Supplementary Material for On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings

A. Proof of Theorem 1

We begin from (14). By taking the squared norm on both sides, and recalling that the random vectors ζ_k have zero mean and are mutually independent, we have

$$\begin{aligned}
\mathbb{E}\left\|y_{k+1}-x^\star\right\|^2 &\le \mathbb{E}\left\|\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix}\right\|^2 \\
&= \mathbb{E}_{\zeta_k,\ldots,\zeta_1}\left[\left\|A^k\begin{bmatrix} x_1-x^\star \\ 0\end{bmatrix} - \alpha\sum_{j=1}^{k} A^{k-j}\begin{bmatrix}(1+\beta)I \\ I\end{bmatrix}\zeta_j\right\|^2\right] \\
&= \mathbb{E}_{\zeta_k}\left[\cdots\mathbb{E}_{\zeta_1}\left[\left\|A^k\begin{bmatrix} x_1-x^\star \\ 0\end{bmatrix} - \alpha\sum_{j=1}^{k} A^{k-j}\begin{bmatrix}(1+\beta)I \\ I\end{bmatrix}\zeta_j\right\|^2\right]\cdots\right] \\
&= \left\|A^k\begin{bmatrix} x_1-x^\star \\ 0\end{bmatrix}\right\|^2 + \alpha^2\sum_{j=1}^{k}\mathbb{E}_{\zeta_j}\left\|A^{k-j}\begin{bmatrix}(1+\beta)I \\ I\end{bmatrix}\zeta_j\right\|^2 \\
&\le \left\|A^k\right\|^2\left\|x_1-x^\star\right\|^2 + \alpha^2((1+\beta)^2+1)\sigma^2\sum_{j=1}^{k}\left\|A^{k-j}\right\|^2. \qquad (25)
\end{aligned}$$

Recall that the spectral radius of a square matrix $A\in\mathbb{R}^{2d\times 2d}$ is defined as $\max_{i=1,\ldots,2d}|\lambda_i(A)|$, where $\lambda_i(A)$ is the ith eigenvalue of A. The spectral radius satisfies (Horn & Johnson, 2013)

$$\rho(A)^k \le \left\|A^k\right\| \quad \text{for all } k,$$

and (Gelfand's theorem)

$$\lim_{k\to\infty}\left\|A^k\right\|^{1/k} = \rho(A).$$

Hence, for any $\epsilon>0$, there exists a K_ϵ such that $\left\|A^k\right\|^{1/k}\le(\rho(A)+\epsilon)$ for all $k\ge K_\epsilon$. Let

$$C_\epsilon = \max_{k<K_\epsilon}\max\left\{1, \frac{\left\|A^k\right\|}{(\rho(A)+\epsilon)^k}\right\}. \qquad (26)$$

Then $\left\|A^k\right\|\le C_\epsilon(\rho(A)+\epsilon)^k$ for all k. Moreover, if $\left\|A^k\right\|^{1/k}$ converges monotonically to $\rho(A)$, then $C_\epsilon\le\|A\|/\rho(A)$.

Now, recall that we have assumed $f(x)=\frac{1}{2}x^\top Hx - b^\top x + c$ where $H\in\mathbb{R}^{d\times d}$ is symmetric, and we have also assumed that f is L-smooth and μ-strongly convex. Thus all eigenvalues of H satisfy $\mu\le\lambda_i(H)\le L$.

Lemma 2. *For A as defined in* (15)*, we have $\rho(A)=\max\{\rho_\mu(\alpha,\beta),\rho_L(\alpha,\beta)\}$ where*

$$\rho_\lambda(\alpha,\beta) = \begin{cases} \frac{1}{2}\left|(1+\beta)(1-\alpha\lambda)\right| + \frac{1}{2}\sqrt{\Delta_\lambda} & \text{if } \Delta_\lambda\ge 0, \\ \sqrt{\beta(1-\alpha\lambda)} & \text{otherwise,}\end{cases}$$

and $\Delta_\lambda=(1+\beta)^2(1-\alpha\lambda)^2-4\beta(1-\alpha\lambda)$.

