A. Omitted Proofs from Section 3

Proof of Lemma 3.2. Let $m_k(\mathbf{x}) = \sum_{i=1}^k a_i \left\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i \right\rangle + D_\psi(\mathbf{u}, \mathbf{x}_0)$ denote the function under the minimum in the lower bound. By Proposition 3.1, $\mathbf{v}_k = \nabla\psi^*(\mathbf{z}_k) = \arg\min_{\mathbf{x}\in\mathcal{X}} m_k(\mathbf{x})$. Observe that $m_k(\mathbf{x}) = a_k \left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \right\rangle + m_{k-1}(\mathbf{x})$. By the definition of Bregman divergence:

$$m_{k-1}(\mathbf{v}_k) = m_{k-1}(\mathbf{v}_{k-1}) + \langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{v}_k - \mathbf{v}_{k-1}\rangle + D_{m_{k-1}}(\mathbf{v}_k, \mathbf{v}_{k-1}).$$

As Bregman divergence is blind to linear and zero-order terms, we have that $D_{m_{k-1}}(\mathbf{v}_k, \mathbf{v}_{k-1}) = D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1})$. By Proposition 3.1, $\mathbf{v}_{k-1} = \arg\min_{\mathbf{x}\in\mathcal{X}} m_{k-1}(\mathbf{x})$, and hence $\langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{v}_k - \mathbf{v}_{k-1}\rangle \geq 0$. Therefore,

$$m_k(\mathbf{v}_k) \geq m_{k-1}(\mathbf{v}_{k-1}) + a_k \left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \right\rangle + D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}).$$

Using the definition of $\widetilde{\nabla} f(\mathbf{x}_k)$, the change in the lower bound is:

$$A_k L_k - A_{k-1} L_{k-1} \geq a_k f(\mathbf{x}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) - a_k \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\rangle . \tag{A.1}$$

For the change in the upper bound, we have:

$$\begin{aligned} A_k U_k - A_{k-1} U_{k-1} =& A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}) \\ =& a_k f(\mathbf{x}_k) + A_k (f(\mathbf{y}_k) - f(\mathbf{x}_k)) + A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})). \end{aligned} \tag{A.2}$$

By convexity of $f(\cdot)$:

$$f(\mathbf{x}_k) - f(\mathbf{y}_{k-1}) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_{k-1}\rangle . \tag{A.3}$$

Combining (A.1)-(A.3) and (AGD+):

$$A_k G_k - A_{k-1} G_{k-1} \leq a_k \langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\rangle - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) + A_k (f(\mathbf{y}_k) - f(\mathbf{x}_k)) - A_k \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle ,$$

as claimed. □

B. AGD+ for Smooth and Strongly Convex Minimization

Here we show that AGD+ can be extended to the setting of smooth and strongly convex minimization. As is customary (Nesterov, 2013), in this setting we assume that $\|\cdot\| = \|\cdot\|_2$ so that $f(\cdot)$ is L-smooth and μ-strongly convex w.r.t. the ℓ_2 norm, for $L < \infty$ and $\mu > 0$. To distinguish from the non-strongly-convex case, we refer to AGD+ for smooth and strongly convex minimization as μAGD+.

To analyze AGD+ in this setting, we need to use a stronger lower bound L_k, which is constructed by the same arguments as before, but now using strong convexity instead of regular convexity. Such a construction gives:

$$L_k = \frac{\sum_{i=1}^k a_i f(\mathbf{x}_i) + \min_{\mathbf{u}\in\mathcal{X}} m(\mathbf{u}) - \sum_{i=1}^k a_i \langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{x}_i\rangle - D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k}, \tag{B.1}$$

where

$$m_k(\mathbf{u}) = \frac{\sum_{i=1}^k a_i \left(\langle \nabla f(\mathbf{x}_i), \mathbf{u} - \mathbf{x}_i\rangle + \frac{\mu}{2}\|\mathbf{u} - \mathbf{x}_i\|^2\right) + D_\psi(\mathbf{u}, \mathbf{x}_0)}{A_k}.$$

While it suffices to have ψ be an arbitrary function that is strongly convex w.r.t. the $\|\cdot\|_2$, for simplicity, we take $\psi(\mathbf{x}) = \frac{\mu_0}{2}\|\mathbf{x}\|^2$, where μ_0 will be specified later.

For $\theta_k = \frac{a_k}{A_k}$, the algorithm can now be stated as follows:

$$\boxed{\begin{aligned} \mathbf{v}_k &= \operatorname*{argmin}_{\mathbf{u}\in\mathcal{X}} m_k(\mathbf{u}), \\ \mathbf{x}_k &= \frac{1}{1+\theta_k}\mathbf{y}_{k-1} + \frac{\theta_k}{1+\theta_k}\mathbf{v}_{k-1}, \\ \mathbf{y}_k &= (1-\theta_k)\mathbf{y}_{k-1} + \theta_k \mathbf{v}_k, \end{aligned} \qquad (\mu\text{AGD+})}$$

where, $\mathbf{x}_1 = \mathbf{x}_0 = \mathbf{y}_0 = \mathbf{v}_0$ is an arbitrary initial point from $\mathcal{X}$.

As before, the main convergence argument is to show that $A_k G_k \leq A_{k-1} G_{k-1}$ and combine it with the bound on the initial gap G_1. We start with bounding the initial gap, as follows.

