Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Shubham Gupta
Instructor: Purushottam Kar
Date: September 29, 2016

LECTURE
15

Quasi-Newton methods

Introduction

The previous lecture discussed the conjugate gradient method and the various formulations of Newton's metod in different problem settings and we saw that Newton method exhibits quadratic rate of convergence under certain constrains.
In this lecture we will look at Quasi-Newton methods of optimization. We will introduce Schatten norms which are widely used in functional analysis. We will look at the secant method and the good and bad broyden methods. We will also introduce the Sherman-Morisson-Woodbury formula used for updation of the inverse of a matrix when it undergoes a perutrbation.

Newton Method Variant

Consider an objective function f which is L_1 lipshitz, $\alpha-$strongly convex and has L_2 lipshitz hessians. Define an optimization algorithm on f which first executes gradient descent until
$\nabla F(x^t) < \epsilon_o$ for some $\epsilon_o > 0$ and then Newton method with $\alpha_t = 1$.

Theorem 15.1. There exists some $\epsilon_o > 0$ such that $\forall \epsilon > 0$, this variant algorithm converges to an $\epsilon-$optimal solution (i.e. $|f - f^*| \le \epsilon$) in $O(\log(\frac{1}{\epsilon_o}) + \log\log(\frac{1}{\epsilon}))$ steps.

Proof. It follows from the linear and quadratic rates of convergence for gradient descent(steepest descent) and newton's method respectively for the given function as shown in previous lectures. Refer to Sec 9.5 of Boyd Vandenberghe (2004) for a detailed proof. □

Remark 15.1. Newton's method doesn't work well in constrained optimization problems.

Schatten Norms

The Schatten norm arise as a generalization of p-integrability similar to the trace class norm and the HilbertSchmidt norm

Definition 15.1. The Schatten p-norms arise when applying the p-norm to the vector of singular values of a matrix. If the singular values of $T \in R^{m\times n}$ are denoted by σ_i, then the Schatten p-norm is defined by

$$\|T\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p \right)^{1/p} \tag{1}$$

Remark 15.2. The Frobenius norm of T defined as $||T||_F = \sqrt{Tr(T^*T)}$ is the schatten-2 norm.

Remark 15.3. Schatten-∞ norm is called spectral norm and schatten-1 norm is the nuclear norm of a matrix. Nuclear norms are widely used in matrix rank minimization problems as convex approximations to the otherwise non-convex problem.

Quasi-Newton Methods

Quasi-Newton methods are an alternative to the Newton method. They can be used if the Jacobian or Hessian is unavailable or is too expensive to compute at every iteration.
Quasi-Newton methods are a generalization of the secant method to find the root of the first derivative for multidimensional problems.

Secant Method

Secant method is a root-finding algorithm which assumes a function to be approximately linear in the region of interest. Each improvement is taken as the point where the approximating line crosses the axis. This succession of roots of secant lines is used to approximate the root of a function.

Definition 15.2. Secant method approximates the first derivative f at x_n with the following finite difference approximation(Boyd Vandenberghe (2004)):

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} \tag{2}$$

$$x_{n+1} = x_n - \frac{1}{f'(x_n)} f(x_n)$$

In general for $g(x) = 0$ where $g : R^n \to R^n$ is a function which takes as input a vector

$$J(x_n)(x_n - x_{n-1}) = g(x_n) - g(x_{n-1}) \tag{3}$$

Approximating $J(x_n)$ using $H_n \in R^{n \times n}$ gives

$$\begin{aligned} H_n(x_n - x_{n-1}) &= g(x_n) - g(x_{n-1}) \\ x_n - x_{n-1} &= S_n(g(x_n) - g(x_{n-1})) \text{ where } [S_n = H_n^{-1}] \end{aligned}$$

Algorithm 1: Quasi-Newton method

```
Input: $f(x)$
Output: $x^t$
Initialize $x^o, S_o$
for $t = 0, 1, 2, \ldots, T$ do
    $\mathbf{d}^t \leftarrow S_t \nabla F(x^t)$
    $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t - \alpha_t \mathbf{d}^t$        //Applying Armijo's rule for $\alpha_t$
    Update $S^{t+1} \leftarrow S^t$
end for
return $\mathbf{x}^T$
```

In the quasi-newton method, we need to ensure that secant rule is obeyed in each iteration

$$\begin{aligned} H_{t+1}(x_{t+1} - x_t) &= \nabla F(x_{t+1}) - \nabla F(x_t) \\ x_{t+1} - x_t &= S_{t+1}(\nabla F(x_{t+1}) - \nabla F(x_t)) \end{aligned}$$