Proof. Since H is real and symmetric, it has a real eigenvalue decomposition $H = U\Lambda_H U^\top$, where $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and Λ_H is the diagonal matrix of eigenvalues of H. Observe that A can be viewed as a 2×2 block matrix with $d \times d$ blocks that all commute with each other, since each block is an affine matrix function of H. Thus, by Polyak (1964, Lemma 5), ξ is an eigenvalue of A if and only if there is an eigenvalue λ of H, such that ξ is an eigenvalue of the 2×2 matrix

$$B(\lambda) \coloneqq \begin{bmatrix} 1-\alpha(1+\beta)\lambda & \beta^2 \\ -\alpha\lambda & \beta \end{bmatrix}. \tag{27}$$

The characteristic polynomial of $B(\lambda)$ is

$$\xi^2 - (1+\beta)(1-\alpha\lambda)\xi + \beta(1-\alpha\lambda) = 0,$$

from which it follows that eigenvalues of $B(\lambda)$ are given by $\rho_\lambda(\alpha,\beta)$; see, e.g.., Lessard et al. (2016, Appendix A). Note that the characteristic polynomial of $B(\lambda)$ is the same as the characteristic polynomial of a different matrix appearing in Lessard et al. (2016), that arises from a different analysis of the AG method. Finally, as discussed in Lessard et al. (2016), for any fixed values of α and β, the function $\rho_\lambda(\alpha,\beta)$ is quasi-convex in λ, and hence the maximum over all eigenvalues of A is achieved at one of the extremes $\lambda = \mu$ or $\lambda = L$. □

To complete the proof of Theorem 1, use Lemma 2 with (25) to obtain that, for any $\epsilon > 0$, there is a positive constant C_ϵ such that

$$\begin{aligned}\mathbb{E}[\|y_{k+1} - x^\star\|^2] &\le C_\epsilon \left((\rho(A)+\epsilon)^{2k} \|x_0 - x^\star\|^2 + \alpha^2((1+\beta)^2+1)\sigma^2 \sum_{j=1}^{k} (\rho(A)+\epsilon)^{2(k-j)} \right) \\ &\le C_\epsilon \left((\rho(A)+\epsilon)^{2k} \|x_0 - x^\star\|^2 + \frac{\alpha^2((1+\beta)^2+1)}{1-(\rho(A)+\epsilon)^2}\sigma^2 \right).\end{aligned}$$

A.1. Estimating the constant C_ϵ

For the theoretical plots in the numerical experiments in Section 3.2 and in Appendix G below, we estimate the constant C_ϵ by taking $K_\epsilon \approx 2$ in (26). That is, for arbitrarily small ϵ and all $k \ge 2$, we approximate $\left\|A^k\right\|^{1/k}$ by $(\rho(A)+\epsilon)$. Therefore, the summation term in (25) is approximated as

$$\alpha^2((1+\beta)^2+1)\left(\frac{1}{1-\rho(\alpha,\beta)^2} + (\|A\|^2 - \rho(\alpha,\beta)^2)\right), \tag{28}$$

where $\|A\|$ denotes the largest singular value of A in (15), and $\rho(\alpha,\beta)$ is the largest eigenvalue of A. The first term in (28) corresponds to the geometric limit of the summation term in (25) after taking matrix norms and approximating the norms of matrix products by powers of the spectral radius for all products $k \ge 2$. The difference term in (28) is simply used to correct for the case $k = 1$. Setting $C_\epsilon \frac{\alpha^2((1+\beta)^2+1)}{1-\rho(\alpha,\beta)^2}$ equal to (28) and solving for C_ϵ gives us the approximate expression for C_ϵ used in the theoretical plots in Section 3.2.

B. Proofs of Corollary 1.1 and Theorem 2

Taking $\alpha = 1/L$ and $\beta = \frac{\sqrt{Q}-1}{\sqrt{Q}+1}$, we find that $\rho(\alpha,\beta) = \frac{\sqrt{Q}-1}{\sqrt{Q}}$. Since $f(x) = \frac{1}{2}x^T Hx - b^T x + c$ is an L-smooth μ-strongly convex quadratic, all eigenvalues of H are bounded between μ and L. Therefore, from Polyak (1964, Lemma 5), we have that $\left\|A^k\right\|_2 \le \max_{\lambda\in[\mu,L]} \left\|B(\lambda)^k\right\|_2 \le \max_{\lambda\in[\mu,L]} \sqrt{d}\left\|B(\lambda)^k\right\|_\infty$, where $B(\lambda)$ is as defined in (27). The eigenvalues of $B(\lambda)^k$ are maximized at $\lambda = \mu$ for $k > 1$, therefore, for large k, $\left\|B(\lambda)^k\right\|_\infty$ is maximized at $\lambda = \mu$.

