothAdam. The proof primarily relies on the analysis in the proof of the previous theorem.

Theorem C.2 (Theorem 9 from [2])**.** *Let $f = \frac{1}{K}\sum_{k=1}^{K} f_k$ be L-smooth and f^* denote the minimum of f. Suppose that $E(\|\nabla f_k\|^2) \leq \lambda$ for any $k \in [K]$. Let $(\boldsymbol{x}_t)_{t\geq 1}$ be generated by GSmoothAdam. Suppose*

$$\sum_{t=1}^{\infty} \frac{\eta_t}{\sum_{i=1}^{t-1}\eta_i} = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty, \quad \sum_{t=1}^{\infty} \eta_t(1-\theta_t) < \infty,$$

and η_t is decreasing. If $\sum_{t=1}^{\infty} |\sigma_t^2 - \sigma_{t+1}^2| < \infty$, then for any $T \geq 1$, we have

$$\min_{1\leq t\leq T} \|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2 = o\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right) \text{ a.s.}$$

Proof. The proof is exactly the same, just replace $f(\boldsymbol{x}^k)$ with $f_{\sigma_{t+1}}(\boldsymbol{x}_t)$. Also, in order to use Lebesgue's Monotone Convergence Theorem, we need to use our assumption about the summability of $|\sigma_t^2 - \sigma_{t+1}^2|$. □

With most of the heavy lifting completed in the proofs of the previous results, we are ready to show that the iterates from GSmoothAdam converge to a stationary point a.s. in $\boldsymbol{x}$.

Proof of Theorem 2. Once again, this proof follows the same structure as in [2] with changes for smoothing. Since f satisfies the assumptions of Theorem C.2, from the proof of Theorem C.2 we have that

$$\sum_{t=1}^{\infty} \eta_{t+1} \|\nabla f_{\sigma_{t+2}}(\boldsymbol{x}_{t+1})\|^2 < \infty \text{ a.s.}$$

Since f is L-smooth,

$$\begin{aligned}
\left| \|\nabla f_{\sigma_{t+2}}(\boldsymbol{x}_{t+1})\| - \|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\| \right| &\leq \|\nabla f_{\sigma_{t+2}}(\boldsymbol{x}_{t+1}) - \nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\| \\
&\leq L\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| + L\left(\frac{3+d}{2}\right)^{3/2} \sqrt{|\sigma_{t+2}^2 - \sigma_{t+1}^2|} \\
&= L\eta_{t+1} \left\| \frac{\boldsymbol{m}_{t+1}}{\sqrt{\boldsymbol{v}_{t+1} + \epsilon}} \right\| + L\left(\frac{3+d}{2}\right)^{3/2} \sqrt{|\sigma_{t+2}^2 - \sigma_{t+1}^2|} \\
&\leq \frac{LM}{\sqrt{\epsilon}} \eta_{t+1} + L\left(\frac{3+d}{2}\right)^{3/2} \sqrt{|\sigma_{t+2}^2 - \sigma_{t+1}^2|} \text{ a.s.} \\
&\leq \frac{LM}{\sqrt{\epsilon}} \eta_{t+1} + L\left(\frac{3+d}{2}\right)^{3/2} \eta_{t+1} \text{ a.s.} \\
&= \eta_{t+1} \left(\frac{LM}{\sqrt{\epsilon}} + L\left(\frac{3+d}{2}\right)^{3/2} \right).
\end{aligned}$$

By Lemma 21 of [2],

$$\lim_{t\to\infty} \|\nabla f_{\sigma_{t+1}}(\boldsymbol{x}_t)\|^2 = 0 \text{ a.s..}$$

By Lemma 2, since $\sigma_t \to 0$, we know that

$$\lim_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 = 0 \text{ a.s..}$$

Since $\|\nabla f(\boldsymbol{x}_t)\| \leq M$ a.s., by Lebesgue's Dominated Convergence Theorem, we have that

$$\lim_{t\to\infty} E(\|\nabla f(\boldsymbol{x}_t)\|^2) = E\left(\lim_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2\right) = 0.$$

□

D Details Regarding Explicitly Smoothing Neural Networks

In order to calculate the explicit form that a smoothed neural network has, we need to be able to combine all of the model's components (including its constraint equations) into a single function that we can take the expectation of. We begin our analysis by writing out the constrained and unconstrained optimization problems for both FFNNs and CNNs.

A FFNN satisfies the following constrained optimization problem:

$$\begin{aligned} \min_{\theta,b} \quad & \sum_{n=1}^{N} \|x_L^n - y^n\|^2 \\ \text{subject to} \quad & x_1^n = \theta_1 x_0^n + b_1 \\ & x_l^n = \theta_l h(x_{l-1}^n) + b_l \text{ for } l = 2, ..., L \end{aligned}$$

Converting this to an unconstrained optimization problem means that a FFNN satisfies the following:

$$\min_{\theta,b} \sum_{n=1}^{N} \|x_L^n - y^n\|^2 + \lambda_1 \|\theta_1 x_0^n + b_1 - x_1^n\|^2 + \sum_{l=2}^{L} \lambda_l \|\theta_l h(x_{l-1}^n) + b_l - x_l^n\|^2. \tag{5}$$

A CNN satisfies the following constrained optimization problem:

$$\begin{aligned} \min_{\theta,b} \quad & \sum_{n=1}^{N} \|x_{L+C}^n - y^n\|^2 \\ \text{subject to} \quad & x_1^n = x_0^n * \theta_1 + b_1 \\ & x_l^n = h(x_{l-1}^n) * \theta_l + b_l \text{ for } 2 \le l \le C \\ & x_l^n = \theta_l h(x_{l-1}^n) + b_l \text{ for } C+1 \le l \le C+L \end{aligned}$$

Again converting this into an unconstrained problem shows that a CNN satisfies:

$$\begin{aligned} \min_{\theta,b} \sum_{n=1}^{N} \|x_{L+C}^n - y^n\|^2 + \lambda_1 \|x_0^n * \theta_1 + b_1 - x_1^n\|_F^2 + \sum_{l=2}^{C} \lambda_l \|h(x_{l-1}^n) * \theta_l + b_l - x_l^n\|_F^2 \\ + \sum_{l=C+1}^{C+L} \lambda_l \|\theta_l h(x_{l-1}^n) + b_l l - x_l^n\|^2. \end{aligned} \tag{6}$$

To smooth these loss functions, we need to smooth each term of its sum. The summary of the smoothed terms and the propositions where they are proven can be found in Table 1.

In order to smooth the functions from the unconstrained problems, we need the following results primarily from [4].

Proposition D.1. *The following are the results from Gaussian smoothing:*

(a) [4] $\text{relu}_\sigma(x) = \frac{x}{2}\left(1 + \text{erf}(x/\sigma)\right) + \frac{\sigma}{2\sqrt{\pi}} e^{-x^2/\sigma^2}$

(b) $\text{relu}_\sigma^2(x) = \frac{1}{4}(1 + \text{erf}(x/\sigma))(\sigma^2 + 2x^2) + \frac{\sigma x}{2\sqrt{\pi}} e^{-x^2/\sigma^2}$

Table 1: Summary of layer-wise Gaussian smoothing (constants removed)

Layer No. (l)	**Original Constraint**	**Smoothed**		**Prop.**
		Constraint	**Regularizer**	
$l=1$	$x_l = \theta_l h(x_{l-1}) + b_l$	$x_l = \theta_l h_\sigma(x_{l-1}) + b_l$		D.2
$l>1$	$x_l = \theta_l h(x_{l-1}) + b_l$	$x_l = \theta_l h_\sigma(x_{l-1}) + b_l$	$\|\theta_l \mathrm{diag}(\sqrt{(h^2)_\sigma}(x_{l-1}))\|_F^2$ $+\|\theta_l \mathrm{diag}(h_\sigma(x_{l-1}))\|_F^2$	D.2
$l>1$	$x_l = \theta_l x_{l-1} + b_l$	$x_l = \theta_l \sigma(x_{l-1}) + b_l$	$\frac{\sigma^2}{2}\|\theta_l\|_F^2 + \frac{\sigma^2 d_l}{2}\|x_{l-1}\|^2$	D.2
$l=1$	$x_l = x_{l-1} * \theta_l + b_l - x_l$	$x_l = x_{l-1} * \theta_l + b_l - x_l$		D.3
$l>1$	$x_l = x_{l-1} * \theta_l + b_l - x_l$	$x_l = x_{l-1} * \theta_l + b_l - x_l$	$\frac{\sigma^2 w_l^2}{2}\|\theta_l\|_F^2 + \frac{\sigma^2}{2}\|x_{l-1}\|_{C_l}^2$	D.3
$l \geq 1$	$x_l = \mathrm{Drop}(x_{l-1})$	$x_l = x_{l-1}$	$p\|x_l\|^2$	D.4
$l \geq 1$	$x_l = \mathrm{AvgPool}(x_{l-1})$	$x_l = \mathrm{AvgPool}(x_{l-1})$		D.5

(c) [4] For a matrix A, vector b, and function h,

$$\Big(\|Ah(\boldsymbol{y})+b\|^2\Big)_\sigma(\boldsymbol{x}) = \Big\|Ah_\sigma(\boldsymbol{x})+b\Big\|^2 + \Big\|A\mathrm{diag}\Big(\sqrt{(h^2)_\sigma}(\boldsymbol{x})\Big)\Big\|_F^2 - \Big\|A\mathrm{diag}(h_\sigma(\boldsymbol{x}))\Big\|_F^2$$

(d) $(\|\boldsymbol{y}\|^2)_\sigma(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 + \frac{\sigma^2 d}{2}$

