6 In the stable regime, convergence-rate of (GD+M) is $\frac{1}{(1-\beta)}$ larger than (GD) using classical convergence analysis

In this section, we show that when (GD) with a learning-rate h and (GD+M) with an effective learning rate $\frac{h}{(1-\beta)}$ both fall inside the stable regime of (GD),then the convergence-rate of (GD+M) is $\frac{1}{(1-\beta)}$ larger than (GD).

Classical convergence of (GD) and (GD+M) is considered in a locally quadratic surface. On a standard quadratic, the minimization is $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - b^T\mathbf{x} + c$, where $\mathbf{A}$ is positive semi-definite matrix with eigen-values in $[\mu, L]$. A simple change of variable would mean doing a minimization of the form $\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}$, where Σ contains the eigenvalues of A on the diagonal. Hence $\nabla f(\mathbf{x}) = \Sigma\mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = \Sigma$. Furthermore, the condition number of the objective function is denoted as $\kappa = \frac{L}{\mu}$.

For Heavy-Ball method, the iterates follow:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla f(\mathbf{x}^k) + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) \tag{6.1}$$

On a locally quadratic, the iterates roughly follow

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\Sigma\mathbf{x} + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) = ((1+\beta)\mathbf{I} - h\Sigma)\mathbf{x}^k - \beta\mathbf{x}^{k-1} \tag{6.2}$$

With slight rearrangement, which could be written as :

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix} \tag{6.3}$$

Denoting $\mathbf{y}^k = \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix}$ and $\mathbf{T} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, the norm of $\|\mathbf{y}^k\|_2$ is derived as follows:

$$\|\mathbf{y}^k\| = \|\mathbf{T}\mathbf{y}^{k-1}\| = \|\mathbf{T}^k\mathbf{y}^0\| \le \|\mathbf{T}^k\|_2\|\mathbf{y}^0\| \le (\rho(\mathbf{T}))^k\kappa(V)\|\mathbf{y}^0\| \tag{6.4}$$

where $\rho(\mathbf{T})$ is the spectral radius of $\mathbf{T}$ and $\mathbf{T}$ has an eigen-decomposition $\mathbf{T} = VDV^{-1}$, $\kappa(V)$ being the condition number of V. $\mathbf{T}$ is permutation-similar to the block-diagonal matrix $\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & . & . & \mathbf{0} \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \mathbf{T}_n \end{bmatrix}$, where $\mathbf{T}_j = \begin{bmatrix} 1+\beta-\alpha\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}$ is a 2×2 matrix for $j = 1, 2..n$. Letting r_j denote the eigen-values for each block matrix $\mathbf{T}_j$ and would satisfy $r_j = \begin{cases} \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm \sqrt{(1+\beta-h\lambda_j)^2 - 4\beta}), & \text{if } (1+\beta-h\lambda_j)^2 - 4\beta = \Delta_j > 0 \\ \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm i\sqrt{|\Delta_j|}, & \text{otherwise} \end{cases}$ where $i = \sqrt{-1}$. Due to the block-matrix structure of $\mathbf{T}$, the convergence factor $\rho(\mathbf{T})$ is determined by the largest vectors among all the block matrices $\mathbf{T}_j$, i.e, $\rho(\mathbf{T}) = \max_j r_j = \max r_1, r_n$.

Now depending upon the 4 conditions $\Delta_j \le 0 \equiv \beta \ge (1-\sqrt{h\lambda_j})$, $\Delta_j > 0 \equiv \beta \le (1-\sqrt{h\lambda_j})$, $|1-\sqrt{h\mu}| < |1-\sqrt{hL}|$ and $|1-\sqrt{h\mu}| > |1-\sqrt{hL}|$, we have four sub-cases to determine $\rho(\mathbf{T})$:

1. If $0 < h \le (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta \ge (1-\sqrt{h\mu})^2$
2. If $0 < h \le (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta < (1-\sqrt{h\mu})^2$
3. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta \ge (\sqrt{hL}-1)^2$
4. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ and $\beta < (\sqrt{hL}-1)^2$

For a small h and fixed β, satisfies condition-2 and the effective learning rate lies in the stability regime of GD. Under this particular condition (2), we have $\Delta_1 > 0$, hence the spectral radius $\rho(\mathbf{T})$ becomes (by taking the larger r_j) :

$$\rho^{(GD+M)} = \frac{1}{2}(1+\beta-h\mu+\sqrt{(1+\beta-h\mu)^2 - 4\beta}) \quad \text{[considering the larger term]} \tag{6.5}$$

$$= \frac{1}{2}(1+\beta-h\mu+\sqrt{(1-\beta)^2 - 2h\mu(1+\beta) + h^2\mu^2}) \tag{6.6}$$

$$= \frac{1}{2}(1+\beta-h\mu+(1-\beta)(\underbrace{\sqrt{1-\frac{2h\mu(1+\beta)+h^2\mu^2}{(1-\beta)^2}}}_{1-\frac{1}{2}\frac{2h\mu(1+\beta)}{(1-\beta)^2}+O(h^2)} -1) + (1-\beta)) \tag{6.7}$$

$$\approx \frac{1}{2}(1+\beta-h\mu-\frac{h\mu(1+\beta)}{(1-\beta)} + (1-\beta)) \quad \text{[small } h \text{ approximation]} \tag{6.8}$$

$$= 1 - \frac{h\mu}{(1-\beta)} \tag{6.9}$$

Similarly, for (GD) with learning-rate $\tilde{h}$ minimizing a locally quadratic function, using the classical convergence approach, we have $\|\mathbf{x}^k\| \le \rho_{\tilde{h}}^k\|\mathbf{x}^0\|$ where $\rho_{\tilde{h}} = \max(|1-\tilde{h}\mu|, |1-\tilde{h}L|)$. Hence for a small enough h i.e,($0 < \tilde{h} \le \frac{2}{L+\mu}$), we have for the convergence rate for GD to be :

$$\rho^{GD} = 1 - \tilde{h}\mu \tag{6.10}$$

Putting $\tilde{h} = \frac{h}{(1-\beta)}$, we see that $\rho^{(GD+M)} \approx \rho^{(GD)}$. Which means if we use a learning rate $\frac{1}{(1-\beta)}$ times larger for GD, it will match the convergence rate of (GD+M).

Equivalently under the same learning rate for (GD) and (GD+M) (say h), the convergence rate of (GD+M) is $\frac{1}{(1-\beta)}$ times larger than that of (GD),i.e, $\rho^{(GD+M)} \approx \frac{1}{(1-\beta)}\rho^{(GD)}$.