Note that the Jordan form of $B(\mu)$ is given by VJV^{-1}, where

$$V = \begin{bmatrix} \frac{\sqrt{Q}(\sqrt{Q}-1)}{\sqrt{Q}+1} & Q \\ -1 & 0 \end{bmatrix} \qquad \text{and} \qquad J = \begin{bmatrix} \frac{\sqrt{Q}-1}{\sqrt{Q}} & 1 \\ 0 & \frac{\sqrt{Q}-1}{\sqrt{Q}} \end{bmatrix}.$$

Using the Jordan form, we determine that $B(\mu)^k$ is

$$B(\mu)^k = \begin{bmatrix} \left(1+\frac{k}{\sqrt{Q}+1}\right)\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^k & k\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^2\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k-1} \\ -\frac{k}{Q}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k-1} & \left(1-\frac{k}{\sqrt{Q}+1}\right)\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^k \end{bmatrix}.$$

Therefore, we have that

$$\left\|B(\mu)^k\right\|_\infty \leq \left(1+\frac{k}{\sqrt{Q}+1}\right)\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^k + k\max\left\{\frac{1}{Q}, \left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^2\right\}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k-1}. \tag{29}$$

Therefore for large k

$$\left\|A^k\right\|_2^2 \leq d\left\|B(\mu)^k\right\|_\infty^2 = \left(\frac{\sqrt{Q}-1}{\sqrt{Q}}+\epsilon_k\right)^{2k},$$

where $\epsilon_k \sim (\sqrt[k]{k}-1)$. Also observe that

$$\begin{aligned}\frac{\alpha^2((1+\beta)^2+1)}{1-\rho(\alpha,\beta)^2} &= \frac{1}{L^2}\frac{\left(\frac{2\sqrt{Q}}{\sqrt{Q}+1}\right)^2+1}{1-\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^2} \\ &= \frac{1}{L^2}\frac{5Q^2+2Q^{3/2}+Q}{(\sqrt{Q}+1)^2(2\sqrt{Q}-1)}.\end{aligned}$$

Since f is L-smooth,

$$f(y_{k+1})-f^\star \leq \frac{L}{2}\left\|y_{k+1}-x^\star\right\|^2.$$

Thus, by Theorem 1 we have

$$\mathbb{E}[f(y_{k+1})]-f^\star \leq \frac{L}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}+\epsilon_k\right)^{2k}\left\|x_0-x^\star\right\|^2 + C_\epsilon\frac{5Q^2+2Q^{3/2}+Q}{2L(2\sqrt{Q}-1)(\sqrt{Q}+1)^2}\sigma^2,$$

which completes the proof of Corollary 1.1.

To prove Theorem 2, first observe that when $\beta=0$, the recursion simplifies significantly. Specifically, then $y_{k+1}=x_k$, $v_k=-\alpha g_k$, and we have (using similar notation as in the proof of Theorem 1)

$$\begin{aligned} r_{k+1} &= (I-\alpha H_k)r_k - \alpha\zeta_k \\ &= \prod_{j=1}^{k}(I-\alpha H_j)r_1 - \alpha\zeta_k - \alpha\sum_{j=1}^{k-1}\prod_{l=j+1}^{k}(I-\alpha H_l)\zeta_j, \end{aligned}$$

where

$$H_j = \int_0^1 \nabla f^2(x^\star - t\, r_j)\mathrm{d}t.$$

Of course, since f is L-smooth and μ-strongly convex, all eigenvalues of H_j lie in the interval $[\mu, L]$ for all $j \geq 0$.

Now, taking the squared norm on both sides, and recalling that the random vectors ζ_k have zero mean and are mutually

independent, we have

$$\begin{aligned}
\mathbb{E}\left\|y_{k+1}-x^\star\right\|^2 &= \mathbb{E}\left\|r_{k+1}\right\|^2 \\
&= \mathbb{E}_{\zeta_k,\ldots,\zeta_1}\left\|\prod_{j=1}^{k}(I-\alpha H_j)r_1-\alpha\zeta_k-\alpha\sum_{j=1}^{k-1}\prod_{l=j+1}^{k}(I-\alpha H_l)\zeta_j\right\|^2 \\
&= \mathbb{E}_{\zeta_k}\left[\cdots\mathbb{E}_{\zeta_1}\left[\left\|\prod_{j=1}^{k}(I-\alpha H_j)r_1-\alpha\zeta_k-\alpha\sum_{j=1}^{k-1}\prod_{l=j+1}^{k}(I-\alpha H_l)\zeta_j\right\|^2\right]\cdots\right] \\
&= \left(\prod_{j=1}^{k}\left\|I-\alpha H_j\right\|^2\right)\left\|x_1-x^\star\right\|^2+\alpha^2\mathbb{E}_{\zeta_k}\left\|\zeta_k\right\|^2+\alpha^2\sum_{j=1}^{k-1}\mathbb{E}_{\zeta_j}\left\|\left(\prod_{l=j+1}^{k}I-\alpha H_l\right)\zeta_j\right\|^2 \\
&\le \left(\prod_{j=1}^{k}\left\|I-\alpha H_j\right\|^2\right)\left\|x_1-x^\star\right\|^2+\alpha^2\sigma^2+\alpha^2\sigma^2\sum_{j=1}^{k-1}\left(\prod_{l=j+1}^{k}\left\|I-\alpha H_l\right\|^2\right).
\end{aligned}$$

Now, since $I-\alpha H_j$ is symmetric, we have $\left\|(I-\alpha H_j)\right\|^2=\rho(I-\alpha H_j)^2$, where $\rho(I-\alpha H_j)$ denotes the spectral radius of $I-\alpha H_j$ (the largest magnitude of an eigenvalue of $I-\alpha H_j$). For $\alpha=\frac{2}{\mu+L}$, and since the eigenvalues of H_j lie in the interval $[\mu,L]$, it is straightforward to show that $\rho(I-\alpha H_j)=\frac{Q-1}{Q+1}$.