Proposition B.1. *If $\psi(\mathbf{x}) = \frac{\mu_0}{2}\|\mathbf{x}\|^2$, where $\mu_0 = a_1(L-\mu)$, then $A_1 G_1 \leq \frac{A_1(L-\mu)}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + E_1^\eta$, where $E_1^\eta = a_1 \langle \boldsymbol{\eta}_1, \mathbf{x}_* - \mathbf{v}_1\rangle$.*

Proof. As $\mathbf{x}_1 = \mathbf{x}_0$, the initial lower bound is:

$$A_1 L_1 = a_1 f(\mathbf{x}_1) + a_1 \langle \nabla f(\mathbf{x}_1), \mathbf{v}_1 - \mathbf{x}_1\rangle + \left(\frac{a_1\mu}{2} + \frac{\mu_0}{2}\right)\|\mathbf{v}_1 - \mathbf{x}_1\|^2 - \frac{\mu_0}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 - a_1\langle \boldsymbol{\eta}_1, \mathbf{x}_* - \mathbf{x}_1\rangle .$$

As $a_1 = A_1$, it follows that $\mathbf{y}_1 = \mathbf{v}_1$, and hence:

$$\begin{aligned} A_1 U_1 =& a_1 f(\mathbf{v}_1) \\ \leq& a_1 f(\mathbf{x}_1) + a_1 \langle \nabla f(\mathbf{x}_1), \mathbf{v}_1 - \mathbf{x}_1\rangle + \frac{a_1 L}{2}\|\mathbf{v}_1 - \mathbf{x}_1\|^2 \\ =& a_1 f(\mathbf{x}_1) + a_1 \left\langle \widetilde{\nabla} f(\mathbf{x}_1), \mathbf{v}_1 - \mathbf{x}_1\right\rangle + \frac{a_1 L}{2}\|\mathbf{v}_1 - \mathbf{x}_1\|^2 - a_1\langle \boldsymbol{\eta}_1, \mathbf{v}_1 - \mathbf{x}_1\rangle , \end{aligned}$$

where the inequality is by the smoothness of $f(\cdot)$. Combining the bounds on the initial upper and lower bounds, it follows:

$$\begin{aligned} A_1 G_1 &\leq \frac{\mu_0}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + \frac{a_1 L - a_1\mu - \mu_0}{2}\|\mathbf{v}_1 - \mathbf{x}_1\|^2 + a_1\langle \boldsymbol{\eta}_1, \mathbf{x}_* - \mathbf{v}_1\rangle \\ &\leq \frac{a_1(L-\mu)}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + E_1^\eta, \end{aligned}$$

as, by the initial assumption, $\mu_0 = a_1(\mu - L)$ and $E_1^\eta = a_1\langle \boldsymbol{\eta}_1, \mathbf{x}_* - \mathbf{v}_1\rangle$. □

To bound the change in the lower bound, it is useful to first bound $m_k(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1})$, as in the following technical proposition.

Proposition B.2. *Let $\psi(\mathbf{x}) = \frac{a_1(L-\mu)}{2}\|\mathbf{x}\|^2$. Then:*

$$m_k(\mathbf{v}_k) \geq m_{k-1}(\mathbf{v}_{k-1}) + a_k\left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\right\rangle + \frac{A_{k-1}\mu}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 + \frac{a_k\mu}{2}\|\mathbf{v}_k - \mathbf{x}_k\|^2.$$

Proof. Observe that, by the definition of $m_k(\cdot)$, $m_k(\mathbf{v}_k) = m_{k-1}(\mathbf{v}_k) + a_k\left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\right\rangle + a_k\frac{\mu}{2}\|\mathbf{v}_k - \mathbf{x}_k\|$.

The rest of the proof bounds $m_{k-1}(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1})$. Observe that, as $\mathbf{v}_{k-1} = \operatorname{argmin}_{\mathbf{u}\in\mathcal{X}} m_{k-1}(\mathbf{u})$, it must be $\langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{u} - \mathbf{v}_{k-1}\rangle \geq 0$, $\forall \mathbf{u} \in \mathcal{X}$. As Bregman divergence is blind to linear terms:

$$\begin{aligned} m_{k-1}(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1}) &= \langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{u} - \mathbf{v}_{k-1}\rangle + D_{m_{k-1}}(\mathbf{v}_k, \mathbf{v}_{k-1}) \\ &\geq D_{m_{k-1}}(\mathbf{v}_k, \mathbf{v}_{k-1}) \\ &= \frac{A_{k-1}\mu}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 + \frac{a_1(L-\mu)}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2. \end{aligned}$$

The rest of the proof is by $\frac{a_1(L-\mu)}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \geq 0$. □

We are now ready to move to the main part of the convergence argument, namely, to show that $A_k G_k \leq A_{k-1} G_{k-1}$ for a certain choice of a_k.

Lemma B.3. *Let $\psi(\mathbf{x}) = \frac{a_1(L-\mu)}{2}\|\mathbf{x}\|^2$ and $0 < a_k \leq A_k\sqrt{\frac{\mu}{L}}$. Then: $A_k G_k \leq A_{k-1} G_{k-1} + E_k^\eta$, where $E_k^\eta = a_k\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\rangle$.*

Proof. As $U_k = f(\mathbf{y}_k)$, we have that:

$$A_k U_k - A_{k-1} U_{k-1} \leq A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}). \tag{B.2}$$

Using Proposition B.2, the change in the lower bound is:

$$\begin{aligned}A_kL_k - A_{k-1}L_{k-1} \geq & a_k f(\mathbf{x}_k) + a_k\left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\right\rangle + \frac{A_{k-1}\mu}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 + \frac{a_k\mu}{2}\|\mathbf{v}_k - \mathbf{x}_k\|^2 \\ & + a_k\left\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{x}_k\right\rangle.\end{aligned} \tag{B.3}$$

Denote $\mathbf{w}_k = \frac{A_{k-1}}{A_k}\mathbf{v}_{k-1} + \frac{a_k}{A_k}\mathbf{x}_k$. By Jensen's Inequality:

$$\frac{A_{k-1}\mu}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 + \frac{a_k\mu}{2}\|\mathbf{v}_k - \mathbf{x}_k\|^2 \geq \frac{\mu A_k}{2}\|\mathbf{v}_k - \mathbf{w}_k\|^2. \tag{B.4}$$

Write $a_k\left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\right\rangle$ as:

$$a_k\left\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\right\rangle = a_k\left\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{w}_k\right\rangle + a_k\left\langle \nabla f(\mathbf{x}_k), \frac{A_{k-1}}{A_k}(\mathbf{v}_{k-1} - \mathbf{x}_k)\right\rangle + a_k\left\langle \boldsymbol{\eta}_k, \mathbf{v}_k - \mathbf{x}_k\right\rangle. \tag{B.5}$$

