CS726 - Lyapunov analysis and the Heavy Ball Method

Benjamin Recht

Department of Computer Sciences, University of Wisconsin-Madison
1210 W Dayton St, Madison, WI 53706
email: brecht@cs.wisc.edu

October 15, 2012

1 Heavy Ball Method

The *Heavy Ball Method* is a two-step procedure defined by the following state transitions:

$$p_k = -\nabla f(x_k) + \beta_k p_{k-1} \tag{1.1a}$$
$$x_{k+1} = x_k + \alpha_k p_k \tag{1.1b}$$

for some initial points x_0 and p_0, and some positive sequences α_k and β_k. Typically, we just set $p_0 = 0$. This algorithm can be re-written as the iteration

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}). \tag{1.2}$$

where the term $x_k - x_{k-1}$ is referred to as *momentum*. The iteration is equivalent to a discretization of the second order ODE

$$\ddot{x} + a\dot{x} + b\nabla f(x) = 0 \tag{1.3}$$

which models the motion of a body in a potential field given by f with friction. This motivated the initial naming of the algorithm by Polyak. In these notes, compare the Heavy Ball Method to Steepest descent on quadratic functions, showing that the former achieves an asymptotically optimal convergence rate.

2 Problem Set-up

We aim to minimize $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = \tfrac{1}{2} x^T A x - b^T x + c$$

where A is an $n \times n$ positive definite matrix, b is a vector and c is a constant. We assume that $\ell I \preceq A \preceq LI$. This problem has a unique minimizer given by $x_* = A^{-1} b$.

3 Analysis of the Gradient Method

Consider the standard gradient method

$$x_{k+1} = x_k - t_k \nabla f(x_k) \tag{3.1}$$

for some starting point x_0 and some sequence of steps t_k.

Let x_* denote the minimizer of f. x_* is unique because of strong convexity. We will now show that the iterates of the gradient method converge to x_* at a linear rate. We will do this by examining the function $\|x_k - x_*\|$ and show that this function is monotonically decreasing if we select the proper step size t. A function which decreases along the trajectory of an optimization algorithm is commonly called a *Lyapunov function*.

Observe that

$$\begin{aligned}
\|x_{k+1} - x_*\| &= \|x_k - t_k \nabla f(x_k) - x_*\| && \text{(3.2a)}\\
&= \|x_k - t_k(Ax_k - b) - x_*\| && \text{(3.2b)}\\
&= \|(I - t_k A)(x_k - x_*)\| && \text{(3.2c)}\\
&\leq \|I - t_k A\| \|x_k - x_*\| && \text{(3.2d)}\\
&\leq \max\{|1 - t_k L|, |1 - t_k \ell|\} \|x - x_*\| \,. && \text{(3.2e)}
\end{aligned}$$

Here, the first equality follows by the definition of x_{k+1}, the second follows from plugging in the gradient, and the third follows because $b = Ax_*$. Inequality (3.2d) follows from the definition of the operator norm. The final inequality follows because

$$(1 - t_k \ell) I \preceq I - t_k A \preceq (1 - t_k L) I \,.$$

Note that $t_k = \frac{2}{L+\ell}$ minimizes (3.2e) for all k. Setting t_k to this value, we find that

$$\|x_{k+1} - x_*\| \leq \left(\frac{L - \ell}{L + \ell}\right) \|x_k - x_*\| \tag{3.3}$$

or, denoting $\kappa = \frac{L}{\ell}$ and $D_0 = \|x_0 - x_*\|$,

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k D_0 \tag{3.4}$$

That is, a constant step-size policy converges at a linear rate.