Therefore we have

$$\begin{aligned}
\mathbb{E}\left[\left\|y_{k+1}-x^\star\right\|^2\right] &\le \left(\frac{Q-1}{Q+1}\right)^{2k}\left\|x_0-x^\star\right\|^2+\alpha^2\sigma^2\sum_{j=1}^{k}\left(\frac{Q-1}{Q+1}\right)^{2(k-j)} \\
&\le \left(\frac{Q-1}{Q+1}\right)^{2k}\left\|x_0-x^\star\right\|^2+\frac{\alpha^2\sigma^2}{1-(\frac{Q-1}{Q+1})^2} \\
&= \left(\frac{Q-1}{Q+1}\right)^{2k}\left\|x_0-x^\star\right\|^2+\frac{Q}{2L}\sigma^2,
\end{aligned}$$

which completes the proof of Theorem 2.

C. Permutation Matrix Construction

For a vector $x\in\mathbb{R}^d$, let $\mathrm{diag}(x)$ denote a $d\times d$ diagonal matrix with its ith diagonal entry equal to x_i. Let $a,b,c,d\in\mathbb{R}^d$ and suppose $M\in\mathbb{R}^{2d\times 2d}$ is the matrix

$$M=\begin{bmatrix}\mathrm{diag}(a) & \mathrm{diag}(b)\\ \mathrm{diag}(c) & \mathrm{diag}(d)\end{bmatrix}.$$

Let $P\in\{0,1\}^{2d\times 2d}$ be the permutation matrix with entries $P_{i,j}$ for $i,j=1,\ldots,2d$ given by

$$P_{i,j}=\begin{cases}1 & \text{if } i \text{ is odd and } j=(i-1)/2+1\\ 1 & \text{if } i \text{ is even and } j=d+\lfloor\frac{i-1}{2}\rfloor+1\\ 0 & \text{otherwise.}\end{cases}$$

Then one can verify that

$$PMP^\top=\begin{bmatrix}T_1 & 0 & \cdots & 0\\ 0 & T_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & T_d\end{bmatrix}$$

where, for $j=1,\ldots,d$, T_j is the 2×2 matrix

$$T_j=\begin{bmatrix}a_j & b_j\\ c_j & d_j\end{bmatrix}.$$

D. Proof of Lemma 1

Recall that $\alpha = 1/L$ and $\beta = \frac{\sqrt{Q}-1}{\sqrt{Q}+1}$. For matrices of the form

$$T_k = B(L)B(\mu)^{k_1}B(L)B(\mu)^{k_2}\cdots B(L)B(\mu)^{k_s}B(L),$$

where

$$B(\lambda) = \begin{bmatrix} 1-\alpha(1+\beta)\lambda & \beta^2 \\ -\alpha\lambda & \beta \end{bmatrix},$$

we would like to show that the spectral radius $\rho(T_k)$ is equal to

$$\rho(T_k) = \left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^k \times k_1k_2\cdots k_s.$$

To see this, first note that the Jordan form of $B(\mu)$ is given by VJV^{-1}, where

$$V = \begin{bmatrix} \frac{\sqrt{Q}(\sqrt{Q}-1)}{\sqrt{Q}+1} & Q \\ -1 & 0 \end{bmatrix} \qquad \text{and} \qquad J = \begin{bmatrix} \frac{\sqrt{Q}-1}{\sqrt{Q}} & 1 \\ 0 & \frac{\sqrt{Q}-1}{\sqrt{Q}} \end{bmatrix}.$$

Using the Jordan form, we determine that $B(\mu)^{k_\ell}$ is

$$B(\mu)^{k_\ell} = \begin{bmatrix} \left(1+\frac{k_\ell}{\sqrt{Q}+1}\right)\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_\ell} & k_\ell\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^2\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_\ell-1} \\ -\frac{k_\ell}{Q}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_\ell-1} & \left(1-\frac{k_\ell}{\sqrt{Q}+1}\right)\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_\ell} \end{bmatrix}.$$