As $\frac{a_k}{A_k} \leq \sqrt{\frac{\mu}{L}}$ and by smoothness of $f(\cdot)$:

$$\begin{aligned}a_k\left\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{w}_k\right\rangle + \frac{\mu A_k}{2}\|\mathbf{v}_k - \mathbf{w}_k\|^2 &\geq A_k\left(\left\langle \nabla f(\mathbf{x}_k), \left(\mathbf{x}_k + \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{w}_k)\right) - \mathbf{x}_k\right\rangle + \frac{L}{2}\|\frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{w}_k)\|^2\right) \\ &\geq A_k\left(f\left(\mathbf{x}_k + \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{w}_k)\right) - f(\mathbf{x}_k)\right).\end{aligned} \tag{B.6}$$

Combining (B.3)-(B.6), we have the following bound for the change in the lower bound:

$$\begin{aligned}A_kL_k - A_{k-1}L_{k-1} \geq & A_k f\left(\mathbf{x}_k + \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{w}_k)\right) - A_{k-1}f(\mathbf{x}_k) + a_k\left\langle \nabla f(\mathbf{x}_k), \frac{A_{k-1}}{A_k}(\mathbf{v}_{k-1} - \mathbf{x}_k)\right\rangle \\ & + a_k\left\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\right\rangle.\end{aligned}$$

Using the definition of $\mathbf{w}_k$, $\theta_k = \frac{a_k}{A_k}$, and ($\mu$AGD+), it is not hard to show that $\mathbf{y}_k = \mathbf{x}_k + \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{w}_k)$ and $\frac{a_k}{A_k}(\mathbf{v}_{k-1} - \mathbf{x}_k) = \mathbf{x}_k - \mathbf{y}_{k-1}$, which, using the convexity of $f(\cdot)$, gives:

$$A_kL_k - A_{k-1}L_{k-1} \geq A_k f(\mathbf{y}_k) - A_{k-1}f(\mathbf{y}_{k-1}) - a_k\left\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\right\rangle. \tag{B.7}$$

Combining (B.7) and (B.2), the proof follows. □

Theorem B.4. *Let $\psi(\mathbf{x}) = \frac{L-\mu}{2}\|\mathbf{x}\|^2$, $a_1 = 1$, $\frac{a_i}{A_i} = \gamma_i \leq \sqrt{\frac{\mu}{L}}$ for $i \geq 2$, and let $\mathbf{y}_k, \mathbf{x}_k$ evolve according to (μAGD+). Then, $\forall k \geq 1$:*

$$\begin{aligned}f(\mathbf{y}_k) - f(\mathbf{x}_*) &= \frac{A_1}{A_k}\cdot\frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \frac{\sum_{i=1}^k a_i\left\langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{v}_i\right\rangle}{A_k} \\ &\leq \left(\Pi_{i=1}^k(1-\gamma_i)\right)\frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \frac{\sum_{i=1}^k a_i\left\langle \boldsymbol{\eta}_i, \mathbf{x}_* - \mathbf{v}_i\right\rangle}{A_k}.\end{aligned}$$

Proof. Applying Lemma B.3, it follows that $G_k \leq \frac{A_1G_1}{A_k} + \frac{\sum_{i=1}^k E_i^\eta}{A_k} = \frac{A_1}{A_2}\cdot\frac{A_2}{A_3}\cdot\dots\cdot\frac{A_{k-1}}{A_k}G_1 + \frac{\sum_{i=1}^k E_i^\eta}{A_k}$. As $\frac{A_{i-1}}{A_i} = 1 - \frac{a_i}{A_i} = 1 - \gamma_i$, we have $G_k \leq \left(\Pi_{i=1}^k(1-\gamma_i)\right)G_1 + \frac{\sum_{i=1}^k E_i^\eta}{A_k}$. The rest of the proof is by applying Proposition B.1 and using that $f(\mathbf{y}_k) - f(\mathbf{x}_*) \leq G_k$. □

Using the same arguments for bounding the noise term as in the case of smooth minimization (Section 3), we have the following corollary.

Corollary B.5 (of Theorem B.4)**.** *If $\boldsymbol{\eta}_i = \mathbf{0}$ (the noiseless gradient case), setting $\gamma_i = \sqrt{\frac{\mu}{L}}$, we recover the standard convergence result for accelerated smooth and strongly convex minimization:*

$$f(\mathbf{y}_k) - f(\mathbf{x}_*) \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{L-\mu}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2.$$

If $\mathbb{E}[\|\boldsymbol{\eta}_i\|] \leq M_i$ and $\max_{\mathbf{u}\in\mathcal{X}} \|\mathbf{x}_* - \mathbf{u}\| \leq R_{\mathbf{x}_*}$, then setting $\gamma_i = \sqrt{\frac{\mu}{L}}$:

$$\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{L-\mu}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + \frac{R_{\mathbf{x}_*}\sum_{i=1}^k a_i M_i}{A_k}.$$

If $\boldsymbol{\eta}_i$'s are zero-mean and independent and $\mathbb{E}[\|\boldsymbol{\eta}_i\|^2] \leq \sigma^2$, then:

$$\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] \leq \left(\Pi_{i=1}^k (1-\gamma_i)\right)\frac{L-\mu}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + \frac{\sigma^2\sum_{i=1}^k \frac{{a_i}^2}{A_i}}{\mu A_k}$$

In particular, setting:

- $\frac{a_i}{A_i} = \gamma_i = \sqrt{\frac{\mu}{L}}$,
$$\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{L-\mu}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + \frac{\sigma^2}{\sqrt{\mu L}}$$
- $a_i = i^p$ *for* $p \in \mathbb{Z}_+$,
$$\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] = O\left(\frac{p+1}{k^{p+1}} \cdot \frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \frac{(p+1)^2}{pk}\cdot\frac{\sigma^2}{\mu}\right).$$

Proof. The bounds for $\boldsymbol{\eta}_i = \mathbf{0}$ and for $\mathbb{E}[\|\boldsymbol{\eta}_i\|] \leq M_i$ and $\max_{\mathbf{u}\in\mathcal{X}} \|\mathbf{x}_* - \mathbf{u}\| \leq R_{\mathbf{x}_*}$ are straightforward.

