EE-599: Mathematics of High Dimensional Data

University of Southern California

Lecturer: Mahdi Soltanolkotabi September 2018 parts are adapted from student scribes

Lecture Note 3

1 Recap: L-smoothness and m-strong convexity and its properties

Definition 1 *(L-smooth functions) A function $f(x)$ is L-smooth if its gradient is Lipschitz continuous. For any $x, y \in \Omega$,*

$$|\nabla f(x) - \nabla f(y)| \leq L\|x - y\|,$$

where $L \geq 0$. If $f(x)$ is both convex and L-smooth, then it has the following properties:

$$g(x) = \frac{L}{2}x^T x - f(x) \quad is \;\; convex,$$
$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_{\ell 2}^2,$$
$$\nabla^2 f(x) \preceq LI.$$

Definition 2 *(m-strongly convex functions) A function $f(x)$ is m-strongly convex iff one of the following equivalent conditions holds. For any $x, y \in \Omega$,*

$$g(x) = f(x) - \frac{m}{2}x^T x \quad is \;\; convex,$$
$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq m\|x - y\|^2,$$
$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}\|y - x\|^2,$$
$$\nabla^2 f(x) \succeq mI.$$

2 Quadratic Lower Bound

With m-strong convexity, the following is true

$$\frac{m}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|^2$$

Implications:

- $\nabla f(x)$ is small $\implies$ $f(x) - f(x^*)$ and $\|x - x^*\|$ are small.
- f has a unique minimizer.

Theorem 1 *If $f(x)$ is m-strongly convex, and L-smooth, then for $\mu = \frac{1}{L}$, and the gradient descent update being*

$$x_{t+1} = x_t - \mu \nabla f(x_t).$$

The following convergence result can be achieved

$$f(x_t - x^*) \leq \left(1 - \frac{m}{L}\right)^t (f(x_1) - f(x^*))$$

Proof The quadratic bound gives us:

$$\begin{aligned}
f(x-\mu\nabla f(x)) &\leq f(x) - \mu\left(1-\frac{\mu L}{2}\right)\|\nabla f(x)\|_{\ell 2}^2 \\
\implies f(x^+) - f(x^*) &\leq f(x) - f(x^*) - \mu\left(1-\frac{\mu L}{2}\right)\|\nabla f(x)\|_{\ell 2}^2 \\
\implies f(x^+) - f(x^*) &\leq (1-\mu m(2-\mu L))(f(x)-f(x^*)) \\
&\leq \left(1-\frac{m}{L}\right)(f(x)-f(x^*)) \\
\implies f(x_t - x^*) &\leq \left(1-\frac{m}{L}\right)^t (f(x_1)-f(x^*))
\end{aligned}$$

which concludes the proof. ■

The convergence rate can be obtained as

$$\begin{aligned}
f(x_t - x^*) &\leq \left(1-\frac{m}{L}\right)^t (f(x_1)-f(x^*)) \leq \epsilon \\
\implies t &\geq \frac{\log\left(\frac{\epsilon}{f(x_1)-f(x^*)}\right)}{\log\left(1-\frac{m}{L}\right)}
\end{aligned}$$

Note: $\frac{m}{L}$ gives an upper bound on the condition number of the hessian of f.

Theorem 2 *If $f(x)$ is twice differentiable, m-strongly convex, and L-smooth, then for $0 < \mu \leq \frac{2}{m+L}$, and the gradient descent iteration as*

$$x_{t+1} = x_t - \mu\nabla f(x_t).$$

The following convergence result can be achieved:

$$\|x_{t+1} - x^*\|^2 \leq \left(1-\frac{2\mu mL}{m+L}\right)^t \times \|x_0 - x^*\|^2.$$

Special case: $\mu = \frac{2}{m+L}$

$$\begin{aligned}
\|x_t - x^*\| &\leq \left(\frac{\kappa-1}{\kappa+1}\right)^t \|x_0 - x^*\| \\
&= \left(1-\frac{2}{\kappa+1}\right)^t \|x_0 - x^*\| \\
&\leq \exp\left(\frac{-2t}{\kappa+1}\right)\|x_0 - x^*\|
\end{aligned}$$