Proof of (b). We have the following:

$$\begin{aligned}
(\mathrm{relu}^2)_\sigma(x) &= \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} \Big(\mathrm{relu}(x+\sigma u)\Big)^2 e^{-u^2}\,du \\
&= \frac{1}{\sqrt{\pi}}\int_{-x/\sigma}^{\infty} (x+\sigma u)^2 e^{-u^2}\,du \\
&= \frac{1}{\sqrt{\pi}}\int_{-x/\sigma}^{\infty} (x^2+2\sigma x u+\sigma^2 u^2) e^{-u^2}\,du \\
&= \frac{x^2}{2}\Big(1+\mathrm{erf}(x/\sigma)\Big) - \frac{\sigma x}{\sqrt{\pi}} e^{-x^2/\sigma^2} + \frac{\sigma x}{2\sqrt{\pi}} e^{-x^2/\sigma^2} + \frac{\sigma^2}{4}\Big(1+\mathrm{erf}(x/\sigma)\Big) \\
&= \frac{1}{4}\Big(1+\mathrm{erf}(x/\sigma)\Big)(\sigma^2+2x^2) + \frac{\sigma x}{2\sqrt{\pi}} e^{-x^2/\sigma^2}
\end{aligned}$$

□

Proof of (d). Observe:

$$\begin{aligned}
\frac{1}{\sqrt{\pi}}\int_{\mathbb{R}^d} \|\boldsymbol{x}+\sigma\boldsymbol{u}\|^2 e^{-\|\boldsymbol{u}\|^2}\,du &= \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}^d} (\|\boldsymbol{x}\|^2 + 2\sigma\langle \boldsymbol{x},\boldsymbol{u}\rangle + \sigma^2\|\boldsymbol{u}\|^2) e^{-\|\boldsymbol{u}\|^2}\,du \\
&= \|\boldsymbol{x}\|^2 + 0 + \frac{\sigma^2 d}{2}.
\end{aligned}$$

□

D.1 Smoothing Terms in Neural Network Unconstrained Optimization

In this section, we smooth each of the terms from the FFNN and CNN unconstrained problems. Let us represent the unconstrained problem as

$$\min_{x,\theta,b} \sum_{n=1}^{N} f(x,\theta,b)$$

where f is a sum of certain norms. Since smoothing is a linear operator, we can focus on a single data point and drop the sum. Hence, we focus on smoothing $f(x,\theta,b)$. Then

$$\begin{aligned}(f * k_\sigma)(x,\theta,b) &= \underbrace{\int_{\mathbb{R}^{d_L}} \cdots \int_{\mathbb{R}^{d_1}}}_{\substack{\text{smoothing}\\ \text{wrt } x}} \underbrace{\int_{\mathbb{R}^{d_L \times d_{L-1}}} \cdots \int_{\mathbb{R}^{d_1 \times d_0}}}_{\substack{\text{smoothing}\\ \text{wrt } \theta}} \underbrace{\int_{\mathbb{R}^{d_L}} \cdots \int_{\mathbb{R}^{d_1}}}_{\substack{\text{smoothing}\\ \text{wrt } b}} f(x+\sigma u_x, \theta+\sigma U_\theta, b+\sigma u_b) \\ &\qquad \cdot e^{-\|u_b\|^2} e^{-\|U_\theta\|_F^2} e^{-\|u_x\|^2} du_b dU_\theta du_x.\end{aligned}$$

Since $f \geq 0$, we are free to switch the order of integration. As such, the process that we will take is to sequentially smooth with respect to the variables.

A summary of the results in this section can be found in Table 1.

Based on the unconstrained problems, we only ever need to focus on a layer and its input. We use the notation x for the input to a layer and x_+ for the output. As such, the weights and biases of a layer are denoted by θ_+ and b_+, respectively.

Proposition D.2 (Smoothing Dense Layer)**.** *Let $x \in \mathbb{R}^d$, $x_+, b_+ \in \mathbb{R}^{d_+}$, $\theta_+ \in \mathbb{R}^{d \times d_+}$, and*

$$l(b_+, x_+, \theta_+, x) = \|\theta_+ h(x) + b_+ - x_+\|^2.$$

If $x = x_0$ represents the input data (i.e. we cannot smooth with respect to x), then

$$l_\sigma(b_1, x_1, \theta_1, x_0) = \|\theta_1 h(x_0) + b_1 - x_1\|^2 + \frac{\sigma^2 d_1}{2}\|h(x_0)\|^2 + \sigma^2 d_1.$$

Otherwise

$$\begin{aligned}l_\sigma(b_+, x_+, \theta_+, x) = \|\theta_+ h_\sigma(x) + b_+ - x_+\|^2 + \|\theta_+ \mathrm{diag}(\sqrt{(h^2)_\sigma}(x))\|_F^2 \\ - \|\theta \mathrm{diag}(h_\sigma(x))\|_F^2 + \sigma^2 d_+ \left(1 + \frac{1}{2}\|\sqrt{(h^2)_\sigma}(x)\|^2\right).\end{aligned}$$

Furthermore, if $h(x) = x$, then

$$l_\sigma(b_+, x_+, \theta_+, x) = \|\theta_+ x + b_+ - x_+\|^2 + \frac{\sigma^2}{2}\|\theta_+\|_F^2 + \frac{\sigma^2 d_+}{2}\|x\|^2 + \frac{\sigma^4 d d_+}{4} + \sigma^2 d_+.$$

Proof. First, we smooth with respect to b_+:

$$\begin{aligned}&\frac{1}{\pi^{d_+/2}} \int_{\mathbb{R}^{d_+}} \|\theta_+ h(x) + (b_+ + \sigma u) - x_+\|^2 e^{-\|u\|^2}\, du \\ &= \frac{1}{\pi^{d_+/2}} \int_{\mathbb{R}^{d_+}} \left(\|\theta_+ h(x) + b_+ - x_+\|^2 + 2\sigma\left\langle \theta_+ h(x) + b_+ - x_+, u\right\rangle + \sigma^2\|u\|^2\right) e^{-\|u\|^2}\, du \\ &= \|\theta_+ h(x) + b_+ - x_+\|^2 + \frac{\sigma^2 d_+}{2}\end{aligned}$$

Second, we smooth with respect to x_+. Since x_+ plays the same role as b_+, smoothing with respect to x_+ is analogous to smoothing with respect to b_+, so we end up with

$$\|\theta_+ h(x) + b_+ - x_+\|^2 + \sigma^2 d_+.$$

Third, we smooth with respect to θ_+. Focusing on the first term in the previous equation, let $d' = \dim(\theta_+) = d_+ \times d$, then

$$\begin{aligned}
&\frac{1}{\pi^{d'/2}} \int_{\mathbb{R}^{d'}} \|(\theta_+ + \sigma U)h(x) + b_+ - x_+\|^2 e^{-\|U\|_F^2}\, dU \\
&= \frac{1}{\pi^{d'}} \int_{\mathbb{R}^{d'}} \Big(\|\theta_+ h(x) + b_+ - x_+\|^2 + \sigma^2 \|Uh(x)\|^2\Big) e^{-\|U\|_F^2}\, dU \\
&= \|\theta_+ h(x) + b_+ - x_+\|^2 + \frac{\sigma^2}{\pi^{d'/2}} \sum_{i=1}^{d_+} \sum_{j=1}^{d} \int_{\mathbb{R}^{d'}} U^{ij} U^{ik} \big(h(x)\big)_j \big(h(x)\big)_k e^{-\|U\|_F^2}\, dU \\
&= \|\theta_+ h(x) + b_+ - x_+\|^2 + \frac{\sigma^2}{\pi^{d'/2}} \sum_{j=1}^{d} \left[\big(h(x)\big)_j^2 \sum_{i=1}^{d_+} \frac{\pi^{d'/2}}{2} \right] \\
&= \|\theta_+ h(x) + b_+ - x_+\|^2 + \frac{\sigma^2 d_+}{2} \|h(x)\|^2.
\end{aligned}$$

So, after smoothing with respect to b, θ, and x_+, we have

$$\|\theta_+ h(x) + b_+ - x_+\|^2 + \frac{\sigma^2 d_+}{2} \|h(x)\|^2 + \sigma^2 d_+. \tag{7}$$

Finally, we smooth with respect to x. By Mobahi's proposition, the first term of (7) smoothes to

$$\|\theta_+ h_\sigma(x) + b_+ - x_+\|^2 + \|\theta_+ \mathrm{diag}(\sqrt{(h^2)_\sigma}(x))\|_F^2 - \|\theta_+ \mathrm{diag}(h_\sigma(x))\|_F^2.$$

Again, by Mobahi's proposition, the second term of (7) smoothes to

$$\begin{aligned}
&\Big\|h_\sigma(x)\Big\|^2 + \Big\|\mathrm{diag}\Big(\sqrt{(h^2)_\sigma}(x)\Big)\Big\|_F^2 - \Big\|\mathrm{diag}(h_\sigma(x))\Big\|_F^2 \\
&\qquad = \|h_\sigma(x)\|^2 + \|\sqrt{(h^2)_\sigma}(x)\|^2 - \|h_\sigma(x)\|^2 = \|\sqrt{(h^2)_\sigma}(x)\|^2.
\end{aligned}$$

Combining these two, we have that after smoothing with respect to everything,

$$\begin{aligned}
&\|\theta_+ h_\sigma(x) + b_+ - x_+\|^2 + \|\theta_+ \mathrm{diag}(\sqrt{(h^2)_\sigma}(x))\|_F^2 - \|\theta_+ \mathrm{diag}(h_\sigma(x))\|_F^2 \\
&\qquad + \sigma^2 d_+ \left(1 + \frac{1}{2} \|\sqrt{(h^2)_\sigma}(x)\|^2\right).
\end{aligned}$$

To prove the furthermore statement, we jump back to (7), which means that if h is the identity function, then after smoothing with respect to b, θ, and x_+, we have

$$\|\theta_+ x + b_+ - x_+\|^2 + \frac{\sigma^2 d_+}{2} \|x\|^2 + \sigma^2 d_+. \tag{8}$$