Assume that $\boldsymbol{\eta}_i$'s are zero-mean and independent and denote $\psi_k(\mathbf{x}) = \sum_{i=1}^k a_i \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_i\|^2 + \frac{\mu_0}{2}\|\mathbf{x} - \mathbf{x}_0\|^2$. Observe that the strong convexity parameter of ψ_k is $\mu A_k + \mu_0 > \mu A_k$. From Fact 2.4, $\mathbf{v}_k = \nabla\psi_k^*(\mathbf{z}_k)$. Similarly as for the case of smooth minimization, let $\hat{\mathbf{v}}_k = \nabla\psi^*(\mathbf{z}_k + a_k\boldsymbol{\eta}_k)$. Then $\hat{\mathbf{v}}_k$ is independent of $\boldsymbol{\eta}_k$, and, using Fact 2.5, we have:

$$\begin{aligned}\mathbb{E}[a_k\langle\boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k\rangle] &= \mathbb{E}[a_k\langle\boldsymbol{\eta}_k, \mathbf{x}_* - \hat{\mathbf{v}}_k\rangle] + \mathbb{E}[a_k\langle\boldsymbol{\eta}_k, \hat{\mathbf{v}}_k - \mathbf{v}_k\rangle]\\ &\leq \frac{{a_k}^2}{\mu A_k}\|\boldsymbol{\eta}_k\|^2.\end{aligned}$$

Combining with Theorem B.4, we get $\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] \leq \left(\Pi_{i=1}^k (1-\gamma_i)\right)\frac{L-\mu}{2}\|\mathbf{x}_* - \mathbf{x}_0\|^2 + \frac{\sigma^2\sum_{i=1}^k \frac{{a_i}^2}{A_i}}{\mu A_k}$. The rest of the proof follows by plugging in particular choices of a_i. □

Let us make a few more remarks here. When $\boldsymbol{\eta}_i$'s are zero-mean, independent, and $\mathbb{E}[\|\boldsymbol{\eta}_i\|^2] \leq \sigma^2$, even the vanilla version of μAGD+ does not accumulate noise (the noise averages out). Under the same assumptions and when $a_i = i$, we recover the asymptotic bound from (Ghadimi & Lan, 2012).[7] More generally, $a_i = i^p$ for any constant integer p gives a convergence bound for which the deterministic term vanishes at rate $1/k^{p+1}$ while the noise term vanishes at rate $1/k$. When $p = \log(k)$ for a fixed number of iterations k of μAGD+, we get $\mathbb{E}[f(\mathbf{y}_k) - f(\mathbf{x}_*)] = O\left(\frac{\log(k)}{k^{\log(k)}} \cdot \frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \frac{\log(k)}{k}\cdot\frac{\sigma^2}{\mu}\right)$, i.e., the deterministic term (independent of noise) decreases super-polynomially with the iteration count, while the noise term decreases at rate $\log(k)/k$. This is a much stronger bound than the one from (Ghadimi & Lan, 2012) and closer to the theoretical lower bound $\Omega\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \cdot \frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \cdot\frac{\sigma^2}{\mu k}\right)$ from (Nemirovskii & Yudin, 1983).

Note that in the setting of *constrained* (bounded-diameter) minimization, (Ghadimi & Lan, 2013) obtained the optimal convergence bound $O\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \cdot \frac{(L-\mu)\|\mathbf{x}_* - \mathbf{x}_0\|^2}{2} + \cdot\frac{\sigma^2}{\mu k}\right)$ by coupling the algorithm from (Ghadimi & Lan, 2012) with a domain-shrinking procedure resulting in a multi-stage algorithm. We expect it is possible to obtain a similar result for μAGD+ by coupling it with the domain-shrinking from (Ghadimi & Lan, 2013).

[7] Note that the independence of $\boldsymbol{\eta}_i$'s is a stronger assumption than used in (Ghadimi & Lan, 2012). Nevertheless, we can obtain the same bounds as stated in Corollary B.5 for the same assumptions as in (Ghadimi & Lan, 2012) using the same arguments as for the case of smooth minimization (see Appendix C.2).

C. Different Models of Inexact Oracle

C.1. Adversarial Models

There are two main adversarial models of inexact gradient oracles that have been used in the convergence analysis of accelerated methods: the approximate gradient model of (d'Aspremont, 2008) and inexact first-order oracle of (Devolder et al., 2014). The approximate gradient model (d'Aspremont, 2008) defines the inexact oracle by a deterministic perturbation satisfying the following condition for all queries:

$$|\langle \boldsymbol{\eta}, \mathbf{y} - \mathbf{z} \rangle| \leq \delta, \ \forall \mathbf{y}, \mathbf{z} \in \mathcal{X}.$$

Hence, this model is only applicable to constrained optimization with bounded-diameter domain and bounded (adversarial) additive noise. Under these assumptions, (d'Aspremont, 2008) proves that it is possible to approximate $f(\mathbf{x}_*)$ up to an error of δ achieving an accelerated rate. We can show the same asymptotic bound[8] by applying the assumption to Equation (3.2) in Proposition 3.5. This yields:

$$G_k \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + \delta$$

whenever $\frac{{a_k}^2}{A_k} \leq \frac{\mu}{L}$ for all k. Setting $a_k = \frac{\mu}{L} \cdot \frac{k+1}{2}$ yields $A_k = k^2 + O(k)$, which establishes the accelerated decrease of the first term in the error bound above.

