APPENDIX

In this section, we consider a simple quadratic function

$$f(x) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i x_i^2 = \frac{1}{2}x^T \Lambda x \tag{1}$$

$$\nabla f(x) = \Lambda x \tag{2}$$

$$\nabla^2 f(x) = \Lambda \tag{3}$$

The results can be generalized to asymptotic analysis of other convex functions based on the following proposition

Proposition 0.1. *Let (f_k) be the sequence defined by the recursion $f_{k+1} = af_k + bf_k^2$, for $k = 1, 2, \ldots$. If $a < 1$ and $f_1 < \frac{1-a}{b}$, then (f_k) converges to 0 at asymptotic rate a.*

Proof. Since (f_k) is strictly decreasing, it is easy to show that , with $a < 1$,

- If $f_1 > \frac{1-a}{b}$, (f_k) diverges.
- If $f_1 = \frac{1-a}{b}$, $(f_k) = \frac{1-a}{b}$.
- If $f_1 < \frac{1-a}{b}$, (f_k) converges to 0.

Consider the case when (f_k) converges to 0. There must exist k_0 such that $f_k < \frac{a(1-a)}{b}$, for all $k \geq k_0$. Suppose that $f_1 = \alpha\frac{1-a}{b}$, where $0 < \alpha < 1$. Let us define a sequence (h_k) as $h_k = \frac{1}{f_1 a^{k-1}} f_k$. Then for $k \geq k_0$

$$h_k = \frac{1}{f_1 a^{k-1}} f_k < \frac{1}{f_1 a^{k-1}} \frac{a(1-a)}{b} = \frac{1}{\alpha a^{k-2}} \tag{4}$$

The recursion for (h_k) is given by

$$\begin{cases} h_1 = 1, \\ h_{k+1} = h_k + \alpha(1-a)a^{k-2}h_k^2. \end{cases}$$

Notice that (h_k) is also strictly increasing, and the following inequalities hold

$$\begin{aligned}
\Rightarrow \quad & \alpha(1-a)a^{k-2} = \frac{h_{k+1} - h_k}{h_k^2} && > \frac{h_{k+1} - h_k}{h_{k+1}h_k} = \frac{1}{h_k} - \frac{1}{h_{k+1}} \\
\Rightarrow \quad & \sum_{i=k_0}^{k-1} \alpha(1-a)a^{i-2} && > \sum_{i=k_0}^{k-1}\left(\frac{1}{h_i} - \frac{1}{h_{i+1}}\right) \\
\Rightarrow \quad & \alpha(1-a)a^{k_0-2}\sum_{j=0}^{k-1-k_0} a^j && > \frac{1}{h_{k_0}} - \frac{1}{h_k} \\
\Rightarrow \quad & \alpha(1-a)a^{k_0-2}\frac{1-a^{k-k_0}}{1-a} && > \frac{1}{h_{k_0}} - \frac{1}{h_k} \\
\Rightarrow \quad & \frac{1}{h_k} && > \frac{1}{h_{k_0}} - \alpha a^{k_0-2}(1-a^{k-k_0}) \\
\Rightarrow \quad & h_k && < \frac{1}{\left(\frac{1}{h_{k_0}} - \alpha a^{k_0-2}\right) + \alpha a^{k-2}}
\end{aligned}$$

From (4), the sequence defined by the RHS must converge to a constant $\frac{1}{\frac{1}{h_{k_0}} - \alpha a^{k_0-2}}$. Consequently, (h_k) is upper-bounded by this sequence and also converges. Finally, we obtain $\lim h_k = \lim \frac{a}{f_1}\frac{f_k}{a^k} < \infty$, yielding the asymptotic convergence rate of (f_k) to 0 is a. $\square$