Through direct matrix multiplication

$$B(L)B(\mu)^{k_\ell}B(L) = -\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_\ell+1} k_\ell B(L).$$

Therefore,

$$T_j = (-1)^{s-1}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k-1-k_s} k_1k_2\cdots k_{s-1}B(L)B(\mu)^{k_s}.$$

Finally, the spectral-radius of $B(L)B(\mu)^{k_s}$ is

$$\rho\left(B(L)B(\mu)^{k_s}\right) = \left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k_s+1} k_s,$$

and hence

$$\rho\left(T_k\right) = \left(\frac{\sqrt{Q}-1}{\sqrt{Q}}\right)^{k} k_1k_2\cdots k_s.$$

E. Proof of Theorem 4

Since the functions f_i are assumed to be twice continuously differentiable, by (10) we can express the mini-batch gradients as

$$g_k = \tilde{H}_k r_k + z_k, \tag{30}$$

where

$$\tilde{H}_k = \sum_{i=1}^{n} v_{k,i}\int_0^1 \nabla^2 f_i(x^\star + tr_k)\mathrm{d}t$$

and

$$z_k = \sum_{i=1}^{n} v_{k,i} \nabla f_i(x^\star).$$

By convexity of norms,

$$\|z_k\| \leq \sum_{i=1}^{n} v_{k,i} \|\nabla f_k(x^\star)\|.$$

Hence, taking expectations gives

$$\begin{aligned}\mathbb{E}_k[\|z_k\|] &\leq \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^\star)\| \\ &= \sigma.\end{aligned}$$

Using (30) in (9) and unrolling, we obtain

$$\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} = A_k \cdots A_1 \begin{bmatrix} r_1 \\ v_0 \end{bmatrix} - \alpha \begin{bmatrix} (1+\beta)I \\ I \end{bmatrix} z_k - \alpha \sum_{j=1}^{k-1} (A_k \cdots A_{j+1}) \begin{bmatrix} (1+\beta)I \\ I \end{bmatrix} z_j, \tag{31}$$

where

$$A_k = \begin{bmatrix} I - \alpha(1+\beta)\tilde{H}_k & \beta^2 I \\ -\alpha\tilde{H}_k & \beta I \end{bmatrix}.$$

By submultiplicativity of matrix norms, $\|A_k \cdots A_{j+1}\| \leq \prod_{l=j+1}^{k} \|A_l\|$. Thus we turn our attention to bounding the spectral norm of A_k.

Lemma 3.

$$\|A_k\| \leq \max_{\lambda \in [\mu, L]} \|B(\lambda)\| = R(\alpha, \beta).$$

Proof. For all $k \geq 0$, every eigenvalue of $\tilde{H}_k$ lies in the interval $[\mu, L]$, based on the assumption that each function f_i is L-smooth and μ-strongly convex. It follows from Polyak (1964, Lemma 5) that there exists an eigenvalue λ of $\tilde{H}_k$ such that $\|A_k\|$ is equal to the spectral norm of

$$B(\lambda) = \begin{bmatrix} 1 - \alpha(1+\beta)\lambda & \beta^2 \\ -\alpha\lambda & \beta \end{bmatrix}.$$

We next compute $\|B(\lambda)\|$, which is equal to the square root of the largest eigenvalue of

$$B(\lambda)^\top B(\lambda) = \begin{bmatrix} (1-\alpha(1+\beta)\lambda)^2 + \alpha^2\lambda^2 & \beta^2(1-\alpha(1+\beta)\lambda) - \alpha\beta\lambda \\ \beta^2(1-\alpha(1+\beta)\lambda) - \alpha\beta\lambda & \beta^2(\beta^2+1) \end{bmatrix}.$$

The characteristic polynomial of $B(\lambda)^\top B(\lambda)$ is

$$\xi^2 - C_\lambda(\alpha, \beta)\xi + \beta^2(1-\alpha\lambda)^2 = 0,$$

where

$$C_\lambda(\alpha, \beta) = (1-\alpha(1+\beta)\lambda)^2 + \alpha^2\lambda^2 + \beta^2(\beta^2+1).$$

The largest root of the characteristic polynomial is equal to

$$R_\lambda(\alpha, \beta)^2 = \frac{1}{2}\left(C_\lambda(\alpha, \beta) + \sqrt{C_\lambda(\alpha, \beta)^2 - 4\beta^2(1-\alpha\lambda)^2}\right)$$

which is equal to $\|B(\lambda)\|^2$. Therefore

$$\|A_k\| \leq \max_{\lambda \in [\mu, L]} R_\lambda(\alpha, \beta).$$

□

Assume that α and β have been chosen so that $R(\alpha,\beta) < 1$. Then for all k and $j+1$, $\|A_k \cdots A_{j+1}\| \leq \prod_{l=j+1}^{k} \|A_l\| \leq R(\alpha,\beta)^{k-j}$.