The inexact first-order oracle (Devolder et al., 2014) is a generalization of the model from (d'Aspremont, 2008) that defines the inexact oracle by:

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

As stated in (Devolder et al., 2014), this model does not apply to noise-corrupted gradients *per se*, but rather to "non-smooth and weakly smooth convex problems". In other words, the model was introduced to characterize the behavior of accelerated methods on objective functions that are non-smooth, but close to smooth. Our results agree with those of (Devolder et al., 2014) and lead to the same kind of error accumulation. To see this, observe that we only use the definition of smoothness when bounding E_k^e in Theorem 3.4. Thus, the error from the inexact oracle would only appear as $E = E_k^e \leq A_k \delta$, leading to:

$$f(\mathbf{y}_k) - f(\mathbf{x}_*) \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + \frac{\sum_{i=1}^k A_i}{A_k} \cdot \delta,$$

This is exactly the same bound as in Theorem 4 of (Devolder et al., 2014), but we obtain it through a generic algorithm with a simple analysis.

C.2. Generalized Stochastic Models

It is possible to generalize the results from Lemma 3.7 to the noise model from (Lan, 2012; Ghadimi & Lan, 2012). In such a model, $\boldsymbol{\eta}_k = G(\mathbf{x}_k, \boldsymbol{\xi}_k) - \nabla f(\mathbf{x}_k)$, where $\{\boldsymbol{\xi}_i\}_{i=1}^k$'s are i.i.d. random vectors, $\mathbb{E}[\boldsymbol{\eta}_k] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{\eta}_k\|_*^2] \leq \sigma^2$. Let $\mathcal{F}_k$ denote the natural filtration up to (and including) iteration k. Then $\hat{\mathbf{v}}_k = \nabla \psi^*(\mathbf{z}_k + a_k \boldsymbol{\eta}_k)$ is measurable w.r.t. $\mathcal{F}_{k-1}$ (as $\{\mathbf{x}_i\}_{i=1}^k$ and $\{\boldsymbol{\xi}_i\}_{i=1}^{k-1}$ are measurable w.r.t. $\mathcal{F}_{k-1}$). It follows that:

$$\begin{aligned} \mathbb{E}[E_k^\eta | \mathcal{F}_{k-1}] &= a_k \mathbb{E}[\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \mathbf{v}_k \rangle | \mathcal{F}_{k-1}] = a_k \mathbb{E}[\langle \boldsymbol{\eta}_k, \mathbf{x}_* - \hat{\mathbf{v}}_k \rangle | \mathcal{F}_{k-1}] + a_k \mathbb{E}[\langle \boldsymbol{\eta}_k, \hat{\mathbf{v}}_k - \mathbf{v}_k \rangle | \mathcal{F}_{k-1}] \\ &\leq \frac{{a_k}^2}{\mu} \mathbb{E}[\|\boldsymbol{\eta}_k\|_*^2] \leq \frac{{a_k}^2 \sigma^2}{\mu}. \end{aligned} \tag{C.1}$$

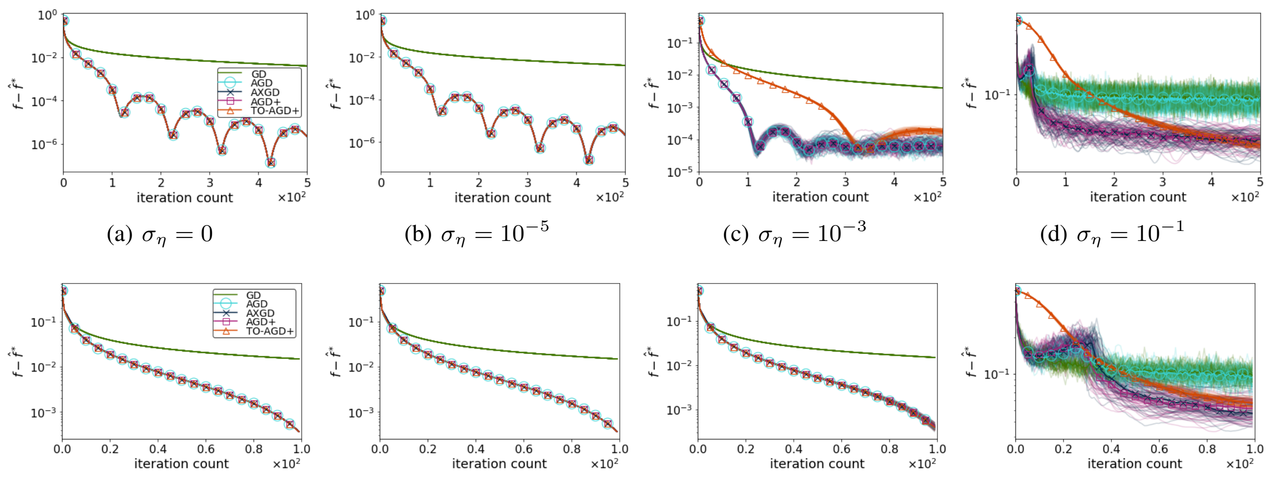
Define $\Gamma_k \stackrel{\text{def}}{=} A_k G_k - \sum_{i=1}^k \frac{{a_k}^2 \sigma^2}{\mu}$. Then, using the results from Section 3 and (C.1):

$$\mathbb{E}[\Gamma_k - \Gamma_{k-1} | \mathcal{F}_{k-1}] = \mathbb{E}\left[A_k G_k - A_{k-1} G_{k-1} - \frac{{a_k}^2 \sigma^2}{\mu} | \mathcal{F}_{k-1}\right] = \mathbb{E}\left[E_k^\eta - \frac{{a_k}^2 \sigma^2}{\mu} | \mathcal{F}_{k-1}\right] \leq 0,$$

i.e., Γ_k is a supermartingale. Hence, we can conclude that $\mathbb{E}[\Gamma_k] \leq \mathbb{E}[\Gamma_1]$, implying $\mathbb{E}[G_k] \leq \frac{A_1}{A_k}\mathbb{E}[G_1] + \frac{\sum_{i=2}^k {a_i}^2 \sigma^2}{\mu A_k}$, and we recover the same bound as in Lemma 3.7:

$$\mathbb{E}[f(\mathbf{y}_k)] - f(\mathbf{x}_*) \leq \frac{D_\psi(\mathbf{x}_*, \mathbf{x}_0)}{A_k} + \frac{\sum_{i=1}^k {a_i}^2 \sigma^2}{\mu A_k}.$$

[8] We actually obtain better constants than those in Theorem 2.2 of (d'Aspremont, 2008).



