EE 381V: Large Scale Optimization **Fall 2012**

Lecture 23 — November 20

Lecturer: Caramanis & Sanghavi *Scribe: John Bridgman and Sifeng Lin*

23.1 Last Time

Last time, we talked about lower and upper bounds on the number of iterations for the algorithms that we have seen so far for optimization. We saw that sub-gradient descent got $O(1/\epsilon^2)$ iterations for error ϵ and that this was optimal. Then, we saw that gradient descent got $O(1/\epsilon)$ and mentioned that the lower bound using the oracle model was $O(1/\sqrt{\epsilon})$. This lower bound is achievable, but one has to use more than just one iteration of information in the update rule. The key idea for most of the algorithms is to add a term that is called momentum. So, the update rule then becomes something of the form:

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

We then saw the FISTA algorithm would get the lower bound with $\beta_k = \frac{k-2}{k+1}$.

We also saw that the gradient descent algorithm for a function that is strongly convex and smooth gives a convergence in terms of $\left(\frac{\kappa-1}{\kappa+1}\right)^n$, where $\kappa = \frac{M}{m}$ and $mI \preceq \bigtriangledown^2 f(x) \preceq MI$. This is the upper bound on the convergence for the algorithm. Last time, we saw the lower bound for this upper bound is $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$.

23.2 Introduction

It will be instructive for later discussion in these notes to recall how we get the above bound for gradient descent.

Setup: f such that gradient of f is M-Lipschitz and f is m-strongly convex. This is the same as saying that $mI \preceq \bigtriangledown^2 f(x) \preceq MI$.

Claim 23.1.

$$\|x_{k+1} - x^\star\|_2 = \|x_k - \alpha_k \bigtriangledown f(x_k) - x^\star\|_2 \leq \max\{|1 - \alpha_k M|, |1 - \alpha_k m|\} \|x_k - x^\star\|_2.$$

Proof:

$$\|x - \alpha \bigtriangledown f(x) - (y - \alpha \bigtriangledown f(y)\| = \left\| \int_0^1 \left[I - \alpha \bigtriangledown^2 f(x + \tau(x-y)) \right] (x-y) d\tau \right\| \tag{23.1}$$

$$\leq \int_0^1 \left\| \left[I - \alpha \bigtriangledown^2 f(x + \tau(x-y)) \right] (x-y) \right\| d\tau \tag{23.2}$$

$$\leq \int_0^1 \left\| I - \alpha \bigtriangledown^2 f(x + \tau(x-y)) \right\| \|(x-y)\| d\tau \tag{23.3}$$

$$\leq \sup_z \left\| I - \alpha \bigtriangledown^2 f(z) \right\| \|x - y\|_2. \tag{23.4}$$

This norm on the matrix evaluates to the largest singular value. So, $\sup_z \|I - \alpha \bigtriangledown^2 f(z)\| \leq \max\{|1 - \alpha_k M|, |1 - \alpha_k m|\}$ by our definition of M and m and this proves the claim above. □

Now, if we take $\alpha = \frac{2}{m+M}$; then, Claim 23.1 is:

$$\|x_{k+1} - x^\star\|_2 \leq \max\left\{ \left| 1 - \frac{2M}{m+M} \right|, \left| 1 - \frac{2m}{m+M} \right| \right\} \|x_k - x^\star\|_2 = \frac{M-m}{M+m} \|x_k - x^\star\|_2.$$

Now, we can multiply them together k times and cancel out terms that are the same on both sides and substitute in $\kappa = M/m$ to get:

$$\|x_{k+1} - x^\star\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x_0 - x^\star\|_2.$$

23.3 Heavy Ball Method

This method is not a descent method. We have seen other methods that were not descent methods, like FISTA and sub-gradient descent. With FISTA, we introduced a Lyapunov function. A valid Lyapunov function is one where the function is non-negative and only zero at $x^\star$. For this case, we will be defining a Lyapunov function to show convergence.

The Heavy Ball Algorithm is:

$$P_k = - \bigtriangledown f(x_k) + \bar{\beta}_k P_{k-1} \qquad x_{k+1} = x_k + \alpha_k P_k.$$

Recall the standard form above where we had the momentum term. These two update equations can then be rewritten in that form as:

$$x_{k+1} = x_k - \alpha_k \bigtriangledown f(x_k) + \frac{\alpha_k}{\alpha_{k-1}} \bar{\beta}_k (x_k - x_{k-1}).$$

For most of the rest of this discussion, we will take $\beta_k = \frac{\alpha_k}{\alpha_{k-1}} \bar{\beta}_k$.