The convergence rate can then be obtained as

$$\begin{aligned}
\exp\left(\frac{-2t}{\kappa+1}\right)\|x_0 - x^*\| &\leq \epsilon \\
\implies t &\geq \frac{K+1}{2}\log\left(\frac{\|x_0 - x^*\|}{\epsilon}\right)
\end{aligned}$$

3 Strong convexity and smoothness is necessary for contractivity

Theorem 3 *Strong convexity and smoothness is necessary for contractivity, which means*

$$\Phi(x) = x - \mu\nabla f(x) \;\; is \;\; contractive \;\; if \quad \|\Phi(x) - \Phi(y)\| \leq \beta \|x - y\|.$$

Furthermore, if f is twice differentiable and $\Phi(x) = x - \mu\nabla f(x)$ is contractive then f must be strongly covnex.

$$\frac{1}{t}\|\Phi(x) - \Phi(y)\| \leq \beta \|x - y\|$$

Let $y = x + t\Delta x$, we have

$$\begin{aligned}
\beta \|\Delta x\| &\geq \lim_{t\to 0}\frac{1}{t}\|\Phi(x + t\Delta x) - \Phi(x)\|, \\
&= \lim_{t\to 0}\left\|\Delta x - \frac{\mu}{t}(\nabla f(x + t\Delta x) - \nabla f(x))\right\|, \\
&= \left\|(I - \mu\nabla^2 f(x))\Delta x\right\|.
\end{aligned}$$

$$\left\|(I - \mu\nabla^2 f(x))\right\| \leq \beta \Rightarrow \frac{1-\beta}{\mu} I \preceq \nabla^2 f(x) \preceq \frac{1+\beta}{\mu} I.$$

Theorem 4 *Let f be m-strong convex and L-lipschitz in Ω. Then PGD with $\mu_s \leq \frac{2}{m(s+1)}$ obeys*

$$f(\sum_{s=1}^{t}\frac{2s}{t(t+1)}x_s) - f(x^*) \leq \frac{2L^2}{m(t+1)}.$$

4 Lower bounds for Black box models

In general, a black-box procedure is a mapping from "history" to the next query point, that is it maps $\{x_1, g_1, \ldots, x_t, g_t\}$ (with $g_s \in \partial f(x_s)$) to x_{t+1}. Assume $x_1 = 0$ and for any $t > 0$, x_{t+1} is in the linear span of $g_1, g_2, \ldots, g_t$, that is

$$x_{t+1} \in Span(g_1, g_2, \ldots, g_t).$$

Theorem 5 *Let $t \leq n, L, R > 0$. There exists a convex and $L-$Lipschitz function $f(x)$ such that for any black-procedure,*

$$\min_{1\leq s\leq t} f(x_s) - \min_{\|x\|\leq R} f(x) \geq \frac{RL}{2(1+\sqrt{t})}.$$

There also exists and m-strongly convex and L-Lipschitz function $f(x)$ such that for any black-box procedure,

$$\min_{1\leq s\leq t} f(x_s) - \min_{\|x\|\leq \frac{L}{2m}} f(x) \geq \frac{L^2}{8mt}.$$

Theorem 6 *Let $t \leq (n-1)/2, L > 0$. There exists a $L-$smooth convex function $f(x)$ such that for any black-box procedure,*

$$\min_{1\leq s\leq t} f(x_s) - f(x^*) \geq \frac{3L}{32}\frac{\|x_1 - x^*\|^2}{(t+1)^2}.$$

Theorem 7 *Let $\kappa > 1$. There exists a m-strongly convex and L-Lipschitz function $f(x) : \ell_2 \to \mathbb{R}$ with $\kappa = L/m$ such that for any $t \geq 1$ one has,*

$$f(x_t) - f(x^*) \geq \frac{m}{2}(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2(t-1)}\|x_1 - x^*\|^2.$$