Once again, we smooth with respect to x. The first term smoothes as follows:

$$\begin{aligned}
&\frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d}\|\theta_+(x+\sigma u)+b_+-x_+\|^2e^{-\|u\|^2}\,du\\
&=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{\pi^{d/2}}\int_{\mathbb{R}^d}\|\theta_+u\|^2e^{-\|u\|^2}\,du\\
&=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{\pi^{d/2}}\sum_{i=1}^{d_+}\sum_{j,k=1}^{d}(\theta_+)_{ij}(\theta_+)_{ik}\int_{\mathbb{R}^d}u_ju_ke^{-\|u\|^2}\,du\\
&=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{\pi^{d/2}}\sum_{i=1}^{d_+}\sum_{j=1}^{d}(\theta_+)_{ij}^2\int_{\mathbb{R}^d}u_j^2e^{-\|u\|^2}\,du\\
&=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{2}\sum_{i=1}^{d_+}\sum_{j=1}^{d}(\theta_+)_{ij}^2\\
&=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{2}\|\theta_+\|_F^2.
\end{aligned}$$

The second term of (8), smoothes to

$$\frac{\sigma^2d_+}{2}\left(\|x\|^2+\frac{\sigma^2d}{2}\right)=\frac{\sigma^2d_+}{2}\|x\|^2+\frac{\sigma^4dd_+}{4},$$

which has been done several times already in this proof. Therefore, we end up with

$$l_\sigma(b_+,x_+,\theta_+,x)=\|\theta_+x+b_+-x_+\|^2+\frac{\sigma^2}{2}\|\theta_+\|_F^2+\frac{\sigma^2d_+}{2}\|x\|^2+\frac{\sigma^4dd_+}{4}+\sigma^2d_+.$$

□

Before we smooth the convolutional layer, we want to write out the "convolutional norm":

$$\|x\|_{C_+}^2=\sum_{a,b=0}^{w_+-1}\sum_{i,j=1}^{k_+}\left(x^{(as_++i)(bs_++j)}\right)^2=\|x*J_{k_+}\|^2.$$

Proposition D.3 (Smoothing Convolutional Layer)**.** *Let $x\in\mathbb{R}^d$, $x_+,b_+\in\mathbb{R}^{d_+}$, $\theta_+\in\mathbb{R}^{c_+\times c_+}$, and*

$$l(\theta_+,x_+,x)=\|x*_+\theta_++b_+-x_+\|_F^2$$

were $*_+$ represents the convolution with stride s_+. Let C_+ represent the convolutional layer with these parameters. If $x=x_0$ represents the input data (i.e. we cannot smooth with respect to x), then

$$l_\sigma(\theta_+,x_+,x)=\|x*_+\theta_++b_+-x_+\|_F^2+\frac{\sigma^2}{2}\|x\|_{C_+}^2+\sigma^2d_+.$$

Otherwise,

$$l_\sigma(\theta_+,x_+,x)=\|x*_+\theta_++b_+-x_+\|_F^2+\frac{\sigma^2}{2}w_+^2\|\theta_+\|_F^2+\frac{\sigma^2}{2}\|x\|_{C_+}^2+\frac{\sigma^4}{4}\|1\|_{C_+}^2+\sigma^2d_+.$$

Proof. For the proof, we will simplify notation and write $*$ instead of $*_+$ since all of the convolutions will use the same stride. Recall that $w_+ = \left(\frac{\text{width}(x)+c_+}{s_+} + 1\right)$. First, we smooth with respect to b_+:

$$\begin{aligned}
&\frac{1}{\pi^{d_+/2}} \int_{\mathbb{R}^{d_+}} \|x * \theta_+ + (b_+ + \sigma U) - x_+\|_F^2 e^{-\|U\|_F^2}\, dU \\
&= \frac{1}{\pi^{d_+/2}} \int_{\mathbb{R}^{d_+}} \left(\|x * \theta_+ + b_+ - x_+\|_F^2 + 2\sigma\langle x * \theta_+ + b_+ - x_+, U\rangle_F + \sigma^2\|U\|_F^2\right) e^{-\|U\|_F^2}\, dU \\
&= \|x * \theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2 d_+}{2}.
\end{aligned}$$

Second, we smooth with respect to θ_+. For first term from the previous equation, we have

$$\begin{aligned}
&\frac{1}{\pi^{c_+^2/2}} \int_{\mathbb{R}^{c_+^2}} \|x * (\theta_+ + \sigma U) + b_+ - x_+\|_F^2 e^{-\|U\|_F^2}\, dU \\
&\qquad = \|x * \theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{\pi^{c_+^2/2}} \int_{\mathbb{R}^{c_+^2}} \|x * U\|_F^2 e^{-\|U\|_F^2}\, dU
\end{aligned}$$

Focusing on the convolution term, (here $a', b' = 0, s_+, 2s_+, ..., w_+ s_+$)

$$\begin{aligned}
&\frac{\sigma^2}{\pi^{c_+^2/2}} \int_{\mathbb{R}^{c_+^2}} \|x * U\|_F^2 e^{-\|U\|_F^2}\, dU \qquad (9) \\
&= \frac{\sigma^2}{\pi^{c_+^2/2}} \int_{\mathbb{R}^{c_+^2}} \sum_{a',b'} \left(\sum_{i,j=1}^{c_+} U^{ij} x^{(a'+i)(b'+j)}\right)^2 e^{-\|U\|_F^2}\, dU \\
&= \frac{\sigma^2}{\pi^{c_+^2/2}} \int_{\mathbb{R}^{c_+^2}} \sum_{a',b'} \left(\sum_{i,j=1}^{c_+} \sum_{k,l=1}^{c_+} U^{ij} U^{kl} x^{(a'+i)(b'+j)} x^{(a'+k)(b'+l)}\right) e^{-\|U\|_F^2}\, dU \\
&= \frac{\sigma^2}{\pi^{c_+^2/2}} \sum_{a',b'} \sum_{i,j=1}^{c_+} \left(x^{(a'+i)(b'+j)}\right)^2 \int_{\mathbb{R}^{c_+^2}} (U^{ij})^2 e^{-\|U\|_F^2}\, dU \\
&= \frac{\sigma^2}{2} \|x\|_{C_+}^2.
\end{aligned}$$

So after smoothing with respect to b_+ and θ_+, we have

$$\|x * \theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}\|x\|_{C_+}^2 + \frac{\sigma^2 d_+}{2}.$$

Third, we smooth with respect to x_+, but as with b_+ we end up with

$$\|x * \theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}\|x\|_{C_+}^2 + \sigma^2 d_+. \qquad (10)$$

Finally, we smooth with respect to x. For the first term of (10), we have

$$\begin{aligned}
&\frac{1}{\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} \|(x + \sigma U) * \theta_+ + b_+ - x_+\|_F^2 e^{-\|U\|_F^2}\, dU \\
&\qquad = \|x * \theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} \|U * \theta_+\|_F^2 e^{-\|U\|_F^2}\, dU
\end{aligned}$$

Now, for the second term, we repeat the process in (9) and arrive at

$$\begin{aligned}
&\frac{1}{\pi^{\frac{d}{2}}}\int_{\mathbb{R}^d}\|(x+\sigma U)*\theta_+ + b_+ - x_+\|_F^2 e^{-\|U\|_F^2}\,dU\\
&= \|x*\theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}\sum_{a,b=1}^{w_+}\sum_{i,j=1}^{c_+}(\theta_+^{ij})^2\\
&= \|x*\theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}w_+^2\sum_{i,j=1}^{c_+}(\theta_+^{ij})^2\\
&= \|x*\theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}w_+^2\|\theta_+\|_F^2.
\end{aligned}$$

For the second term of (10),

$$\begin{aligned}
&\frac{1}{\pi^{\frac{d}{2}}}\int_{\mathbb{R}^d}\|x+\sigma U\|_{C_+}^2 e^{-\|U\|_F^2}\,dU\\
&= \frac{1}{\pi^{\frac{d}{2}}}\int_{\mathbb{R}^d}\sum_{a,b=1}^{w_+}\sum_{i,j=1}^{c_+}\left(x^{(a+i)(b+j)}+\sigma U^{(a+i)(b+j)}\right)^2 e^{-\|U\|_F^2}\,dU\\
&= \sum_{a,b=1}^{w_+}\sum_{i,j}\left(x^{(a+i)(b+j)}\right)^2 + \frac{\sigma^2}{\pi^{\frac{d}{2}}}\sum_{a,b=1}^{w_+}\sum_{i,j}\frac{\pi^{\frac{d}{2}}}{2}\\
&= \|x\|_{C_+}^2 + \frac{\sigma^2}{2}\|1\|_{C_+}^2.
\end{aligned}$$

So, after smoothing with respect to all variables, we end up with

$$\|x*\theta_+ + b_+ - x_+\|_F^2 + \frac{\sigma^2}{2}w_+^2\|\theta_+\|_F^2 + \frac{\sigma^2}{2}\|x\|_{C_+}^2 + \frac{\sigma^4}{4}\|1\|_{C_+}^2 + \sigma^2 d_+.$$

□

Proposition D.4 (Smoothing Dropout Layer)**.** *Let $p \in [0,1]$ and* drop *be given by $\mathcal{P}(\text{drop}(x) = 0) = p$ and $\mathcal{P}(\text{drop}(x) = x) = 1-p$. For $x, x_+ \in \mathbb{R}^d$, define the unconstrained loss associated with the dropout layer as*

$$l(x, x_+) = \|\text{drop}(x) - x_+\|^2.$$

Then

$$l_\sigma(x, x_+) = p\|x_+\|^2 + (1-p)\|x - x_+\|^2 + \left(1 - \tfrac{p}{2}\right)\sigma^2 d.$$