Taking the norm on both sides of (31) and using the triangle inequality, we have

$$\left\| \begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} \right\| \leq R(\alpha,\beta)^k \left\| \begin{bmatrix} r_1 \\ v_0 \end{bmatrix} \right\| + \alpha\sqrt{(1+\beta)^2+1} \sum_{j=1}^{k} R(\alpha,\beta)^{k-j} \|z_k\|. \tag{32}$$

Taking the expectation gives

$$\mathbb{E}_k \|y_{k+1} - x^\star\| \leq \mathbb{E}_k \left\| \begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} \right\| \tag{33}$$

$$\leq R(\alpha,\beta)^k \|x_0 - x^\star\| + \frac{\alpha\sqrt{(1+\beta)^2+1}}{1-R(\alpha,\beta)^2}\sigma. \tag{34}$$

F. Proof of Corollaries 4.2 and 4.1

When $\beta = 0$, we have $y_{k+1} = x_k$ and $v_k = -\alpha g_k$ for all k. In this case we have

$$r_{k+1} = r_k - \alpha g_k.$$

Since the objectives f_i are twice continuously differentiable, the mini-batch gradients can again be written as (using the same notation as in the proof of Theorem 4)

$$g_k = \tilde{H}_k r_k + z_k.$$

Thus, with $A_k = I - \alpha\tilde{H}_k$, we have

$$\begin{aligned} r_{k+1} &= A_k r_k - \alpha z_k \\ &= A_k \cdots A_1 r_1 - \alpha z_k - \alpha \sum_{j=1}^{k-1} (A_k \cdots A_{j+1}) z_k. \end{aligned}$$

Since $\tilde{H}_k$ is symmetric, it follows that A_k is also symmetric, and so $\|A_k\|$ is equal to the largest magnitude of any eigenvalue of A_k. Recall that all eigenvalues of $\tilde{H}_k$ lie in the interval $[\mu, L]$. Therefore, $\|A_k\| \leq \max_{\lambda\in[\mu,L]} |1-\alpha\lambda| = \max\{|1-\alpha\mu|, |1-\alpha L|\}$. Choosing $\alpha < \frac{2}{L}$ and taking the norm and expectation thus yields that

$$\begin{aligned} \mathbb{E}_k \|x_k - x^\star\| &= \mathbb{E}_k \|r_{k+1}\| \\ &\leq \left|1-\alpha\tilde{\lambda}\right|^k \|x_0 - x^\star\| + \frac{\alpha}{1-\left|1-\alpha\tilde{\lambda}\right|}\sigma, \end{aligned} \tag{35}$$

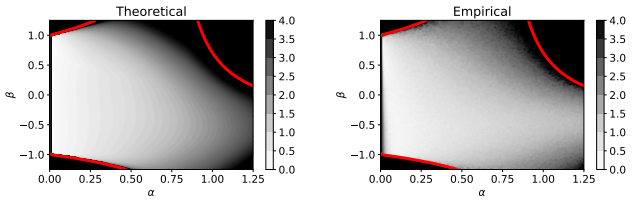
where $\tilde{\lambda} \coloneqq \operatorname{argmax}_{\lambda\in\{\mu,L\}} |1-\alpha\lambda|$. When $\alpha = \frac{2}{\mu+L}$, we have that $\max_{\lambda\in[\mu,L]} |1-\alpha\lambda| = \frac{Q-1}{Q+1}$, and equation (35) simplifies as

$$\begin{aligned} \mathbb{E}_k \|x_k - x^\star\| &= \mathbb{E}_k \|r_{k+1}\| \\ &\leq \left(\frac{Q-1}{Q+1}\right)^k \|x_0 - x^\star\| + \frac{1}{\mu}\sigma. \end{aligned}$$

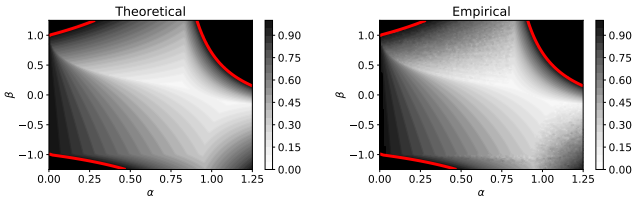
G. Additional Experiments

G.1. Least Squares

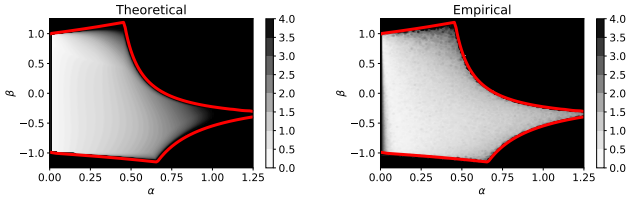
To provide additional experiments illustrating the relationship between empirical observations and the theory developed in Section 3 for the stochastic approximation setting, we conduct additional experiments on randomly-generated least-squares problems. We generate the least-squares problem using the approach described in (Lenard & Minkoff, 1984). Visualizations are shown in Figure G.1.



