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 10, 2010

1 Problem Set-up

Let $f : \mathbb{R}^n \to \mathbb{R}$. Assume

1. f is twice differentiable.
2. f is strongly convex: there exists a constant $\ell > 0$ such that

$$f(z) \geq f(x) + \nabla f(x)^T(z-x) + \frac{\ell}{2}\|z-x\|^2 \,. \tag{1.1}$$

3. ∇f is Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$. Note that, by Taylor's Theorem, this means that

$$f(z) \leq f(x) + \nabla f(x)^T(z-x) + \frac{L}{2}\|z-x\|^2 \,. \tag{1.2}$$

2 Lipschitz gradients

Since f is strongly convex and has a Lipschitz continuous gradient, it follows that for all vectors x and y and all positive scalars t

$$\|x - t\nabla f(x) - (y - t\nabla f(y))\| \leq \max\{|1-tL|, |1-t\ell|\}\|x-y\| \,. \tag{2.1}$$

To see this, note that

$$\|x - t\nabla f(x) - (y - t\nabla f(y))\| \leq \left\| \int_0^\infty (I - t\nabla^2 f(x + t(y-x))(y-x)dt \right\| \tag{2.2}$$

$$\leq \sup_z \|I - t\nabla^2 f(z)\| \|y-z\| \,. \tag{2.3}$$

Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least ℓ and the maximum eigenvalue is at least L. Therefore the eigenvalues of $I - t\nabla^2 f(z)$ are at most $\max(1-tL, 1-t\ell)$ and at least $\min(1-tL, 1-t\ell)$. Therefore, $\|I - t\nabla^2 f(z)\| \leq \max(|1-tL|, |1-t\ell|)$.

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

$$\|x_{k+1} - x_*\| = \|x_k - t_k \nabla f(x_k) - x_*\| \leq \max\{|1 - t_k L|, |1 - t_k \ell|\} \|x - x_*\|. \tag{3.2a}$$

Here, the first equality follows by the definition of x_{k+1} and because x_* is optimal. The inequality follows from (2.1). Note that $t_k = \frac{2}{L+\ell}$ minimizes the right hand side 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 Heavy Ball Method

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

$$p_k = -\nabla f(x_k) + \beta_k p_{k-1} \tag{4.1a}$$

$$x_{k+1} = x_k + \alpha_k p_k \tag{4.1b}$$

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

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}). \tag{4.2}$$

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

$$\ddot{x} + a\dot{x} + b\nabla f(x) = 0 \tag{4.3}$$

which models the motion of a body in a potential field given by f with friction. This motivate the initial naming of the algorithm by Polyak.

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

$$\left\| \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}) - x_* \\ x_k - x_* \end{bmatrix} \right\|^2 \tag{4.4a}$$

$$= \left\| \begin{bmatrix} x_k + \beta_k (x_k - x_{k-1}) - x_* \\ x_k - x_* \end{bmatrix} - \alpha_k \begin{bmatrix} \nabla f(x_k) \\ 0 \end{bmatrix} \right\|^2 \tag{4.4b}$$

$$= \left\| \begin{bmatrix} (1+\beta_k) I & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} - \alpha_k \begin{bmatrix} \nabla f(x_k) \\ 0 \end{bmatrix} \right\|^2 \tag{4.4c}$$

$$\leq \sup_z \left\| \begin{bmatrix} (1+\beta_k) I - \alpha_k \nabla^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix} \right\|^2 \left\| \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} \right\|^2 . \tag{4.4d}$$

Again, the first equality follows by the definition of x_{k+1} and because x_* is optimal. The final inequality follows because we assume ∇f is differentiable. The supremum is finite because we assume ∇f is Lipschitz. Indeed, this supremum can be bounded using the following

Proposition 4.1 *For* $\beta_k = \max\{|1 - \sqrt{\alpha_k \ell}|, |1 - \sqrt{\alpha_k L}|\}^2$,

$$\sup_z \left\| \begin{bmatrix} (1+\beta_k) I - \alpha_k \nabla^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix} \right\| \leq \sqrt{\beta_k} \tag{4.5}$$

Proof Fix z, and let $U \Lambda U^*$ be an eigendecomposition of $\nabla^2 f(z)$. Then by conjugation we have

$$\left\| \begin{bmatrix} (1+\beta_k) I - \alpha_k \nabla^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix} \right\| = \left\| \begin{bmatrix} U^* & 0 \\ 0 & U^* \end{bmatrix} \begin{bmatrix} (1+\beta_k) I - \alpha_k \nabla^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \right\| \tag{4.6a}$$

$$= \left\| \begin{bmatrix} (1+\beta_k) I - \alpha_k \Lambda & -\beta_k I \\ I & 0 \end{bmatrix} \right\| \tag{4.6b}$$

$$= \max_{1 \leq i \leq d} \left\| \begin{bmatrix} 1 + \beta_k - \alpha_k \lambda_i & -\beta_k \\ 1 & 0 \end{bmatrix} \right\| \tag{4.6c}$$

For fixed i, the eigenvalues of the 2×2 matrix are roots of the equation

$$\rho^2 - (1 + \beta_k - \alpha_k \lambda_i) + \beta_k = 0 \tag{4.7}$$

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

$$(1 - \sqrt{\alpha_k \lambda_i})^2 \leq \max \left\{ (1 - \sqrt{\alpha_k \ell})^2, (1 - \sqrt{\alpha_k L})^2 \right\} \tag{4.8}$$

and setting β_k equal to the right hand side completes the proof. ∎

Plugging this bound into (4.4d),

$$\left\| \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} \right\|^2 \leq \max \left\{ (1 - \sqrt{\alpha_k \ell})^2, (1 - \sqrt{\alpha_k L})^2 \right\} \left\| \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} \right\|^2 \tag{4.9}$$

Letting $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\ell})^2}$, we have

$$\left\| \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} \right\|^2 \leq \left(\frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} \right)^2 \left\| \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} \right\|^2 \tag{4.10}$$

Or, in other words

$$\|x_k - x_*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k D_0 \tag{4.11}$$

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.

References