Proof. Let $x, x_+ \in \mathbb{R}^d$. Note that if x and x_+ are matrices, the norms and inner products in this proof are the Frobenius-type. First, we will smooth with respect to x_+. So,

$$E_u[l(x, x_+ + \sigma u)] = \frac{p}{\pi^{d/2}}\int_{\mathbb{R}^d}\|x_+ + \sigma u\|^2 e^{-\|u\|^2}\,du + \frac{1-p}{\pi^{d/2}}\int_{\mathbb{R}^d}\|x - (x_+ + \sigma u)\|^2 e^{-\|u\|^2}\,du.$$

The first integral can be computed as

$$\frac{p}{\pi^{d/2}}\int_{\mathbb{R}^d}\|x_+ + \sigma u\|^2 e^{-\|u\|^2}\,du = p\|x_+\|^2 + \frac{p\sigma^2 d}{2}.$$

The second integral is

$$\frac{1-p}{\pi^{d/2}}\int_{\mathbb{R}^d}\|x-(x_+ + \sigma u)\|^2 e^{-\|u\|^2}\,du = (1-p)\|x-x_+\|^2 + \frac{(1-p)\sigma^2 d}{2}.$$

So,

$$\begin{aligned}E_u[l(x, x_+ + \sigma u)] &= p\|x_+\|^2 + \frac{p\sigma^2 d}{2} + (1-p)\|x-x_+\|^2 + \frac{(1-p)\sigma^2 d}{2}\\ &= p\|x_+\|^2 + (1-p)\|x-x_+\|^2 + \frac{\sigma^2 d}{2}.\end{aligned}$$

Second, we will smooth with respect to x. We have

$$\begin{aligned}&E_u(l(x+\sigma u, x_+))\\ &= \frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d}\left(p\|x_+\|^2 + (1-p)\|x+\sigma u - x_+\|^2 + \frac{\sigma^2 d}{2}\right)e^{-\|u\|^2}\,du\\ &= p\|x_+\|^2 + \frac{\sigma^2 d}{2} + \frac{1-p}{\pi^{d/2}}\int_{\mathbb{R}^d}\|x+\sigma u - x_+\|^2 e^{-\|u\|^2}\,du\\ &= p\|x_+\|^2 + \frac{\sigma^2 d}{2} + (1-p)\left(\|x-x_+\|^2 + \frac{\sigma^2 d}{2}\right)\\ &= (1-p)\|x-x_+\|^2 + p\|x_+\|^2 + (2-p)\frac{\sigma^2 d}{2}.\end{aligned}$$

This shows the claimed result. □

Proposition D.5 (Smoothing Average Pooling)**.** *Define*

$$\Big(\text{pool}(x)\Big)_i = \frac{1}{k}\sum_{j=1}^{k} x^{i_j},$$

that is the pooling layer averages over x^{i_j} for $j = 1, ..., k$. Let

$$l(x, x_+) = \|\text{pool}(x) - x_+\|^2.$$

Then

$$l_\sigma(x, x_+) = \|\text{pool}(x) - x_+\|^2 + \frac{\sigma^2 d_+}{2k} + \frac{\sigma^2 d_+}{2}.$$

Proof. First, we smooth with respect to x_+, which has been done numerous times throughout this note and we end up with

$$\|\text{pool}(x) - x_+\|^2 + \frac{\sigma^2 d_+}{2}.$$

Second, we smooth with respect to x:

$$\begin{aligned}&\frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d}\|\text{pool}(x+\sigma u) - x_+\|^2 e^{-\|u\|^2}\,du\\ &\qquad = \|\text{pool}(x+\sigma u) - x_+\|^2 + \frac{\sigma^2}{\pi^{d/2}}\int_{\mathbb{R}^d}\|\text{pool}(u)\|^2 e^{-\|u\|^2}\,du.\end{aligned}$$

Focusing on the last term,

$$\begin{aligned}\int_{\mathbb{R}^d} \|\text{pool}(u)\|^2 e^{-\|u\|^2}\, du &= \sum_{i=1}^{d_+} \int_{\mathbb{R}^d} \left(\frac{1}{k}\sum_{j=1}^{k} u^{i_j}\right)^2 e^{-\|u\|^2}\, du \\ &= \frac{1}{k^2}\sum_{i=1}^{d_+}\sum_{j,m=1}^{k}\int_{\mathbb{R}^d} u^{i_j}u^{i_m}e^{-\|u\|^2}\, du \\ &= \frac{1}{k^2}\sum_{i=1}^{d_+}\sum_{j=1}^{k}\int_{\mathbb{R}^d} (u^{i_j})^2 e^{-\|u\|^2}\, du \\ &= \frac{d_+\pi^{d/2}}{2k}.\end{aligned}$$

So,

$$\frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d} \|\text{pool}(x+\sigma u) - x_+\|^2 e^{-\|u\|^2}\, du = \|\text{pool}(x+\sigma u) - x_+\|^2 + \frac{d_+\sigma^2}{2k}.$$

Combining this with the constant term from the previous equation gives the result. □

Note D.1 (Comment about pooling layers)**.** *If a max pooling layer is used, then the smoothing function is difficult (if not impossible to compute). In particular, let $\mathcal{I}$ is the index that a single part of the pooling layer takes the maximum over and*

$$h(x) = \max_{i,j\in\mathcal{I}} x^{ij}.$$

Since h is nonlinear, we need to use Proposition D.1 (e) which requires computing h_σ. Then

$$\begin{aligned}h_\sigma(x) &= \frac{1}{\pi^{|\mathcal{I}|/2}}\int_{\mathbb{R}^{|\mathcal{I}|}} \max_{i,j\in\mathcal{I}} \left(x^{ij} + \sigma U_{ij}\right) e^{-\|U\|_F^2}\, dU \\ &= E\left(\max_{i,j\in\mathcal{I}} X_{ij}\right)\end{aligned}$$

where $X_{ij} \sim \mathcal{N}(x_l^{ij}, \sigma^2)$. Even in the case where all of the x_l^{ij} are the same, this does not have a closed form. Hence, we cannot compute h_σ.

D.2 Proofs of Mathematical Formulation of Smoothed Neural Networks

Now that we have smoothed all of the components of both of the unconstrained problems (5) and (6), we are ready to explicitly write out their smoothed mathematical formulations.

Proof of Theorem 3. The unconstrained FFNN is given in (5), which finds the minimum of the sum (over n) of

$$\|x_L^n - y^n\|^2 + \lambda_1\|\theta_1 x_0^n + b_1 - x_1^n\|^2 + \sum_{l=2}^{L}\lambda_l\|\theta_l h(x_{l-1}^n) + b_l - x_l^n\|^2.$$

Based on the previous section, this smoothes to

$$\begin{aligned}\|x_L^n - y^n\|^2 + \frac{\sigma^2 d_L}{2} + \lambda_1 \left(\|\theta_1 x_0^n + b_1 - x_1^n\|^2 + \frac{\sigma^2 d_1}{2} \|x_0\|^2 + \sigma^2 d_1 \right) \\ + \sum_{l=2}^{L} \lambda_l \Bigg(\|\theta_l h(x_{l-1}^n) + b_l - x_l^n\|^2 + \|\theta_l \mathrm{diag}(\sqrt{(h^2)_\sigma}(x_{l-1}))\|_F^2 \\ - \|\theta_l \mathrm{diag}(h_\sigma(x_{l-1}))\|_F^2 + \sigma^2 d_l \left(1 + \frac{1}{2} \|\sqrt{(h^2)_\sigma}(x_{l-1})\|^2 \right) \Bigg).\end{aligned}$$

Now, we can reconstrain to get the result. The proof for the CNN case is exactly the same. □

Table 2: Example converting CNN layers to smoothed counterparts (additive constants omitted).

Original Constraint	**Smoothed**	
	Constraint	**Regularizer**
$\|x_0 * \theta_1 + b_1 - x_1\|_F^2$	$\|x_0 * \theta_1 + b_1 - x_1\|_F^2$	
$\|h(x_1) - x_1^h\|_F^2$	$\|h_\sigma(x_1) - x_1^h\|_F^2$	$+\|\sqrt{(h^2)_\sigma}(x_1)\|_F^2 - \|h_\sigma(x_1)\|_F^2$
$\|\text{Pool}(x_1^h) - x_2\|_F^2$	$\|\text{Pool}(x_1^h) - x_2\|_F^2$	
$\|\text{Flatten}(x_2) - x_3\|^2$	$\|\text{Flatten}(x_2) - x_3\|^2$	
$\|\theta_4 x_3 + b_4 - x_4\|^2$	$\|\theta_4 x_3 + b_4 - x_4\|^2$	$+\frac{\sigma^2}{2}\|\theta_4\|_F^2 + \frac{\sigma^2 d_4}{2}\|x_3\|^2$
$\|h(x_4) - x_4^h\|^2$	$\|h_\sigma(x_4) - x_4^h\|^2$	$+\|\sqrt{(h^2)_\sigma}(x_4)\|^2 - \|h_\sigma(x_4)\|^2$
$\|\theta_5 x_4^h + b_5 - x_5\|^2$	$\|\theta_5 x_4^h + b_5 - x_5\|^2$	$+\frac{\sigma^2}{2}\|\theta_5\|_F^2 + \frac{\sigma^2 d_5}{2}\|x_4\|^2$

E Additional Details of Numerical Experiments and Practical Guide to Smooth Neural Network Implementation

We begin with a description of how to implement the explicitly smooth neural networks. In order to use TensorFlow's built-in layer regularizers, we avoid regularization terms that mix weights and layer inputs (e.g., $\|\theta \text{diag}(\sqrt{(h^2)_\sigma}(x))\|_F^2$). This is why took a non-standard approach and we separated the activation function from the layer. Hence, we split up terms like $\|\theta h(x) + b\|$ into two terms like $\|h(x) - y\|$ and $\|\theta y + b\|$. Then we use the results from Section D.1 to smooth. The original unconstrained terms and their smoothed constraints and regularization terms can be found in Table 2.