Now, we will derive the Lyapunov function for the Heavy Ball method. Start with:

$$\|x_{k+1} - x^\star\|_2^2 + \|x_k - x^\star\|_2^2.$$

This can be rewritten in a vector form:

$$\left\| \begin{bmatrix} x_{k+1} - x^\star \\ x_k - x^\star \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} x_k - \alpha_k \bigtriangledown f(x_k) + \beta_k(x_k - x_{k-1}) - x^\star \\ x_k - x^\star \end{bmatrix} \right\|_2^2 \tag{23.5}$$

$$= \left\| \begin{bmatrix} (1+\beta_k)I & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} - \alpha_k \begin{bmatrix} \bigtriangledown f(x_k) \\ 0 \end{bmatrix} \right\|_2^2. \tag{23.6}$$

We have that:

$$\bigtriangledown f(x_k) = \int_0^1 \bigtriangledown^2 f(x_k + \tau(x_k - x^*))(x_k - x^*)d\tau = \int_0^1 \bigtriangledown^2 f(x_k + \tau(x_k - x^*))d\tau (x_k - x^*).$$

So, we can rewrite the above expression as:

$$\begin{aligned}
&\left\| \begin{bmatrix} (1+\beta_k)I & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} - \alpha_k \begin{bmatrix} \int_0^1 \bigtriangledown^2 f(x_k + \tau(x_k - x^*))d\tau & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2 \\
&= \left\| \int_0^1 \begin{bmatrix} (1+\beta_k)I - \alpha_k \bigtriangledown^2 f(x_k + \tau(x_k - x^*)) & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} d\tau \right\|_2^2 \\
&\leq \int_0^1 \left\| \begin{bmatrix} (1+\beta_k)I - \alpha_k \bigtriangledown^2 f(x_k + \tau(x_k - x^*)) & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2 d\tau \\
&\leq \int_0^1 \left\| \begin{bmatrix} (1+\beta_k)I - \alpha_k \bigtriangledown^2 f(x_k + \tau(x_k - x^*)) & -\beta_k I \\ I & 0 \end{bmatrix} \right\|_2^2 d\tau \left\| \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2 \\
&\leq \sup_z \left\| \begin{bmatrix} (1+\beta_k)I - \alpha_k \bigtriangledown^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix} \right\|_2^2 \left\| \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2.
\end{aligned}$$

Notice how similar this was to our earlier example for gradient descent. Let

$$A(z) = \begin{bmatrix} (1+\beta_k)I - \alpha_k \bigtriangledown^2 f(z) & -\beta_k I \\ I & 0 \end{bmatrix}.$$

Then, we will show that $\sup_z \|A(z)\| \leq \sqrt{\beta_k}$, where we take

$$\beta_k = \left(\max\{|1 - \sqrt{\alpha_k m}|, |1 - \sqrt{\alpha_k M}|\} \right)^2.$$

We know that $\bigtriangledown^2 f(z)$ is positive semi-definite. This means that it admits an eigenvalue decomposition. That is to say, there exists some matrix U_z that is orthonormal (where the columns have magnitude one and all columns are perpendicular to each other) such that $\bigtriangledown^2 f(z) = U_z \Lambda_z U_z^T$, where Λ_z is non-zero only on its diagonal.

We will use the following property:

Remark 23.1. *If U is an orthonormal matrix and A is any matrix, then*

$$\sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|=1} \|UAU^\star x\|.$$

Then,

$$\|A(z)\| = \left\| \begin{bmatrix} U_z^\star & 0 \\ 0 & U_z^\star \end{bmatrix} A(z) \begin{bmatrix} U_z & 0 \\ 0 & U_z \end{bmatrix} \right\|.$$

$U_z^\star$ is the conjugate of U_z.

Then, because $U_z^\star U_z = I$ and $U_z^\star \bigtriangledown^2 f(z) U_z = \Lambda_z$; this simplifies to:

$$\left\| \begin{bmatrix} (1+\beta_k)I - \alpha_k \Lambda_k & -\beta_k I \\ I & 0 \end{bmatrix} \right\|.$$

Notice, that this array is a block matrix of diagonal matrices. So, it looks like:

$$\begin{bmatrix} \ddots & \ddots \\ \ddots & 0 \end{bmatrix}.$$

Then,

$$\|A(z)\| = \max_i \left\| \begin{array}{cc} 1+\beta_k - \alpha_k \lambda_{i,z} & -\beta_k \\ 1 & 0 \end{array} \right\|.$$