5 Momentum methods

Intuition: Look at different algorithms as differential equations. Gradient is sort of like

$$\frac{dx}{dt} = -\nabla f(x)$$

A fixed point occurs when $\nabla f(x) = 0$. But more often we have damping in the differential equation: $\alpha\frac{d^2x}{dt^2} = -\nabla f(x) - b\frac{dx}{dt}$ Discretizing the above equation gives:

$$\frac{x(t+\Delta t) - 2x(t) + x(t-\Delta t)}{\Delta t^2} \approx -\nabla f(x(t)) - b\frac{x(t) - x(t-\nabla t)}{\nabla t}$$

$$\Rightarrow x(t+\Delta t) = x(t) - \frac{\Delta t^2}{\alpha}\nabla f(x(t)) + \left(1 - \frac{b}{a}\Delta t\right)(x(t) - x(t-\Delta t))$$

$$\Rightarrow x_{t+1} = x_t - \alpha\nabla f(x_t) + \eta(x_t - x_{t-1}).$$

Where α, η are two parameters,

$$\begin{aligned} y_t &= x_{t+1} - x_t \\ &= -\mu\nabla f(x_t) + \eta(x_t - x_{t-1}) \\ &= -\mu\nabla f(x_t) + \eta y_{t-1} \end{aligned}$$

This can also be re-written slightly differently:

$$\begin{aligned} x_{t+1} &= x_t + y_t \\ y_t &= -\mu\nabla f(x_t) + \eta(y_{t-1}) \qquad \text{(Heavy Ball)} \end{aligned}$$

6 Nesterov's Accelerated Gradient Descent

Previously, we said that the gradient descent has a rate of convergence $1/t$ after t steps for an L-smooth convex function. With Nesterov's Accelerated Gradient, we can attain a better rate of order $1/t^2$. For the case of L-smooth and m-strongly convex function, the accelerated scheme provides better convergence rate as well.

Algorithm:

$$\begin{cases} x_{t+1} = x_t + y_t \\ y_t = -\mu \nabla f\left(x_t + \eta_t y_{t-1}\right)) + \eta_t y_{t-1} \end{cases}$$

or

$$\begin{cases} z_t & = x_t + \eta_t (x_t - x_{t-1}) \\ x_{t+1} & = z_t - \mu \nabla f(z_t) \end{cases}$$

Nesterov is better for general functions f, for quadratic the convergence is the same.

Theorem 8 *Let f be an L-smooth and m-strongly convex function, if we run Nesterov's accelerated gradient descent with*

$$\begin{cases} z_t & = x_t + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(x_t - x_{t-1}) \\ x_{t+1} & = z_t - \frac{1}{L}\nabla f(z_t) \end{cases},$$

then the following is satisfied

$$f(x_t) - f(x^*) \leq \frac{m+L}{2} \|x_1 - x^*\|^2 \exp\left(\frac{-t-1}{\sqrt{\kappa}}\right).$$

Theorem 9 *Let f be an L-smooth convex function, if we run Nesterov's accelerated gradient descent with*

$$\begin{cases} z_t & = x_t + \eta_t (x_t - x_{t-1}) \\ x_{t+1} & = z_t - \frac{1}{L}\nabla f(z_t) \\ \eta_t & = \theta_t (\frac{1}{\theta_t} - 1) \\ \theta_t & = \frac{1}{2}\left(-\theta_{t-1}^2 + \sqrt{\theta_{t-1}^4 + \theta_{t-1}^2}\right), \theta_0 = 1 \end{cases},$$

then the following is satisfied

$$f(x_t) - f(x^*) \leq \frac{4L}{(t+2)^2} \|x_0 - x^*\|^2.$$

7 Conjugate Gradient Method

The conjugate gradient method is an algorithm for finding the nearest local minimum of a function of n variables which presupposes that the gradient of the function can be computed. It uses conjugate directions instead of the local gradient for going downhill.