Practically, here are the steps that we follow to get the explicitly smooth network:

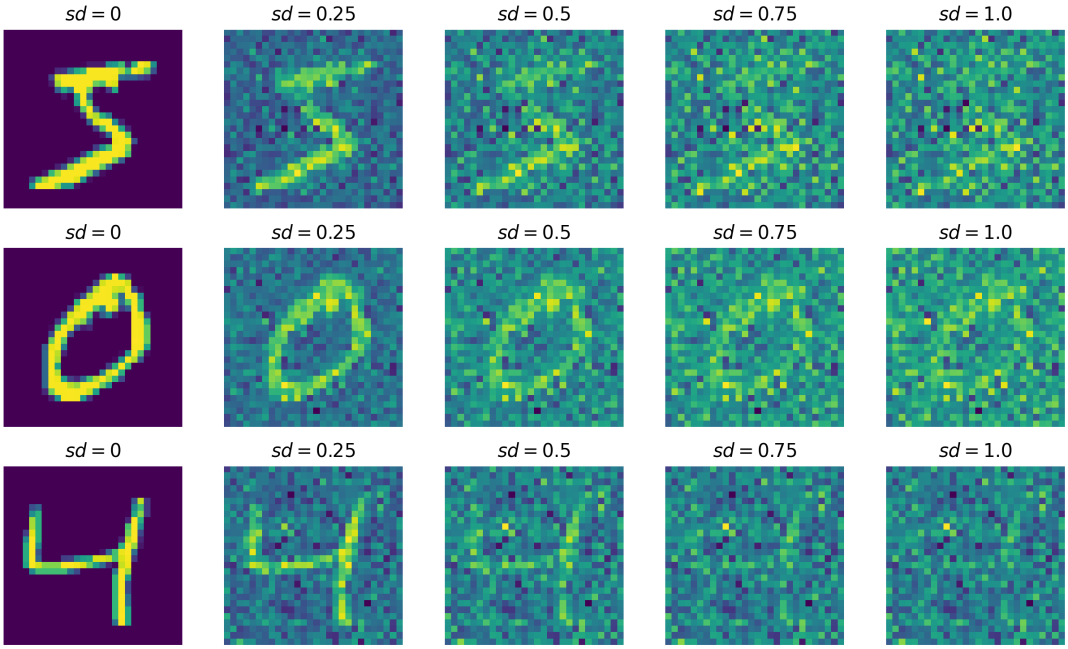
1. Decide on the network structure making sure to split activation functions into their own terms (as mentioned before)
2. Use Table 1to convert to the smoothed network
3. Create code for smooth network using appropriate regularization terms

Moving onto the details of our experiments, for all of our experiments we use the CNN architecture in Table 3. The resulting smooth CNN architecture can be found in Table . All training used a batch size of 1 and early stopping with a patience of 2 epochs based on validation loss (up to 25 epochs). For the smooth architecture, four regularization weights need to be chosen, one for each of the following: first ReLU layer (λ_2), first dense layer (λ_5), second ReLU layer (λ_6), and output layer (λ_7). Based on the construction of our particular CNN, no regularization term was needed for the convolutional layer. A coarse hyperparameter search was done for the learning rate (options: 10^{-3}, 10^{-4}, and 10^{-5}) and regularization coefficients (options: 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}) using 5-fold cross validation on the training set. The optimal hyperparameters that were used for the experiments are shown in Table . The default values for any additional hyperparameters available in Tensorflow, but not mentioned here were used for these methods. Additionally, the hyperparameters for when $\sigma = 0.01$ are those found for $\sigma = 0.1$, no additional search was performed for this σ-value. Finally, the learning rate for the CIFAR-10 experiments using Adam were reduced by a factor of 10^{-1} based on additional 5-fold cross validation on the training set using only the unsmoothed CNN. Code for the experiments is available at `https://github.com/acstarnes/GSmoothSGD`.

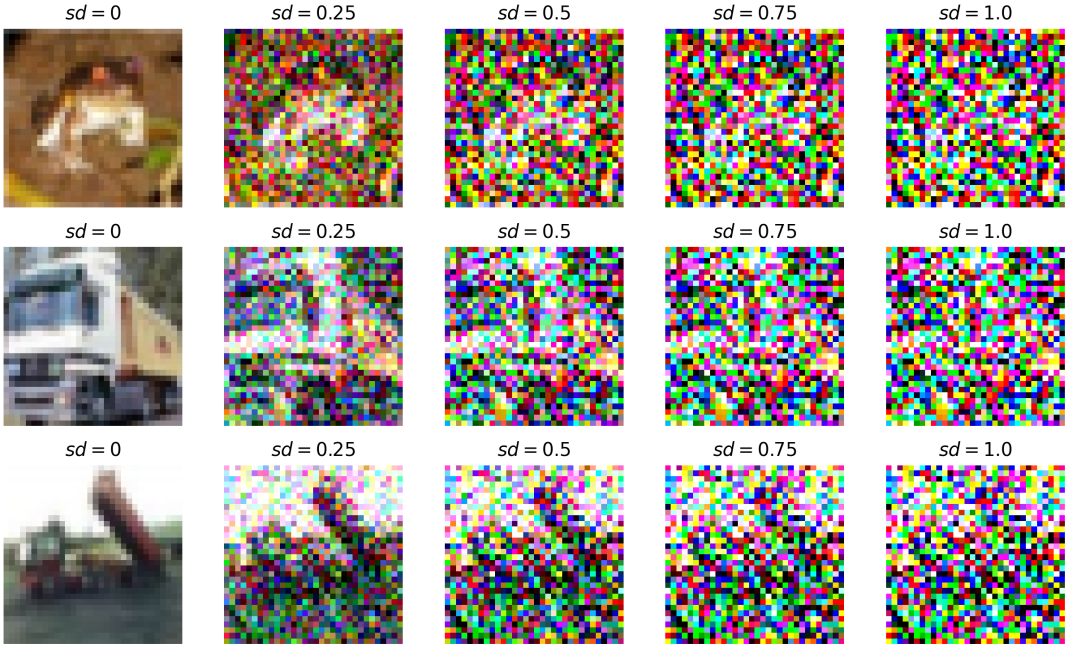
Table 3: CNN architecture for experiments

Layer Number	**Layer Name**	**Layer Properties**
0	Input	MNIST or CIFAR10 images
1	Convolution	32 4×4 convolutions with stride of 1 and 'valid' padding
2	Activation	ReLU
3	Pooling	Average pooling with 2×2 pool size and stride of 2
4	Flatten	
5	Dense	128 neurons
6	Activation	ReLU
7	Dense	10 neurons

We ran the MNIST and CIFAR-10 experiments using $\sigma = 0, 0.01, 0.1, 0.5, 1$. Figure 1a contains example of the noisy MNIST images and Figure 1b shows examples for CIFAR-10. The full heatmaps of the experiments are shown in Figures 2 and 3.



(a) MNIST Noisy Images



(b) CIFAR-10 Noisy Images

Figure 1: Examples of noisy images of MNIST and CIFAR-10 for noise standard deviations (sd)