This matrix is invertible (non-zero determinant), so the norm just returns the maximum eigenvalue. The eigenvalues of the 2 by 2 matrix are the solutions to

$$\lambda^2 - \lambda(1 + \beta_k - \alpha_k \lambda_{i,z}) + \beta_k = 0.$$

The roots of this equation are:

$$\frac{1 + \beta_k - \alpha_k \lambda_{i,z} \pm \sqrt{(1+\beta_k - \alpha_k \lambda_{i,z})^2 - 4\beta_k}}{2}.$$

We now will take $\beta_k = (1 - \sqrt{\alpha_k \lambda_{i,z}})^2$ and this becomes:

$$\lambda^2 - \lambda(1 + (1 - \sqrt{\alpha_k \lambda_{i,z}})^2 - \alpha_k \lambda_{i,z}) + (1 - \sqrt{\alpha_k \lambda_{i,z}})^2 = 0.$$

When $\beta_k = (1 - \sqrt{\alpha_k \lambda_{i,z}})^2$, the roots are $1 - \sqrt{\alpha_k \lambda_{i,z}}$, which is $\sqrt{\beta_k}$.

Consider the term under the square root: $(1 + \beta_k - \alpha_k \lambda_{i,z})^2 - 4\beta_k$. When this is negative, the roots of the characteristic equation are imaginary.

$$(1 + \beta_k - \alpha_k \lambda_{i,z})^2 - 4\beta_k < 0 \tag{23.7}$$

$$(\beta_k - \alpha_k \lambda_{i,z} - 1)^2 - 4\alpha_k \lambda_{i,z} < 0 \tag{23.8}$$

$$(\beta_k - \alpha_k \lambda_{i,z} - 1)^2 < 4\alpha_k \lambda_{i,z} \tag{23.9}$$

$$-2\sqrt{\alpha_k \lambda_{i,z}} < \beta_k - \alpha_k \lambda_{i,z} - 1 < 2\sqrt{\alpha_k \lambda_{i,z}} \tag{23.10}$$

$$(1 - \sqrt{\alpha_k \lambda_{i,z}})^2 < \beta_k < (1 + \sqrt{\alpha_k \lambda_{i,z}})^2. \tag{23.11}$$

When the roots are imaginary, the magnitude is

$$\sqrt{\frac{4\beta_k - (1 + \beta_k - \alpha_k \lambda_{i,z})^2 + (1 + \beta_k - \alpha_k \lambda_{i,z})^2}{4}} = \sqrt{\beta_k}.$$

So, we can conclude that if $\beta_k \geq (1 - \sqrt{\alpha_k \lambda_{i,z}})^2$; then, $\sup_z \|A(z)\| \leq \sqrt{\beta_k}$.
We also assume that

$$(1 - \sqrt{\alpha_k \lambda_{i,z}})^2 \leq \max\{(1 - \sqrt{\alpha_k m})^2, (1 - \sqrt{\alpha_k M})^2\}.$$

Then we can plug this into the equation above to get

$$\left\| \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2 \leq \max\{(1 - \sqrt{\alpha_k m})^2, (1 - \sqrt{\alpha_k M})^2\} \left\| \begin{bmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{bmatrix} \right\|_2^2.$$

If we let $\alpha_k = \frac{4}{(\sqrt{m}+\sqrt{M})^2}$, we get

$$(1 - \sqrt{\alpha_k m})^2 = (1 - \frac{\sqrt{4m}}{\sqrt{m} + \sqrt{M}})^2 = \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^2 = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^2.$$

Substitute this back in and iterate the equation to get

$$\|x_k - x^\star\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^\star\|_2,$$

which is the result we desired.

23.4 Mirror Descent

Mirror descent is primarily concerned with optimizing general convex functions. In order to motivate mirror descent, we will step back and think about the expression for the bound on the number of iterations for general convex functions. Recall that for sub-gradient descent, the bound was

$$f_N - f^\star \leq \frac{LR}{\sqrt{N+1}}.$$

We saw earlier that the $O(1/\sqrt{N})$ was optimal; but, what about the other constants? These other constants may depend upon the dimension and other parameters of the problem. Primarily in this course, we have been looking at these algorithms through the lens of the problem being very large in some sense. For example, minimizing $f(x)$ with $x \in \mathbb{R}^n$ such that n is large enough that algorithms with complexity $O(n^2)$ are effectively intractable. So, this motivates seeking ways to decrease constants that are parameterizable by n. The parameter R is the error of our initial guess and the parameter L is a bound on the magnitude of the gradient. Both of these terms can possibly be of $O(n)$ or more. This motivates the idea of finding a way to decrease these constants. The key idea behind mirror descent is to notice for sub-gradient descent both these constants are in terms of the l^2-norm. It may be that these constants could be considerably less if expressed in terms of some other norm.