(a) $\sigma_\eta = 0$ (b) $\sigma_\eta = 10^{-5}$ (c) $\sigma_\eta = 10^{-3}$ (d) $\sigma_\eta = 10^{-1}$

(e) $\sigma_\eta = 0$ (f) $\sigma_\eta = 10^{-5}$ (g) $\sigma_\eta = 10^{-3}$ (h) $\sigma_\eta = 10^{-1}$

Figure 2: Median performance of gradient descent (GD) and accelerated algorithms (AGD, AXGD, AGD+ with restart and slow-down semi-heuristics and TO-AGD+) over 50 repeated runs on a hard instance for unconstrained smooth minimization for $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_\eta I)$ over $\mathbb{R}^n$ and for: (a)-(d) 500 iterations; (e)-(h) 100 iterations.

D. Additional Experiments

D.1. "Hard Instance" for Unconstrained Smooth Minimization

Unconstrained Minimization In Section 5, we compared various accelerated algorithms with gradient descent (GD) when restart and slow-down semi-heuristics are disabled and enabled. We also included a comparison with AGD+ when its parameters are set according to Corollary 3.9 (TO-AGD+). Since the step sizes of TO-AGD+ depend on the number of iterations, one may hope that TO-AGD+ could outperform other accelerated algorithms for a smaller number of iterations. However, this is not true – for a smaller number of iterations, in the best case TO-AGD+ matches the performance of other algorithms with restart and slow-down. When all algorithms have the same performance, they all essentially run their vanilla versions – without restart and slow down and for their standard accelerated step sizes. This is illustrated in Fig. 2.

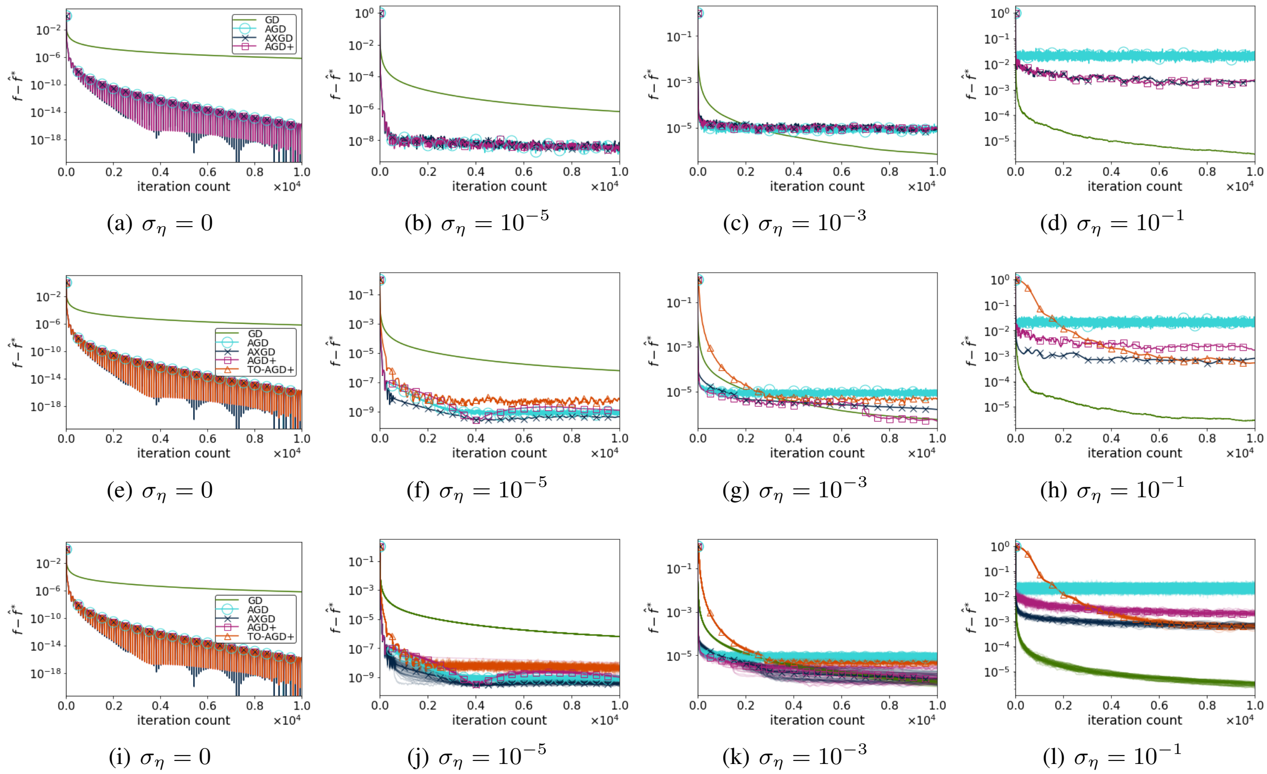
"Hard Instance" over Simplex The set of experiments in Figure 3 correspond to the minimization of the hard instance function for smooth optimization, constrained over the probability simplex. It should be compared to the unconstrained version in Figure 1. As predicted, we observe that the presence of constraints decreases the effect of error accumulation as the boundary of the feasible set limits the variance. Given the low variance due to the constraints, the effect of RESTART+SLOWDOWN is less evident for this batch of experiments.

D.2. Regression on Epileptic Seizure Dataset

For the second set of experiments, we used the Epileptic Seizure Recognition Dataset (Andrzejak et al., 2001) obtained from (Lichman, 2013). The dataset consists of brain activity EEG recordings for 100 patients at different time points and in five different states, of which only one indicates epileptic seizure. The dataset contains 11500 rows and 179 columns, of which the first 178 columns are features while the last column indicates whether the patient was in seizure. Before running the experiments, we standardize the data using a standard preprocessing function from the Python Sci-kit library.