Algorithm:

1. Initialize $p_1 = -\nabla f(x_0)$
2. In step t, $p_t = -\nabla f(x_t) + \beta p_{t-1}$ and $x_{t+1} = x_t + \alpha p_t$,

   where

$$\alpha = \underset{\eta}{\operatorname{argmin}} f(x_t + \eta p_t) \text{ and } \beta = \frac{\|\nabla f(x_{t-1})\|_{\ell 2}^2}{\|\nabla f(x_{t-2})\|_{\ell 2}^2}$$

The most famous variant of conjugate gradient is applied to quadratic losses where

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}.$$

In this case the algorithm takes the following form

- **Initialization:** $\boldsymbol{x}_0 = 0$ and $\boldsymbol{r}_0 = \boldsymbol{b}$
- FOR $t = 1, 2, \ldots$
  1. If $t = 1$, take $\boldsymbol{p}_t = r_0$; otherwise, take
  $$\boldsymbol{p}_t = \boldsymbol{r}_{t-1} + \beta\boldsymbol{p}_{t-1} \quad \text{where} \quad \beta = -\frac{\boldsymbol{p}_{t-1}^T\boldsymbol{A}\boldsymbol{r}_{t-1}}{\boldsymbol{p}_{t-1}^T\boldsymbol{A}\boldsymbol{p}_{t-1}}$$
  2. Compute
  $$\alpha = \frac{\|\boldsymbol{r}_{t-1}\|_{\ell_2}^2}{\boldsymbol{p}_t^T\boldsymbol{A}\boldsymbol{p}_t}, \quad \boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \alpha\boldsymbol{p}_t, \quad \boldsymbol{r}_t = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_t.$$

  ENDFOR

8 Newton's method

8.1 Newton's method for solving equations

Want $f(x) = 0$.

$$\phi(x) = \phi(x_0) + \phi'(x_0)(x - x_0) + o(|x - x_0|)$$

Plugging in $x = x_0 + \Delta x$ we get,

$$\phi(x_0 + \Delta x) = \phi(x_0) + \phi'(x_0)\Delta x + o(|\Delta x|)$$

If we can solve

$$\phi(x_0) + \phi'(x_0)\Delta x = 0$$

then $\phi(x_0 + \Delta x) = o(|\Delta x|)$ and we get fast convergence. So we want,

$$\Delta x = -\frac{\phi(x_0)}{\phi'(x_0)}$$

Hence the iteration becomes:

$$x_{t+1} = x_t - \frac{\phi(x_t)}{\phi'(x_t)}$$

8.2 Newton's method in $\mathbb{R}^d$

Given a non-linear map $F : \mathbb{R}^d \to \mathbb{R}^d$ and we want to solve $F(x) = 0$. With $J_F(x)$ being the jacobian of F at x, the first order Taylor's approximation is:

$$F(x + \Delta x) = F(x) + J_F(x)\Delta x + o(\|\Delta x\|)$$

Solving this for $F(x + \Delta x) = 0$ we get,

$$\Delta x = -J_F^{-1}(x)F(x)$$

Iteration: $x_{t+1} = x_t - J_F^{-1}(x_t)F(x_t)$

In general for solving $\nabla f(x) = 0$,

$$\nabla f(x + \Delta x) \approx \nabla f(x) + \nabla^2 f(x)\Delta x$$
$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1}\nabla f(x_t)$$

More generally one uses the damped iterations

$$x_{t+1} = x_t - \mu_t[\nabla^2 f(x_t)]^{-1}\nabla f(x_t)$$

Theorem 10 *If the following are true:*

- f *is twice continuously differentiable*
- $\nabla^2 f$ *is L-Lipschitz in the operator norm:*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$$

  for all x and y

- $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succeq mI, mn > 0$
- $\|x_0 - x^*\| \leq \frac{2m}{3L}$

then Newton' method shows the following properties:

1. $\|x_t - x^*\| \leq \frac{2m}{3L} \quad \forall t$
2. $\|x_{t+1} - x^*\| \leq \frac{2m}{3L}\|x_t - x^*\|^2 \quad \forall t$

Some points to observe about Newton's method:

- If f is not convex, we get a local minimum in the neighborhood.
- This method has quadratic convergence i.e. it has $O\left(\log\log\frac{1}{\epsilon}\right)$ iterations required for ϵ-optimality.