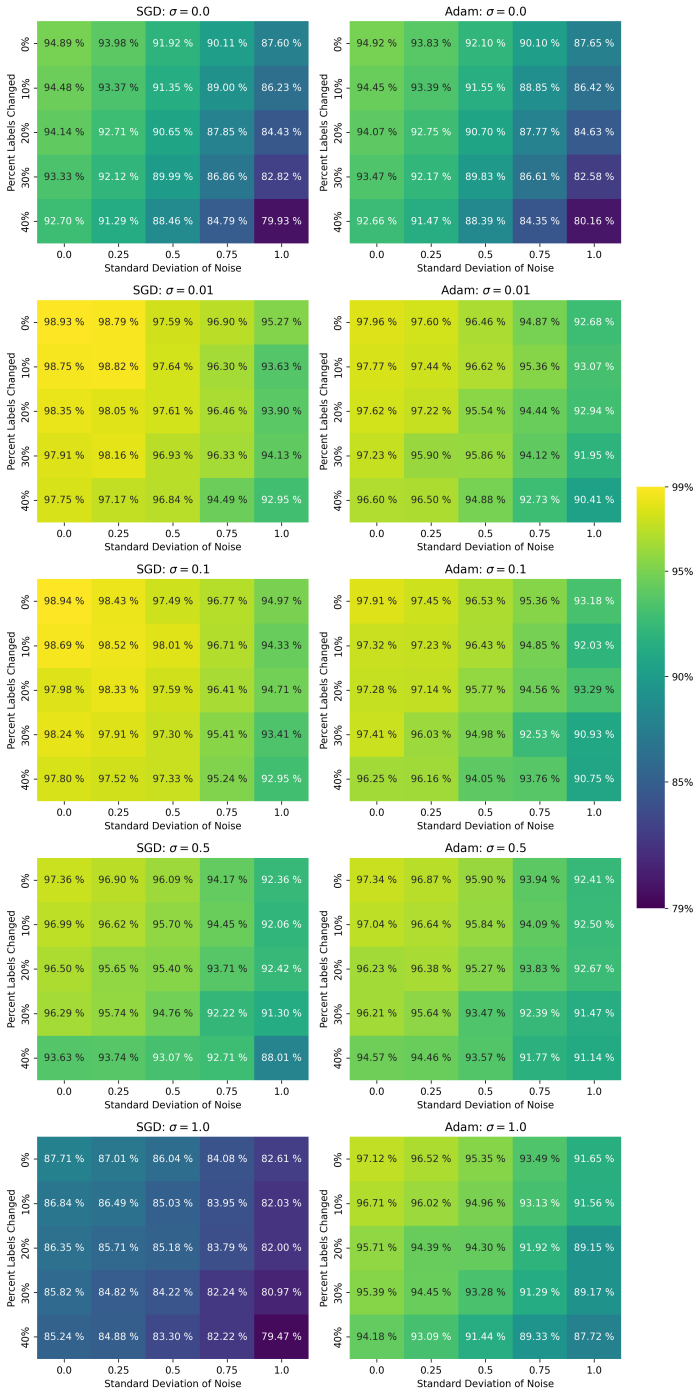


Figure 2: Test Accuracy for 25 SGD and Adam Experiments using MNIST

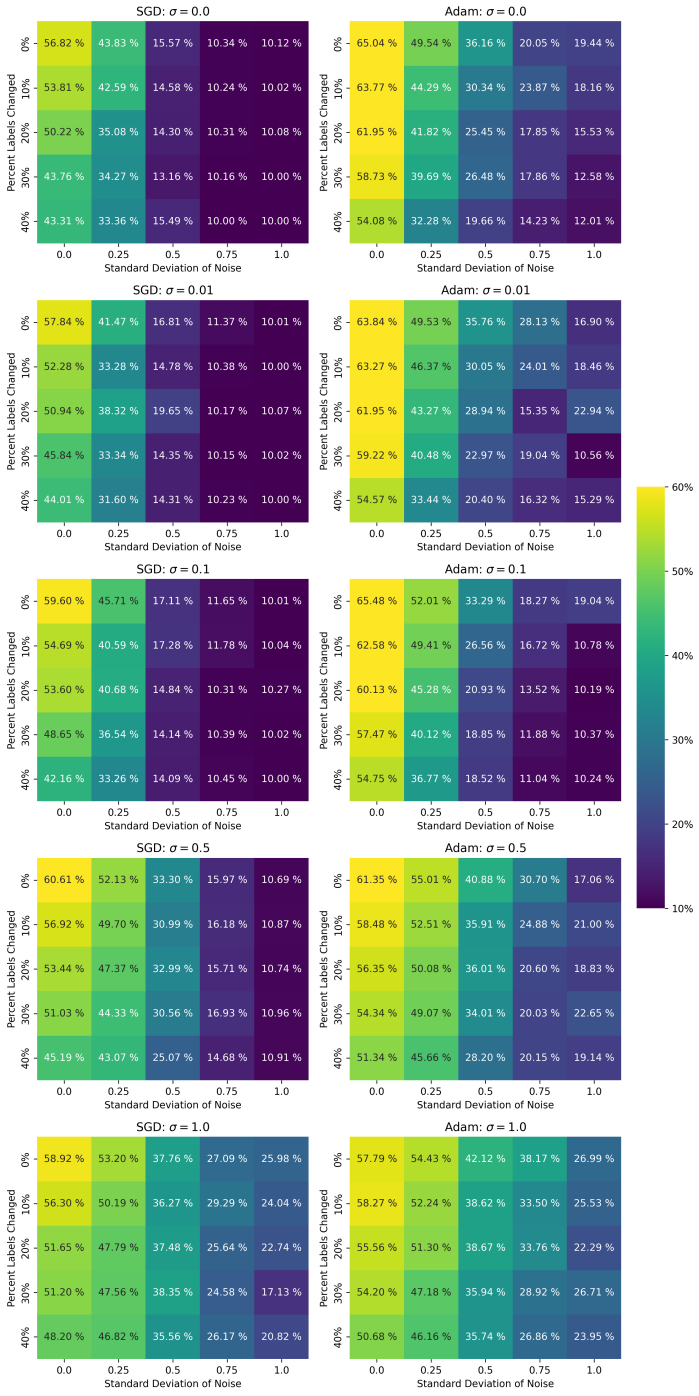


Figure 3: Test Accuracy for 25 SGD and Adam Experiments using CIFAR-10

Algorithm 1 GSmoothSVRG

Require: $\widetilde{\boldsymbol{x}}_0 \in \mathbb{R}^d$, $\sigma_s \geq 0$ for $s = 0, 1, ...$

for $s = 1, 2, ...$ **do**
  $\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{x}}_{s-1}$, $\boldsymbol{x}_0 = \widetilde{\boldsymbol{x}}$, $\tau = \sigma_s$
  $\widetilde{\boldsymbol{\mu}}_{\sigma_s} = \frac{1}{K}\sum_{i=1}^{K} \nabla f_{i,\sigma_s}(\widetilde{\boldsymbol{x}})$
  for $t = 1, ..., m$ **do**
    $i_t \sim \text{Unif}[K]$
    $\boldsymbol{v}_t^{\sigma_s,\tau} = \nabla f_{i_t,\sigma_s}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}_\tau$
    $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta \boldsymbol{v}_t^{\sigma_s,\tau}$
  end for
  $\widetilde{\boldsymbol{x}}_s = \boldsymbol{x}_t$ for $t \sim \text{Unif}[K]$
end for

F Gaussian smoothing stochastic variance reduced gradient (GSmoothSVRG)

Another modification of SGD that addresses this the variation in the gradient near a minimum is stochastic variance reduced gradient (SVRG) [3], which modifies the gradient used in the update step using inner and outer loops. The outer loop computes a control variate that reduces the variance of the iterates in the inner loop. The inner loop performs SGD steps, but with the control variate added to each step. So, the inner update becomes

$$\begin{aligned}\boldsymbol{v}_t &= \nabla f_{k_t}(\boldsymbol{x}_{t-1}) - \nabla f_{k_t}(\widetilde{\boldsymbol{x}}) + \widetilde{\mu}\\ \boldsymbol{x}_t &= \boldsymbol{x}_{t-1} - \eta \boldsymbol{v}_t,\end{aligned}$$

where $\widetilde{x}$ is the output of the previous inner iteration and $\widetilde{\mu}$ is the full gradient at $\widetilde{x}$. Since the motivation for SVRG is variance reduction, just like SGD, SVRG has a tendency converge to non-global minima, making it another good candidate for using Gaussian smoothing. GSmoothSGD suffers the same variance issues as SGD, so in order to combine the benefits of variance reduction and smoothing, we propose Gaussian smoothed SVRG (GSmoothSVRG) which can be found in Algorithm 1.

In this section, we provide convergence results for GSmoothSVRG and then provide numerical experiments that show how the variance is reduced compared to both SGD and GSmoothSGD.

F.1 Convergence of GSmoothSVRG

This section provides the proof of Theorem F.1 and the additional background results needed for the proof. The original convergence result for SVRG is stated for strongly convex function, we will do the same for GSmoothSVRG, which means we need to show that smoothing preserves strong convexity as well. Our second result shows just that.

Lemma F.1. *If f is γ-strongly convex, then so is f_σ.*

The proof is the same as the proof when f is convex (see Lemma 1) where only a minor modification is made for the strong convexity which now includes the quadratic function that f grows as fast as. We still provide the proof here.

Proof of Lemma F.1. Assume that f is γ-strongly convex. Then

$$\begin{aligned}
f_\sigma(t\boldsymbol{x}+(1-t)\boldsymbol{y}) &= \frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d} f(t\boldsymbol{x}+(1-t)\boldsymbol{y}+\sigma\boldsymbol{u})e^{-\|\boldsymbol{u}\|^2}\,du \\
&= \frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d} f\Big(t(\boldsymbol{x}+\sigma\boldsymbol{u})+(1-t)(\boldsymbol{y}+\sigma\boldsymbol{u})\Big)e^{-\|\boldsymbol{u}\|^2}\,du \\
&\le \frac{1}{\pi^{d/2}}\int_{\mathbb{R}^d} \Big(tf(\boldsymbol{x}+\sigma\boldsymbol{u})+(1-t)f(\boldsymbol{y}+\sigma\boldsymbol{u}) \\
&\qquad + \frac{\gamma}{2}t(1-t)\|(\boldsymbol{x}+\sigma\boldsymbol{u})-(\boldsymbol{y}+\sigma\boldsymbol{u})\|^2\Big)e^{-\|\boldsymbol{u}\|^2}\,du \\
&= tf_\sigma(\boldsymbol{x})+(1-t)f_\sigma(\boldsymbol{y})+\frac{\gamma}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2
\end{aligned}$$

shows that f_σ is also γ-strongly convex. □

We now show that adding variance reduction using SVRG to GSmoothSGD or equivalently smoothing SVRG does not change the convergence rate from the original SVRG for strongly convex and L-smooth functions. In particular, compared with GSmoothSGD, GSmoothSVRG will converge for a fixed learning rate.

Theorem F.1. *Consider GSmoothSVRG in Algorithm 1. Assume f_i is convex and L-smooth and $f=\frac{1}{K}\sum_{k=1}^K f_k$ is γ-strongly convex (for some $\gamma>0$). Assume m is sufficiently large so that*

$$\alpha=\frac{1+2L\eta^2 m}{\eta\gamma(1-2L\eta)m}<1.$$

Then

$$E(f(\widetilde{\boldsymbol{x}}_s)-f(\boldsymbol{x}_*))\le\alpha^s E(f_{\sigma_0}(\widetilde{\boldsymbol{x}}_0)-f(\boldsymbol{x}_*))+\frac{Ld}{2}\sum_{i=1}^s\alpha^i\max(0,\sigma_{i-1}^2-\sigma_i^2).$$

The proof of this theorem is broken into four lemmas whose proofs follow the same structure as in [3] with adaptions for smoothing. In particular, we mimic the proof that SVRG converges for strongly convex and L-smooth functions from [3], which is broken into four lemmmas, and smooth when necessary. The four lemmas build on each other culminating in the fact that the iterates are getting closer to the minimum. The proof of the theorem then iteratively applies this result to get the claimed bound. We begin by justifying the first lemma which states a bound between the difference in the gradients of the iterates and the minimum.