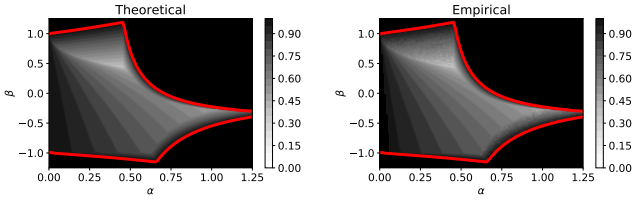
(a) ($Q = 2$): Coefficient multiplying σ^2



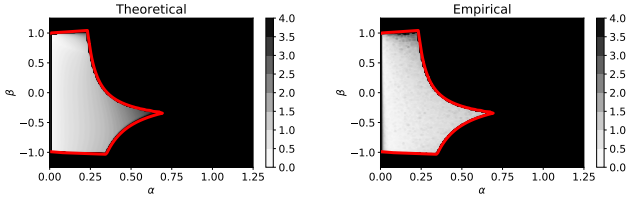
(b) ($Q = 2$): Convergence rate $\rho(\alpha, \beta)$



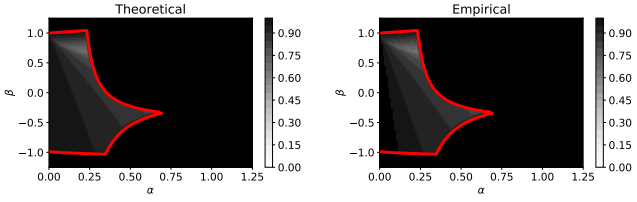
(c) ($Q = 8$): Coefficient multiplying σ^2



(d) ($Q = 8$): Convergence rate $\rho(\alpha, \beta)$



(e) ($Q = 32$): Coefficient multiplying σ^2



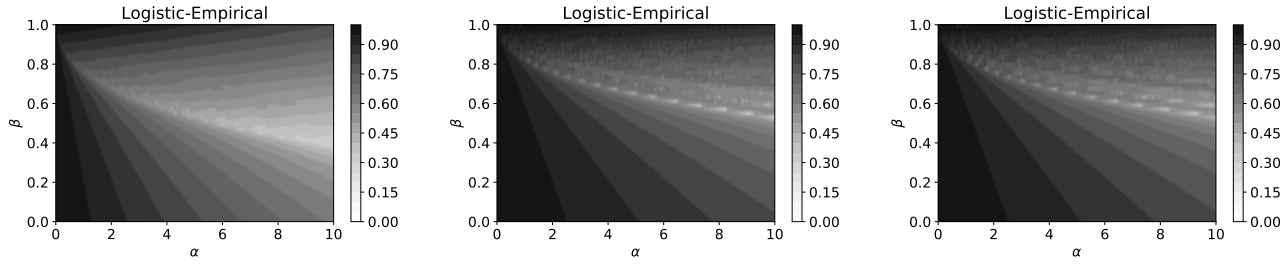
(f) ($Q = 32$) Convergence rate $\rho(\alpha, \beta)$

Figure G.1. Visualizing the accuracy with which the theory predicts the coefficient of the variance term and the convergence rate for different choices of constant step-size and momentum parameters, and various objective condition numbers Q. Plots labeled "Theoretical" depict theoretical results from Theorem 1. Plots labeled "Empirical" depict empirical results when using the ASG method to solve a least-squares regression problem with additive Gaussian noise; each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. In all figures, the area enclosed by the red contour depicts the theoretical stability region from Theorem 1 for which $\rho(\alpha, \beta) < 1$. Fig. G.1a/G.1c/G.1e: Pixel intensities correspond to the coefficient of the variance term in Theorem 1 ($\lim_{k\to\infty} \frac{1}{\sigma}\mathbb{E}\left\|y_k - x^\star\right\|_\infty$), which provides a good characterization of the magnitude of the neighbourhood of convergence, even without explicit knowledge of the constant C_ϵ. Fig. G.1b/G.1d/G.1f: Pixel intensities correspond to the theoretical convergence rates in Theorem 1, which provides a good characterization of the empirical convergence rates. Moreover, the theoretical conditions for convergence in Theorem 1 depicted by the red-contour are tight.