9 Quasi-Newton Methods

Both the gradient method and Newton's method are iterative approximations.
Gradient Method:

$$
\begin{aligned}
f(x) &\approx f(x_t) + < \nabla f(x_t), x - x_t > + \frac{\alpha_t}{2}\|x - x_t\|^2 \\
\text{minimizer:} \quad & x_t - \frac{1}{\alpha_t}\nabla f(x_t)
\end{aligned}
$$

Newton's Method:

$$
\begin{aligned}
f(x) &\approx f(x_t) + < \nabla f(x_t), x - x_t > + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t) \\
\text{minimizer:} \quad & x_t - (\nabla^2 f(x_t))^{-1} \nabla f(x_t) \\
\text{damped:} \quad & x_t - \mu_t (\nabla^2 f(x_t))^{-1} \nabla f(x_t)
\end{aligned}
$$

One interpretation of the gradient method is that it provides an approximation of the hessian as a diagonal scalar matrix. Quasi-newton methods take this analogy one step further and approximate the hessian with some other matrix B_t:

$$
f(x) \approx f(x_t) + < \nabla f(x_t), x - x_t > + \frac{1}{2}(x - x_t)^T B_t (x - x_t)
$$

which leads to the update:

$$
\text{damped:} \quad x_{t+1} = x_t - \mu_t B_t^{-1} \nabla f(x_t)
$$

Instead of computing B_t afresh at every iteration it can be updated in a simple manner to account for curvature measured during the most recent step.

9.1 BFGS Method

Suppose we have generated a new x_{t+1} and wish to construct a new quadratic model:

$$
m_t(x) = f(x_t) + < \nabla f(x_t), x - x_t > + \frac{1}{2}(x - x_t)^T B_{t+1} (x - x_t)
$$

We can impose the following requirements on m_t to get B_{t+1}:

1. $\nabla m_t(x_t) = \nabla f(x_t)$
2. $\nabla m_t(x_{t-1}) = \nabla f(x_{t-1})$

These requirements capture the fact that if the last two gradients are correct then the hessian should also be pretty good. This gives:

$$
\nabla m_t(x_{t-1}) = \nabla f(x_t) - B_t(x_{t+1} - x_t)
$$

If we let $s_t = x_{t+1} - x_t$, $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$, $B_{t+1}s_t = y_t$, $H_t = B_t^{-1}$, this implies $s_t = H_{t+1}y_t$. Lastly, we want B_{t+1} to be close to B_t. This can be done in the W-norm sense to get an elegant analytical update for B_{t+1} in terms of B_t. We define W-norm as:

$$\textbf{W-norm}: \|A\|_W = \|W^{\frac{1}{2}} A W^{\frac{1}{2}}\|_F$$

For the BFGS method,

$$W = \int_0^1 \nabla^2 f(x_t + ts_t) dt$$

If we then minimize $\|H - H_t\|_W$ we get the following update rule:

$$\begin{aligned} \textbf{BFGS:} \quad & H_{t+1} = (I - P_t s_t y_t^T) H_t (I - P_t y_t s_t^T) + P_t s_t s_t^T \\ \text{where} \quad & P_t = \frac{1}{y_t^T s_t} \end{aligned}$$

It is easy to check that $s_t = H_{t+1} y_t$ is being satisfied.
BFGS Algorithm:
Given starting point x_0. convergence tolerance $\epsilon > 0$ and inverse hessian approximation H_0,

$$\begin{aligned} \text{Initialize } t = 0 \\ \text{while } (\|f(x_t)\| > \epsilon) : \\ & P_t = -H_t \nabla f(x_t) \\ & \text{Choose } \alpha_t \text{ by line search obeying Wolfe} \\ & x_{t+1} = x_t + \alpha_t P_t \\ & s_t = x_{t+1} - x_t \\ & y_t = \nabla f(x_{t+1}) - \nabla f(x_t) \\ & H_{t+1} = (I - P_t s_t y_t^T) H_t (I - P_t y_t s_t^T) + P_t s_t s_t^T \end{aligned}$$