Lemma F.2. *For each $i\in\{1,...,K\}$, let f_i be L-smooth and convex. Then for any $\sigma\ge 0$,*

$$E(\|\nabla f_{i,\sigma}(\boldsymbol{x})-\nabla f_i(\boldsymbol{x}_*)\|^2)\le 2L(f_\sigma(\boldsymbol{x})-f(\boldsymbol{x}_*)).$$

Furthermore, for $\sigma\ge\tau\ge 0$,

$$E(\|\nabla f_{i,\sigma}(\boldsymbol{x})-\nabla f_{i,\tau}(\boldsymbol{x}_*^\tau)\|^2)\le 4L(f_\sigma(\boldsymbol{x})-f(\boldsymbol{x}_*))$$

where $\boldsymbol{x}_*^\tau$ is the minimizer of f_τ.

We can adapt the above to get the following statements as well:

$$E(\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_{i,\tau}(\boldsymbol{x}_*)\|^2) \leq 2L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)) + \frac{1}{2}\tau^2 L^2 d$$

$$E(\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_{i,\tau}(\boldsymbol{x})\|^2) \leq 4L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*))$$

Proof. Note that since each f_k is convex, f and $f_{k,\sigma}$ are convex for any $\sigma \geq 0$. This means that f_σ is also convex for any $\sigma \geq 0$. Let

$$g_i^\sigma(\boldsymbol{x}) = f_{i,\sigma}(\boldsymbol{x}) - f_i(\boldsymbol{x}_*) - \langle \nabla f_i(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle.$$

Then since f_i is convex,

$$f_{i,\sigma}(\boldsymbol{x}) - f_i(\boldsymbol{x}_*) \geq f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_*) \geq \langle \nabla f_i(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle.$$

This means $g_i^\sigma(\boldsymbol{x}) \geq 0$ for any i and σ. Since $f_{i,\sigma}$ is L-smooth, so is g_i^σ. So,

$$0 \leq g_i^\sigma\left(\boldsymbol{x} - \tfrac{1}{L}\nabla g_i^\sigma(\boldsymbol{x})\right) \leq g_i^\sigma(\boldsymbol{x}) - \frac{1}{2L}\|g_i^\sigma(\boldsymbol{x})\|^2$$

and rearranging we have

$$\|g_i^\sigma(\boldsymbol{x})\|^2 \leq 2L g_i^\sigma(\boldsymbol{x}).$$

Since

$$\nabla g_i^\sigma(\boldsymbol{x}) = \nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*),$$

we have

$$\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\|^2 \leq 2L\Big(f_{i,\sigma}(\boldsymbol{x}) - f_i(\boldsymbol{x}_*) - \langle \nabla f_i(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle\Big).$$

Therefore, taking the expectation over i,

$$\begin{aligned} E(\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\|^2) &\leq 2LE(f_{i,\sigma}(\boldsymbol{x}) - f_i(\boldsymbol{x}_*) - \langle \nabla f_i(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle) \\ &= 2L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)). \end{aligned}$$

The furthermore statement can be seen by

$$\begin{aligned} E(\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_{i,\tau}(\boldsymbol{x}_*^\tau)\|^2) &\leq E(\|\nabla f_{i,\sigma}(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}_*)\|^2) + E(\|\nabla f_{i,\tau}(\boldsymbol{x}_*^\tau) - \nabla f_i(\boldsymbol{x}_*)\|^2) \\ &\leq 2L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)) + 2L(f_\tau(\boldsymbol{x}_*^\tau) - f(\boldsymbol{x}_*)) \\ &\leq 2L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)) + 2L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)) \\ &= 4L(f_\sigma(\boldsymbol{x}) - f(\boldsymbol{x}_*)), \end{aligned}$$

since $f_\sigma(\boldsymbol{x}) \geq f_\tau(\boldsymbol{x}_*^\tau)$. □

Next, we bound the GSmoothSVRG gradient update by a linear combination of function outputs. The previous lemma is applied in the proof in order to derive the bound.

Lemma F.3. *For each $i \in \{1, ..., K\}$, let f_i be L-smooth and convex. For $\sigma \geq \tau \geq 0$,*

$$E(\|\boldsymbol{v}_t\|^2 | \boldsymbol{x}_{t-1}) \leq 4L(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*) + f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Recall that

$$\boldsymbol{v}_t^{\sigma,\tau} = \nabla f_{i_t,\sigma}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}_\tau.$$

This means

$$\begin{aligned} E(\boldsymbol{v}_t^{\sigma,\tau}|\boldsymbol{x}_{t-1}) &= \nabla f_\sigma(\boldsymbol{x}_{t-1}) - \nabla f_\tau(\widetilde{\boldsymbol{x}}) + \nabla f_\tau(\widetilde{\boldsymbol{x}}) \\ &= \nabla f_\sigma(\boldsymbol{x}_{t-1}). \end{aligned} \tag{11}$$

Proof. Observe

$$\begin{aligned} E(\|\boldsymbol{v}_t\|^2|\boldsymbol{x}_{t-1}) &= E(\|\nabla f_{i_t,\sigma}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}_\tau\|^2|\boldsymbol{x}_{t-1}) \\ &\leq E(\|\nabla f_{i_t,\sigma}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau)\|^2|\boldsymbol{x}_{t-1}) \\ &\quad + E(\|\nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}) + \widetilde{\boldsymbol{\mu}}_\tau\|^2|\boldsymbol{x}_{t-1}) \\ &\overset{(1)}{\leq} E(\|\nabla f_{i_t,\sigma}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau)\|^2|\boldsymbol{x}_{t-1}) \\ &\quad + E(\|\nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}) - E(f_{i_t,\tau}(\boldsymbol{x}_*^\tau) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}}))\|^2|\boldsymbol{x}_{t-1}) \\ &\overset{(2)}{\leq} E(\|\nabla f_{i_t,\sigma}(\boldsymbol{x}_{t-1}) - \nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau)\|^2|\boldsymbol{x}_{t-1}) + E(\|\nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau) - \nabla f_{i_t,\tau}(\widetilde{\boldsymbol{x}})\|^2|\boldsymbol{x}_{t-1}) \\ &\overset{\text{Lem F.2}}{\leq} 4L(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*)) + 4L(f_\tau(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \\ &\overset{(3)}{\leq} 4L(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*) + f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \end{aligned}$$

where step (1) is due to $E(\nabla f_{i_t,\tau}(\boldsymbol{x}_*^\tau)) = 0$, step (2) follows from $E(\|\xi - E(\xi)\|^2) = E(\|\xi\|^2) - \|E(\xi)\|^2 \leq E(\|\xi\|^2)$ for any random vector ξ, and step (3) is because f is convex and $\sigma \geq \tau$. □

With this bound on the GSmoothSVRG gradient update, we are ready to bound the expected value of the iterates.

Lemma F.4. *For each $i \in \{1, ..., K\}$, let f_i be L-smooth and convex. For $\sigma \geq \tau \geq 0$,*

$$2\eta(1 - 2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*)) \leq E(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2) + 4L\eta^2 mE(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Proof. First,

$$\begin{aligned} E(\|\boldsymbol{x}_t - \boldsymbol{x}_*\|^2|\boldsymbol{x}_{t-1}) &\overset{\text{def. } \boldsymbol{x}_t}{=} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_*\|^2 - 2\eta\langle\boldsymbol{x}_{t-1} - \boldsymbol{x}_*, E(\boldsymbol{v}_t|\boldsymbol{x}_{t-1})\rangle + \eta^2 E(\|\boldsymbol{v}_t\|^2|\boldsymbol{x}_{t-1}) \\ &\overset{\text{eqn. (11)}}{=} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_*\|^2 - 2\eta\langle\boldsymbol{x}_{t-1} - \boldsymbol{x}_*, \nabla f_\sigma(\boldsymbol{x}_{t-1})\rangle + \eta^2 E(\|\boldsymbol{v}_t\|^2|\boldsymbol{x}_{t-1}) \\ &\overset{\text{Lem. F.3}}{\leq} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_*\|^2 - 2\eta\langle\boldsymbol{x}_{t-1} - \boldsymbol{x}_*, \nabla f_\sigma(\boldsymbol{x}_{t-1})\rangle \\ &\quad + 4L\eta^2(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*) + f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \\ &\overset{\text{conv.}}{\leq} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_*\|^2 - 2\eta(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*)) \\ &\quad + 4L\eta^2(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*) + f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \\ &= \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_*\|^2 - 2\eta(1 - 2L\eta)(f_\sigma(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_*)) + 4L\eta^2(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)). \end{aligned}$$

Since $P(\widetilde{\boldsymbol{x}}_s = \boldsymbol{x}_t) = \frac{1}{m}$ for $t = 0, ..., m-1$, then

$$mE(f_\sigma(\widetilde{\boldsymbol{x}}_s)|\boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) = \sum_{t=0}^{m-1} f_\sigma(\boldsymbol{x}_t).$$

So, summing over the m steps gives

$$E(\|\boldsymbol{x}_m - \boldsymbol{x}_*\|^2 | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) \leq \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2 - 2\eta(1-2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*) | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) + 4L\eta^2 m(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Rearranging shows

$$E(\|\boldsymbol{x}_m - \boldsymbol{x}_*\|^2 | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) + 2\eta(1-2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*) | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) \leq \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2 + 4L\eta^2 m(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Since $\|\boldsymbol{x}_m - \boldsymbol{x}_*\|^2 \geq 0$,

$$2\eta(1-2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*) | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1}) \leq \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2 + 4L\eta^2 m(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Finally, taking the expectation gives

$$\begin{aligned}
&2\eta(1-2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*)) \\
&= 2\eta(1-2L\eta)mE(E(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*) | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1})) \\
&\leq E(E(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2 | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1})) + 4L\eta^2 mE(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*) | \boldsymbol{x}_0, ..., \boldsymbol{x}_{m-1})) \\
&= E(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2) + 4L\eta^2 mE(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).
\end{aligned}$$

□

All of the work so far now culminates in showing that, for strongly convex functions, expected difference between the iterates and minimum decreases by a multiplicative factor smaller than one each time.