We run the ASG method on least-squares regression problems with various condition numbers Q. The objectives f correspond to randomly generated least squares problems, consisting of 500 data samples with 10 features each. Stochastic gradients are sampled by adding zero-mean Gaussian noise, with standard-deviation $\sigma = 0.25$, to the true gradient. The left plots in each sub-figure depict theoretical predictions from Theorem 1, while the right plots in each sub-figure depict empirical results. Each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. In all figures, the area enclosed by the red contour depicts the theoretical stability region from Theorem 1 for which $\rho(\alpha, \beta) < 1$.

Figures G.1a/G.1c/G.1e showcase the coefficient multiplying the variance term, which is taken to be $\frac{\alpha^2((1+\beta)^2+1)}{1-\rho(\alpha,\beta)^2}$ in theory. Brighter regions correspond to smaller coefficients, while darker regions correspond to larger coefficients. All sets of figures (theoretical and empirical) use the same color scale. We can see that the coefficient of the variance term in Theorem 1 provides a good characterization of the magnitude of the neighbourhood of convergence. The constant C_ϵ is approximated as $1 + (1 - \rho(\alpha, \beta)^2)(\varrho(\alpha, \beta)^2 - \rho(\alpha, \beta)^2)$, where $\varrho(\alpha, \beta)$ is defined as the largest singular value of A in (15), and $\rho(\alpha, \beta)$ is the largest eigenvalue of A.

Figures. G.1b/G.1d/G.1f showcase the linear convergence rate in theory and in practice. Brighter regions correspond to



(a) $Q = 30$ (b) $Q = 45$ (c) $Q = 60$

Figure G.2. Visualizing the convergence rate for the ASG method (momentum $\beta > 0$) and the SGD method (momentum $\beta = 0$), for various randomly generated ℓ_2 regularized multinomial logistic-regression problems. Multi-class classification problems consist of 5 classes and 100 data samples with 10 features each, only 5 of which are are discriminative. We create one data-cluster per class, and vary the cluster separation and regularization parameter to vary the condition number Q. For reporting purposes, we estimate the condition number Q during training by evaluating the eigenvalues of the Hessian at each iteration. The smoothness constant L is taken to be the maximum eigenvalue seen during training, and the modulus of strong-convexity μ is taken to be the minimum eigenvalue seen during training. The faster convergence rates (brighter regions) correspond to $\beta > 0$, indicating that the ASG method provides acceleration over SGD in this stochastic approximation setting. Moreover, for a given step-size, the contrast between the brighter regions ($\beta > 0$) and darker regions ($\beta = 0$) increases as the condition number grows, supporting theoretical findings that the convergence rate of the ASG method exhibits a better dependence on the condition number than SGD.

faster rates, and darker regions correspond to slower rates. Again, all figures (theoretical and empirical) use the same color scale. We can see that the theoretical linear convergence rates in Theorem 1 provide a good characterization of the empirical convergence rates. Moreover, the theoretical conditions for convergence in Theorem 1 depicted by the red-contour appear to be tight.

G.2. Multinomial Logistic Regression

Next we conduct experiments on ℓ_2 regularized multinomial logistic regression problems with additive Gaussian noise, to examine whether the ASG method still achieves acceleration over SGD for these problems in the stochastic approximation setting, as is predicted by the theory in Section 3. These problems are smooth and strongly-convex, but non-quadratic. Tight estimates of the smoothness constant L and the modulus of strong-convexity μ cannot be computed definitively since the eigenvalues of the Hessian vary throughout the parameter space.

We randomly generate multi-class classification problems consisting of 5 classes and 100 data samples with 10 features each, only five of which are discriminative. We create one data cluster per class, and vary the cluster separation and regularization parameter to vary the condition number Q. For reporting purposes, we estimate the condition number Q during training by evaluating the eigenvalues of the Hessian at each iteration. The smoothness constant L is taken to be the maximum eigenvalue seen during training, and the modulus of strong-convexity μ is taken to be the minimum eigenvalue seen during training. We use the `make_classification()` function in scikit-learn (Pedregosa et al., 2011) to generate random classification problem instances.

Visualizations are provided in Figure G.2. Each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. Pixel intensities denote the linear convergence rates observed in practice. Brighter regions correspond to faster rates, and darker regions correspond to slower rates.

The parameter setting β equals 0 corresponds to SGD, and the parameter setting $\beta > 0$ corresponds to the ASG method. The faster convergence rates (brighter regions) correspond to $\beta > 0$, indicating that the ASG method provides acceleration over SGD in this stochastic approximation setting. Moreover, for a given step-size, the contrast between the brighter regions ($\beta > 0$) and darker regions ($\beta = 0$) increases as the condition number grows, supporting theoretical findings that the convergence rate of the ASG method exhibits a better dependence on the condition number than SGD.