1 Proof of convergence rate for fixed step size gradient descent

From the update $x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}) = x^{(k)} - \alpha\Lambda x^{(k)}$, we have

$$\begin{aligned}
f(x^{(k+1)}) &= \frac{1}{2}(x^{(k)} - \alpha\Lambda x^{(k)})^T \Lambda (x^{(k)} - \alpha\Lambda x^{(k)}) = \frac{1}{2}(x^{(k)})^T (I - \alpha\Lambda)\Lambda(I - \alpha\Lambda)x^{(k)} \\
&= \frac{1}{2}(\Lambda^{1/2}x^{(k)})^T (I - \alpha\Lambda)^2 (\Lambda^{1/2}x^{(k)}) \\
&\leq \frac{1}{2}\left\|I - \alpha\Lambda\right\|_2^2 . \left\|\Lambda^{1/2}x^{(k)}\right\|_2^2 = \max_i (1 - \alpha\lambda_i)^2 . f(x^{(k)})
\end{aligned}$$

By setting $\alpha = \frac{2}{\lambda_1 + \lambda_d}$, we obtain

$$f(x^{(k+1)}) \leq \left(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\right)^2 f(x^{(k)}).$$

2 Proof of convergence rate for fixed step size momentum method

From the update $x^{(k+1)} = x^{(k)} - \alpha\nabla f(x^{(k)}) + \beta(x^{(k)} - x^{(k-1)})$, we have

$$\begin{aligned}
y^{(k+1)} &= \begin{bmatrix} x^{(k+1)} \\ x^{(k)} \end{bmatrix} = \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ x^{(k-1)} \end{bmatrix} = My^{(k)} \\
f_{k+1} &= f(x^{(k+1)}) + f(x^{(k)}) = \frac{1}{2} {y^{(k+1)}}^T \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda \end{bmatrix} y^{(k+1)} = \frac{1}{2} {y^{(k)}}^T M^T \hat{\Lambda} M y^{(k)} = \dots \\
&= \frac{1}{2} {y^{(1)}}^T {M^k}^T \hat{\Lambda} M^k y^{(1)} = \frac{1}{2}(\hat{\Lambda}^{1/2}y^{(1)})^T \left(\hat{\Lambda}^{-1/2} {M^k}^T \hat{\Lambda} M^k \hat{\Lambda}^{-1/2}\right)(\hat{\Lambda}^{1/2}y^{(1)}) \\
&= \frac{1}{2}(\hat{\Lambda}^{1/2}y^{(1)})^T \left((\hat{\Lambda}^{1/2}M^k\hat{\Lambda}^{-1/2})^T(\hat{\Lambda}^{1/2}M^k\hat{\Lambda}^{-1/2})\right)(\hat{\Lambda}^{1/2}y^{(1)}) \\
&\leq \frac{1}{2}\left\|\hat{\Lambda}^{1/2}M^k\hat{\Lambda}^{-1/2}\right\|_2^2 \left\|\hat{\Lambda}^{1/2}y^{(1)}\right\|_2^2 = \left\|\hat{\Lambda}^{1/2}M^k\hat{\Lambda}^{-1/2}\right\|_2^2 f_1 \\
&\leq (\left\|\hat{\Lambda}^{1/2}\right\|_2 \left\|M^k\right\|_2 \left\|\hat{\Lambda}^{-1/2}\right\|_2) f_1 = \left\|M^k\right\|_2^2 \frac{\lambda_1^2}{\lambda_d^2} f_1
\end{aligned}$$

Since $\lim_{k\to\infty} \frac{\|M^k\|_2^2}{\rho(M)^k} = 1$ [1], the spectral radius $\rho(M) = \max_j\{|\lambda_j(M)|\}$ determines the convergence rate of the series (f_k). Recall that $M = \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix}$. We define the permutation π such that

$$\pi(j) = \begin{cases} 2j - 1 \text{ if } j \leq d, \\ 2j - 2d \text{ otherwise.} \end{cases}$$