Lemma F.5. *Assume f_i is convex and L-smooth and f is γ-strongly convex (for some $\gamma > 0$). For $\sigma \geq \tau \geq 0$, if f is γ-strongly convex, then*

$$E(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*)) \leq \frac{1 + 2L\eta^2 m}{\eta\gamma(1-2L\eta)m} E(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Proof. Since f is γ-strongly convex, so is f_σ. As $\boldsymbol{x}_0 = \widetilde{\boldsymbol{x}}$,

$$E(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2) \leq \frac{2}{\gamma} E(f_\sigma(\widetilde{\boldsymbol{x}}) - f_\sigma(\boldsymbol{x}_*)) \leq \frac{2}{\gamma} E(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).$$

Combining this with Lemma F.4 shows

$$\begin{aligned}
2\eta(1-2L\eta)mE(f_\sigma(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*)) &\leq E(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|^2) + 4L\eta^2 mE(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \\
&\leq \frac{2}{\gamma} E(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) + 4L\eta^2 mE(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)) \\
&= 2\left(\frac{1}{\gamma} - 2L\eta^2 m\right) E(f_\sigma(\widetilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*)).
\end{aligned}$$

Arithmetic gives the result. □

With the result of the previous lemma in hand, we are finally ready to prove that GSmoothSVRG converges.

Proof of Theorem F.1. Using Lemma F.5, we have

$$\begin{aligned} E(f_{\sigma_s}(\widetilde{\boldsymbol{x}}_s) - f(\boldsymbol{x}_*)) &\leq \alpha E(f_{\sigma_s}(\widetilde{\boldsymbol{x}}_{s-1}) - f(\boldsymbol{x}_*)) \\ &\leq \alpha E(f_{\sigma_{s-1}}(\widetilde{\boldsymbol{x}}_{s-1}) - f(\boldsymbol{x}_*)) + \frac{Ld}{2}\alpha \max(0, \sigma_{s-1}^2 - \sigma_s^2) \\ &\vdots \\ &\leq \alpha^s E(f_{\sigma_0}(\widetilde{\boldsymbol{x}}_0) - f(\boldsymbol{x}_*)) + \frac{Ld}{2}\sum_{i=1}^{s} \alpha^i \max(0, \sigma_{i-1}^2 - \sigma_i^2). \end{aligned}$$

□

Note that the only condition that τ in GSmoothSVRG needs to satisfy is $\sigma_t \geq \tau \geq 0$ at each iteration. The two obvious choices for τ are σ_s or 0. If $\tau = \sigma_s$, then we are performing SVRG on f_{σ_s}. On the other hand, if $\tau = 0$, then we are making the control variate of SVRG include information about the gradient of the original, non-smoothed function.

F.2 Numerical Experiments for GSmoothSVRG

For the GSmoothSVRG experiments, we repeat the noisy MNIST experiments using a batch size of 1. Since our primary focus was to compare variance of the norm of the iterative update and the rate of convergence between GSmoothSVRG and SGD, we only trained for one epoch (of 5000 steps in total). Additionally, we wanted to use the same learning rate for both algorithms; after a hyperparameter search, we chose to use 0.01 for both methods. A hyperparameter search was also done to pick regularization weights as well as the smoothing parameter for the inner GSmoothSVRG update. The regularization weights for GSmoothSVRG were 10^{-13} for both the first ReLU layer and first dense layer and 10^{-11} for both the second ReLU layer and the output layer, and we use the same smoothing parameter for both the inner and outer updates. The standard deviation of the norm of the gradient update for all of the GSmoothSVRG experiments can be found in Figure 6.

References

[1] Christopher De Sa. Lecture 5: Stochastic gradient descent. https://www.cs.cornell.edu-/courses/cs4787/2019sp/notes/lecture5.pdf, 2019. Accessed: 2024-09-26.

[2] Meixuan He, Yuqing Liang, Jinlan Liu, and Dongpo Xu. Convergence of adam for non-convex objectives: Relaxed hyperparameters and non-ergodic case. *arXiv preprint arXiv 2307.11782*, 2023.

[3] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26:315–323, 2013.

[4] Hossein Mobahi. Training recurrent neural networks by diffusion. *arXiv preprint arXiv:1601.04114*, 2016.

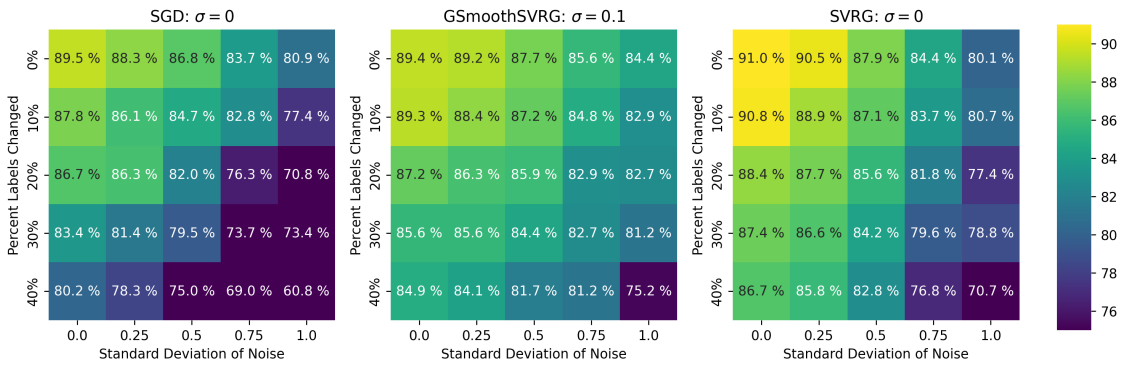


Figure 4: SVRG Noise Heatmaps

[5] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[6] Andrew Starnes, Anton Dereventsov, and Clayton Webster. Gaussian smoothing gradient descent for minimizing high-dimensional non-convex functions. *arXiv preprint arXiv 2311.00521*, 2023.

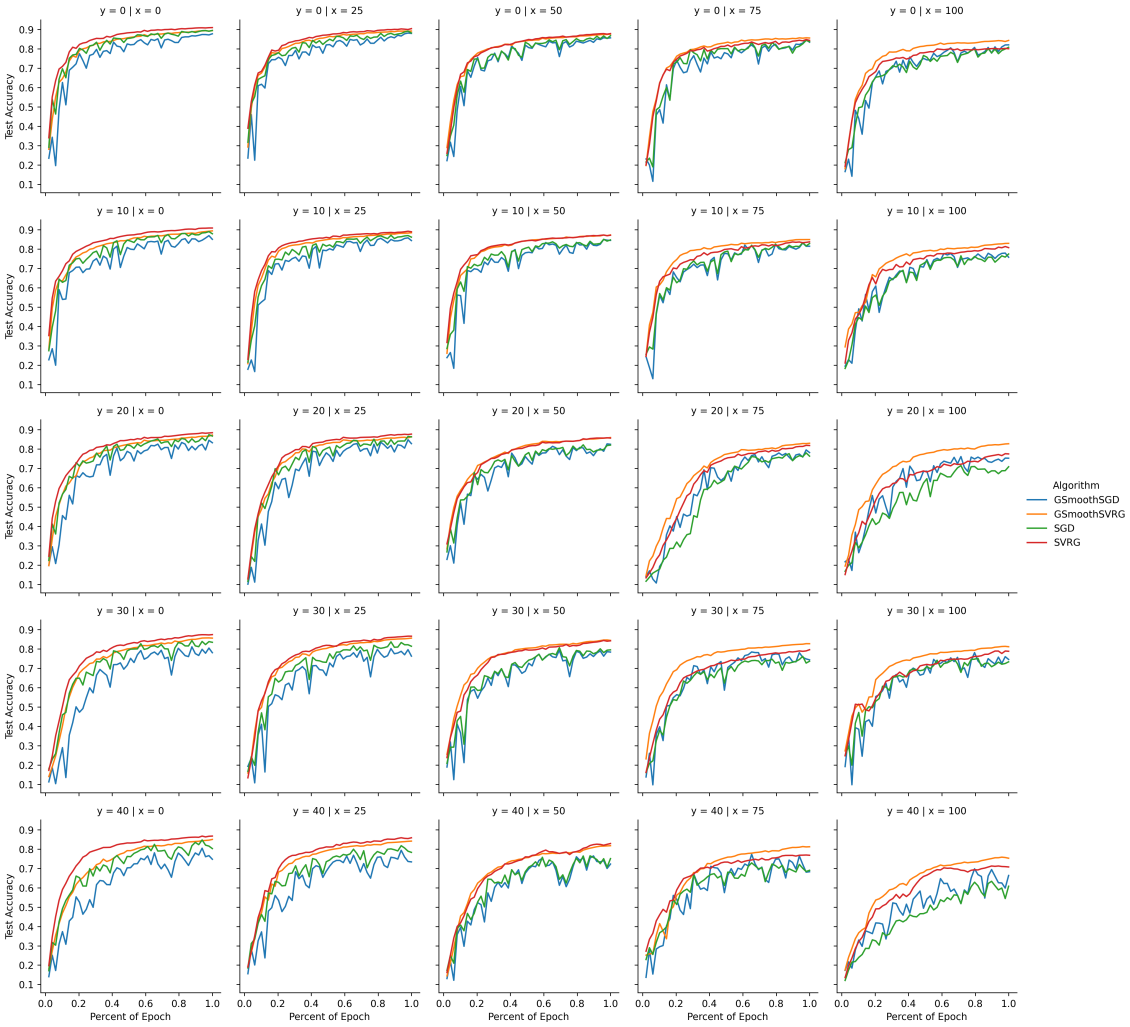


Figure 5: SVRG Noise Heatmap Accuracy

Figure 6: SVRG Noise Heatmap Update Standard Deviation