4 Analysis of the Heavy Ball Method

We restrict our attention to the case where α_k and β_k are fixed constants. To prove this method converges, we use a time-lagged version of the standard Lyapunov function. That is, instead of looking at $\|x_{k+1} - x_*\|^2$, we examine $\|x_{k+1} - x_*\|^2 + \|x_k - x_*\|^2$. For this invariant, we have the chain of inequalities

$$\begin{aligned}
\begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} &= \begin{bmatrix} x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) - x_* \\ x_k - x_* \end{bmatrix} && \text{(4.1a)}\\
&= \begin{bmatrix} x_k + \beta(x_k - x_{k-1}) - \alpha(Ax_k - b) - x_* \\ x_k - x_* \end{bmatrix} && \text{(4.1b)}\\
&= \begin{bmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} \,. && \text{(4.1c)}
\end{aligned}$$

Again, we use the fact that $Ax_* = b$. Define the matrix

$$T = \begin{bmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{bmatrix}.$$

Iterating the above calculation we have that

$$\left\| \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} \right\| = \|T^k\| \left\| \begin{bmatrix} x_1 - x_* \\ x_0 - x_* \end{bmatrix} \right\|. \tag{4.2}$$

Hence, it suffices to bound the norm of T^k to get a convergence rate. We use following theorem from matrix anaylsis:

Proposition 4.1 *Let M be an $n \times n$ matrix. Let $\rho(M) = \max_i |\lambda_i(M)|$. Then there exists a sequence $\epsilon_k \geq 0$ such that*

$$\|M^k\| \leq (\rho(M) + \epsilon_k)^k \text{ and } \lim_{k\to\infty} \epsilon_k = 0\,.$$

$\rho(M)$ is called the *spectral radius* of M and is equal to the maximum magnitude of any eigenvalue of M. We can bound the spectral radius using the following

Proposition 4.2 *For $\beta \geq \max\{|1 - \sqrt{\alpha\ell}|, |1 - \sqrt{\alpha L}|\}^2$, $\rho(T) \leq \beta$.*

Proof Let $U\Lambda U^T$ be an eigendecomposition of A. Let Π be the $2n \times 2n$ matrix with entries

$$\Pi_{i,j} = \begin{cases} 1 & i \text{ odd},\ j = i \\ 1 & i \text{ even},\ j = 2n + i \\ 0 & \text{otherwise} \end{cases}. \tag{4.3}$$

Then, by conjugationm we have

$$\Pi \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \begin{bmatrix} (1+\beta)I - \alpha A & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \Pi^T \tag{4.4a}$$

$$= \Pi \begin{bmatrix} (1+\beta)I - \alpha\Lambda & -\beta I \\ I & 0 \end{bmatrix} \Pi^T \tag{4.4b}$$

$$= \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & T_n \end{bmatrix}. \tag{4.4c}$$

Where

$$T_i := \begin{bmatrix} 1 + \beta - \alpha\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}. \tag{4.5}$$

That is, T is similar to the block diagonal matrix with 2×2 diagonal blocks T_i. To compute the eigenvalues of T, it suffices to compute the eigenvalues of all of the T_i. For fixed i, the eigenvalues of the 2×2 matrix are roots of the equation

$$u^2 - (1 + \beta - \alpha\lambda_i)u + \beta = 0 \tag{4.6}$$

In the case that $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$, the roots of the characteristic equations are imaginary, and both have magnitude β. Note that by assumption

$$(1 - \sqrt{\alpha\lambda_i})^2 \leq \max\left\{(1 - \sqrt{\alpha\ell})^2, (1 - \sqrt{\alpha L})^2\right\} \tag{4.7}$$

and setting β equal to the right hand side completes the proof. ■

Setting $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\ell})^2}$ and $\beta = \frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}$ yields

$$\left\|\begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix}\right\| \leq \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} + \epsilon_k\right)^k \left\|\begin{bmatrix} x_1 - x_* \\ x_0 - x_* \end{bmatrix}\right\| \tag{4.8}$$

Or, in other words

$$\|x_k - x_*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \epsilon_k\right)^k D_0 \tag{4.9}$$

which is the rate attainable by the nonlinear conjugate gradient method. Of course, we need to know L and ℓ to achieve such a rate.