23.4.1 A look at proximal methods

Remember in the proximal method we assumed the function could be decomposed into the sum of two other functions where they had nicer properties. In other words,

$$f(x) = g(x) + h(x),$$

where $g(x)$ is smooth and convex and $h(x)$ is convex and "simple". Then, the update method is

$$x_+ = \arg\min_u h(u) + g(x) + \langle \bigtriangledown g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2.$$

Recall, if $h(x) = 0$, this simplifies to gradient descent. If $h(x) = I_X(x)$ (the indicator function of set X), this simplified to projected gradient descent. Consider

$$x_+ = \arg\min_u \langle \bigtriangledown g(x), u - x\rangle + \frac{1}{2t}\|u - x\|_2^2.$$

What happens when we replace the norm at the end of this expression by some other norm? This is the idea behind mirror descent. The norm we will choose is called a Bregman Divergence. Now, remember that the reason we want to do this is to reduce the complexity. So, if the other norm is not easy to compute, there is no point in considering it. Also if the bounds do not improve, then there is no point in considering it.

Definition 23.1 (Definition: Bregman Divergence).

$$D(u, v) = \omega(u) - \omega(v) - \langle \bigtriangledown \omega(v), u - v\rangle,$$

where ω is a "Distance Generating Function" and ω is continuously differentiable and m-strong convex.

Remark 23.2. *If $\omega(u) = \|u\|_2^2$, then $D(u, v) = \|u - v\|_2^2$.*

Proof:

$$D(u, v) = \langle u, u\rangle - \langle v, v\rangle - \langle v, u - v\rangle = \langle u, u\rangle - 2\langle u, v\rangle = \langle u - v, u - v\rangle = \|u - v\|_2^2.$$

□

Remark 23.3. *KL divergence is also a Bregman Divergence. Recall that the KL divergence is:*

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)},$$

where $p(x)$ and $q(x)$ are two probability distribution. The ω that produces this is $\omega(u) = \sum_x u(x) \log u(x) - \sum_x u(x) = H(u) - 1$.

23.4.2 Algorithm

Assume that we are trying to minimize $f(x)$ subject to $x \in X$, where X is bounded and convex and $f(x)$ is convex. The mirror descent algorithm is:

$$x_+ = \arg\min_{u \in X} \langle \alpha g - \bigtriangledown \omega(x), u \rangle + \omega(u),$$

where $g \in \partial f(x)$.

When $\omega(u) = \frac{1}{2}\|u\|_2^2$, this is just

$$x_+ = Proj_X(x - \alpha g).$$

Projection onto a set X can be written as

$$\arg\min_{u \in X} \|u - x + \alpha g\|^2 = \arg\min_{u \in X} \langle u - x + \alpha g, u - x + \alpha g \rangle \tag{23.12}$$

$$= \arg\min_{u \in X} 2\langle \alpha g - x, u \rangle + \langle u, u \rangle \tag{23.13}$$

$$= \arg\min_{u \in X} \langle \alpha g - \bigtriangledown \|x\|, u \rangle + \frac{1}{2}\|u\|^2. \tag{23.14}$$

Which can be written as

$$x_+ = \arg\min_{u \in X} \langle \alpha g - \bigtriangledown \omega(x), u \rangle + \omega(u),$$

with $\omega(u) = \frac{1}{2}\|u\|_2^2$.

23.4.3 Convergence

In this section, we will not cover the convergence, but just give a little bit of introduction to give the flavor of the convergence. Next class, convergence will be covered in detail. First, we will take a quick look at convergence of sub-gradient descent and discuss how the equations will change for mirror descent.

The equation for sub-gradient descent convergence is

$$\|x_{k+1} - x^\star\|_2^2 \leq \|x_k - \alpha_k g_k - x^\star\|_2^2 = \|x_k - x^\star\|_2^2 - 2\alpha_k \langle g_k, x_k - x^\star \rangle + \alpha_k^2 \|g\|_2^2.$$

Notice that for projected gradient descent, the inequality is only because $x_k - \alpha g$ can be outside the set; otherwise, this would be equality. This implies that the inequality holds for any u in the set X. By rearranging the terms, we can then write:

$$\alpha \langle g, x - u \rangle - \frac{1}{2}\alpha^2 \|g\|_2^2 \leq \frac{1}{2}\|x - u\|_2^2 - \frac{1}{2}\|x_+ - u\|_2^2, \quad \forall u \in X.$$

The second part of this equation can be taken as a Lyapunov function for our iterations. We next replace the necessary terms as we did above with the ω function and iterate, resulting in a very similar bound to what we get for sub-gradient descent.