Linear regression and LASSO. We performed linear regression on the considered dataset mainly to illustrate the performance of the algorithms in the bounded regime (for ℓ_1-constraints – LASSO), which we discuss here. The results for standard, unconstrained linear regression are similar to the results for logistic regression discussed below and are omitted.

Fig. 4(a)-4(d) shows the performance of GD and accelerated algorithms for ℓ_1-constrained linear regression (LASSO) on the Epileptic Seizure Recognition Dataset. Interestingly, the experiments suggest that there are cases when AGD performes better than AXGD and AGD+. In particular, in the noiseless (Fig. 4(a)) and the low-noise (Fig. 4(b)) settings, AGD performs



(a) $\sigma_\eta = 0$

(b) $\sigma_\eta = 10^{-5}$

(c) $\sigma_\eta = 10^{-3}$

(d) $\sigma_\eta = 10^{-1}$

(e) $\sigma_\eta = 0$

(f) $\sigma_\eta = 10^{-5}$

(g) $\sigma_\eta = 10^{-3}$

(h) $\sigma_\eta = 10^{-1}$

(i) $\sigma_\eta = 0$

(j) $\sigma_\eta = 10^{-5}$

(k) $\sigma_\eta = 10^{-3}$

(l) $\sigma_\eta = 10^{-1}$

Figure 3: Performance of gradient descent (GD) and accelerated algorithms (AGD, AXGD, AGD+) on a hard instance for unconstrained smooth minimization for $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_\eta I)$ and over probability simplex (a)-(d) without and (i)-(l) with RESTART+SLOWDOWN and RESTART+SLOWDOWN-2.

much better than worst-case and converges faster than AXGD and AGD+.

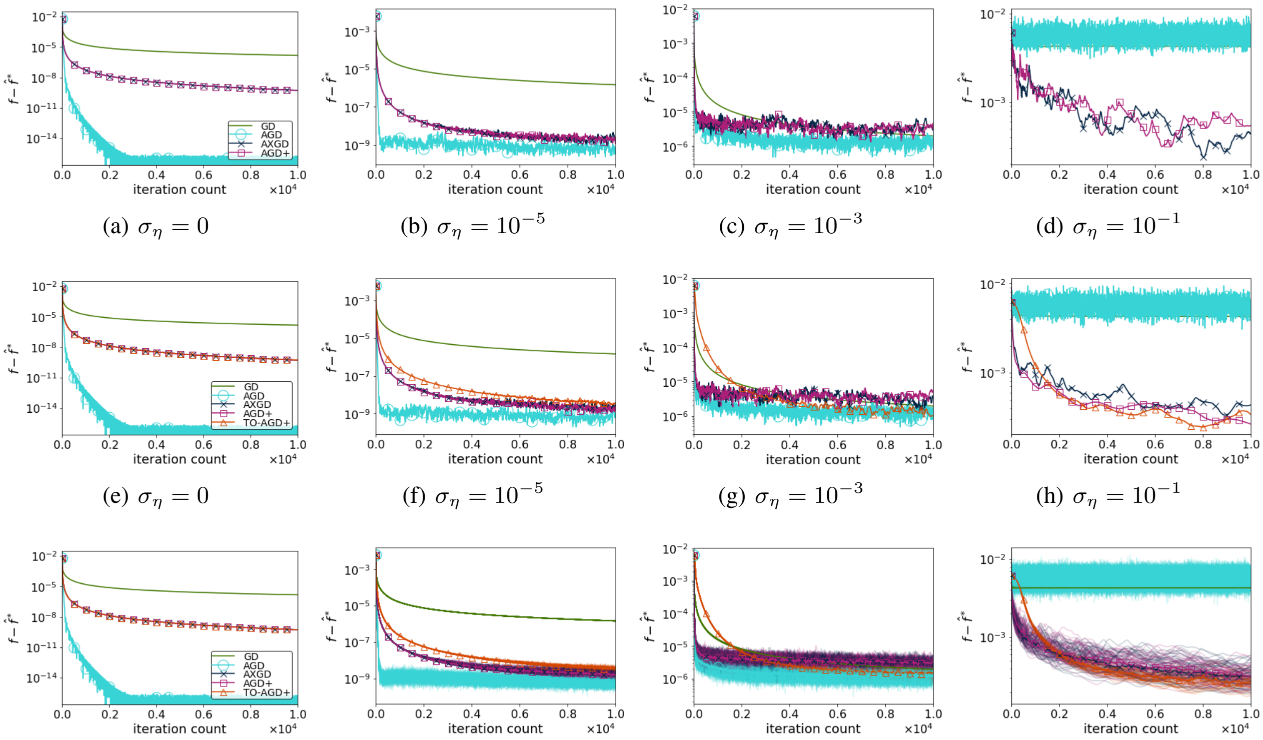
However, the faster convergence comes at the expense of lower stability to noise as the noise becomes higher. Specifically, as the noise is increased, AGD performs only marginally better and with higher variance than AXGD and AGD+ (Fig. 4(c)), and stabilizes to much higher mean and variance in the very high-noise setting (Fig. 1(d)).

Intuitively, “greedy” gradient steps that AGD takes may reduce the function value significantly and lead to faster convergence in the noiseless and low-noise settings, while making the convergence very sensitive to the noise from the last iteration, as the gradient steps only depend on the last seen (noisy) gradient. In contrast, AXGD and AGD+ are more stable to noise, since both of their per-iteration steps depend on the aggregate gradient (and thus, aggregate noise) information.

As expected from the analytical results from Section 3, restart and slow-down does not noticeably improve the mean error of the algorithms (Fig. 4(e)-4(h), 4(i)-4(l)). However, in agreement with the analysis, it can reduce the error variance in the high-noise-variance setting (Fig. 4(l)). We also note that TO-AGD+ is more stable over the repeated methods’ execution, at the expense of slower initial convergence.

Logistic regression. Finally, we evaluated the performance of the accelerated algorithms and GD for (unregularized) logistic regression on the Epileptic Seizure Recognition Dataset. The results are shown in Fig. 5.