Then

$$M \sim P_\pi M P_\pi^T = \begin{bmatrix} M_1 & 0 & \dots & 0 \\ 0 & M_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & M_d \end{bmatrix}$$

is a block diagonal matrix with eigenvalues are simply those of $M_1, M_2, \dots, M_d$. For any $j = 1, \dots, d$, the eigenvalues of M_j are the root of the characteristic polynomial $\sigma^2 - (1 + \beta - \alpha\lambda_j)\sigma + \beta$. Since $\alpha = \left(\frac{2}{\sqrt{\lambda_1}+\sqrt{\lambda_d}}\right)^2, \beta = \left(\frac{\sqrt{\lambda_1}-\sqrt{\lambda_d}}{\sqrt{\lambda_1}+\sqrt{\lambda_d}}\right)^2$, the two complex roots are given by

$$\sigma_{j_1,j_2} = \frac{1}{2}\left(1 + \beta - \alpha\lambda_j \pm \sqrt{(1 + \beta - \alpha\lambda_j)^2 - 4\beta}\right).$$

It follows that the magnitudes of all eigenvalues are equal to $\sqrt{\beta}$. Thus $\rho(M) = \sqrt{\beta}$.

[1] Gelfand's formula.

3 Proof of convergence rate for adaptive step size gradient descent

From (1), we have

$$f(x^{(k+1)}) = f(x^{(k)}) - \alpha_k \nabla f(x^{(k)})^T \nabla f(x^{(k)}) + \frac{1}{2}\alpha_k^2 \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}).$$

Substituting $\alpha_k = \frac{\nabla f(x^{(k)})^T \nabla f(x^{(k)})}{\nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)})}$, we obtain

$$\begin{aligned}
f(x^{(k+1)}) &= f(x^{(k)}) - \frac{1}{2}\frac{\left(\nabla f(x^{(k)})^T \nabla f(x^{(k)})\right)^2}{\nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)})} = f(x^{(k)}) - \frac{1}{2}\frac{\left(x^{(k)T}\Lambda^2 x^{(k)}\right)^2}{x^{(k)T}\Lambda^3 x^{(k)}} \\
&= \left(1 - \frac{\left(x^{(k)T}\Lambda^2 x^{(k)}\right)^2}{\left(x^{(k)T}\Lambda^3 x^{(k)}\right)\left(x^{(k)T}\Lambda x^{(k)}\right)}\right) f(x^{(k)}) \\
&\leq \left(1 - \frac{4\lambda_1\lambda_d}{(\lambda_1+\lambda_d)^2}\right) f(x^{(k)}) = \left(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\right)^2 f(x^{(k)})
\end{aligned}$$

The last inequality uses Kantorovich Inequality

$$\frac{\left(y^T\Lambda^2 y\right)^2}{\left(y\Lambda^3 y\right)\left(y^T\Lambda y\right)} \geq \frac{4\lambda_1\lambda_d}{(\lambda_1+\lambda_d)^2}.$$

4 Proof of convergence rate for adaptive step size momentum method

Proof. For asymptotic analysis, we consider the region near the optimum, in which the objective function can be well-approximated by a quadratic. We know that fixing $\alpha^{(k)}$ to $\frac{2}{\lambda_1+\lambda_d}$ yields

$$\left\|y^{(k+1)}\right\|_2 \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\left\|y^{(k)}\right\|_2.$$

On the other hand, choosing adaptive step size

$$\begin{bmatrix}\hat{\alpha} \\ \hat{\beta}\end{bmatrix} = \begin{bmatrix}\nabla f^T \nabla^2 f \nabla f & -\Delta x^T \nabla^2 f \nabla f \\ -\Delta x^T \nabla^2 f \nabla f & \Delta x^T \nabla^2 f \Delta x\end{bmatrix}^{-1} \begin{bmatrix}\nabla f^T \nabla f \\ -\Delta x^T \nabla f\end{bmatrix}$$

minimizes the quadratic with respect to α, β. That means the resulting $\hat{y}^{(k)}$ satisfies

$$\left\|\hat{y}^{(k+1)}\right\|_2 \leq \left\|y^{(k+1)}\right\|_2 \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\left\|y^{(k)}\right\|_2.$$

Hence, each iteration of adaptive schedule decreases the distance at least as much as each iteration of fixed step size scheme. The convergence rate therefore is upper-bounded by the one of fixed step size scheme inside the quadratic region, which is $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.

□