Similar as in the case of unconstrained minimization from the beginning of this section, in the noiseless and low-noise settings (Fig. 5(a), 5(b)) all accelerated algorithms perform similarly and restart and slow-down does not lead to any noticeable improvements or degradation (Fig. 5(e), 5(i), 5(f), 5(j)). Once the noise is high enough (Fig. 5(c), 5(d)), all accelerated algorithms begin to accumulate noise, while restart and slow-down stabilize their performance to a low error mean and variance. Interestingly, in all the experiments, when RESTART+SLOWDOWN and RESTART+SLOWDOWN-2 are employed all accelerated algorithms perform at least as good as GD in terms of the error mean and variance.

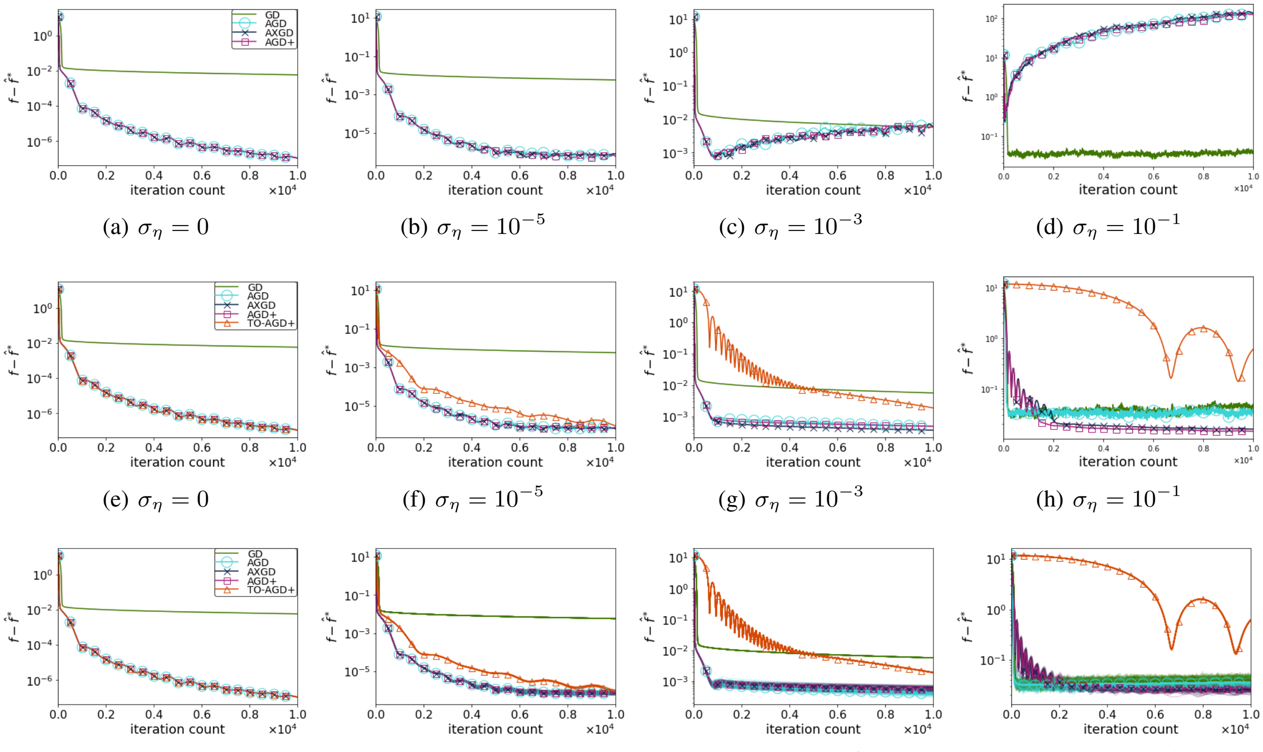


(a) $\sigma_\eta = 0$ (b) $\sigma_\eta = 10^{-5}$ (c) $\sigma_\eta = 10^{-3}$ (d) $\sigma_\eta = 10^{-1}$

(e) $\sigma_\eta = 0$ (f) $\sigma_\eta = 10^{-5}$ (g) $\sigma_\eta = 10^{-3}$ (h) $\sigma_\eta = 10^{-1}$

(i) $\sigma_\eta = 0$ (j) $\sigma_\eta = 10^{-5}$ (k) $\sigma_\eta = 10^{-3}$ (l) $\sigma_\eta = 10^{-1}$

Figure 4: Performance of GD and accelerated algorithms (AGD, AXGD, AGD+, TO-AGD+) for linear regression over ℓ_1-ball (LASSO) on Epileptic Seizure Recognition Dataset (Andrzejak et al., 2001) (a)-(d) without noise reduction; (i)-(l) sample run with TOAGD+, RESTART+SLOWDOWN and RESTART+SLOWDOWN-2; and (i)-(l) 50 repeated runs and the median with TOAGD+, RESTART+SLOWDOWN and RESTART+SLOWDOWN-2.



(a) $\sigma_\eta = 0$

(b) $\sigma_\eta = 10^{-5}$

(c) $\sigma_\eta = 10^{-3}$

(d) $\sigma_\eta = 10^{-1}$

(e) $\sigma_\eta = 0$

(f) $\sigma_\eta = 10^{-5}$

(g) $\sigma_\eta = 10^{-3}$

(h) $\sigma_\eta = 10^{-1}$

(i) $\sigma_\eta = 0$

(j) $\sigma_\eta = 10^{-5}$

(k) $\sigma_\eta = 10^{-3}$

(l) $\sigma_\eta = 10^{-1}$

Figure 5: Performance of gradient descent (GD) and accelerated algorithms (AGD, AXGD, AGD+, TO-AGD+) for logistic regression on Epileptic Seizure Recognition Dataset (Andrzejak et al., 2001) (a)-(d) without noise reduction; (e)-(h) sample run with noise reduction; and (i)-(l) 50 repeated runs and the median run with noise reduction.