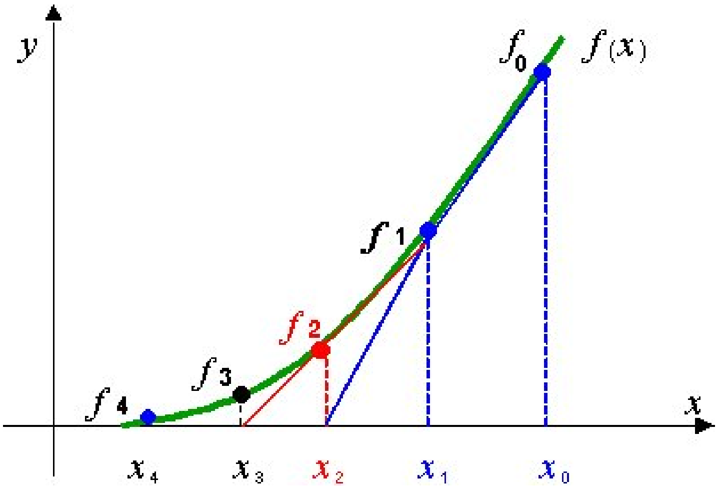

Figure 1: Secant Method: It shows how the quasi-newton method proceeds approximating tangents using secants. Courtesy: `http://ced.kmutnb.ac.th/scc/SlideNumerical/Chapter3/Secant.html`

Broyden's methods

The Broyden Method (1965) developed by C. G. Broyden is an extension of the secant method of root finding to higher dimensions. Computing the Jacobian in each iteration explicitly is an expensive operation. Broyden's method tries to compute the whole Jacobian only at the first iteration, and do a rank-one update at the other iterations.

This method uses the current estimate of the Jacobian H_n and improves it by taking the solution to the secant equation that is a minimal modification to H_n

Definition 15.3. The **Good Broyden update** approximates the jacobian at each step so as to minimize the frobenius norm: $\|H_{t+1} - H_t\|_\text{f}$ such that $H_{t+1}\Delta x^t = \Delta g^t$. Thus,

$$H_{t+1} = \operatorname*{arg\,min}_{H \in R^{n\times n}} \|H - H_t\|_\text{f} = H_t + \frac{\Delta g^t - H_t \Delta x^t}{\|\Delta x^t\|^2} \Delta x^{t\,\text{T}} \tag{4}$$

Lemma 15.2. The optimization problem minimize $\sum_{i=1}^{n} \frac{||a_i||^2}{2}$ such that $A^T X = B$ where $A = [a_1, a_2, ...a_n]^T$, $B = [b_1, b_2, \cdots, b_n]^T$ and $X = [x_1, x_2, \cdots, x_n]^T$ yields $a_i = \frac{b_i x^T}{||x||^2}$.

Proof. Define the langragian for the above problem with dual parameters $\lambda_i' s$ as

$$L(A, \lambda) = \sum_{i=1}^{n} \frac{||a_i||^2}{2} + \sum_{i=1}^{n} \lambda_i ({a_i}^T x - b_i) \tag{5}$$

Stationarity condition (differentiate wrt a_i) gives

$$a_i = \lambda_i x$$

$$a_i^T x = b_i \implies \lambda_i ||x||^2 = b_i$$

Therefore,

$$\lambda_i = \frac{b_i}{||x||^2} \implies a_i = \frac{b_i x^T}{||x||^2}$$

□

Claim 15.3. The good broyden update $H_{t+1} = H_t + \frac{\Delta g^t - H_t \Delta x^t}{\|\Delta x^t\|^2} {\Delta x^t}^{\mathrm{T}}$ minimizes the frobenius norm $\|H_{t+1} - H_t\|_{\mathrm{f}}$.

Proof.

$$\begin{aligned} H_{t+1} &= H_t + E \\ H_t \Delta x^t + E \Delta x^t &= \Delta g^t \\ E \Delta x^t &= \Delta g^t - H_t \Delta x^t \end{aligned}$$

Using lemma 15.2 with $A = E$, $X = \Delta x^t$ and $B = \Delta g^t - H_t \Delta x^t$,

$$\arg\min ||E||_f = \frac{\Delta g^t - H_t \Delta x^t}{\|\Delta x^t\|^2} {\Delta x^t}^{\mathrm{T}}$$

Therefore,

$$H_{t+1} = H_t + \frac{\Delta g^t - H_t \Delta x^t}{\|\Delta x^t\|^2} {\Delta x^t}^{\mathrm{T}}$$

□

Applying Sherman-Morrison-Woodbury formula allows us to update the inverse jacobian estimate which is the required value in the newton update

$$H_{t+1}^{-1} = H_t^{-1} + \frac{\Delta x_t - H_t^{-1} \Delta g^t}{\Delta x_t^{\mathrm{T}} H_t^{-1} \Delta g_t} \Delta x_t^{\mathrm{T}} H_t^{-1} \tag{6}$$

Definition 15.4. The **Bad Broyden method** updates the inverse of the jacobian estimate S_t so as to minimize the Frobenius norm: $\|\mathbf{H}_{t+1}^{-1} - \mathbf{H}_t^{-1}\|_{\mathrm{f}}$ such that $S_{t+1} \Delta g^t = \Delta x^t$

$$S_{t+1} = \underset{S \in R^{n \times n}}{\arg\min} \|\mathbf{S} - \mathbf{S}_t\|_{\mathrm{f}} \tag{7}$$

Using the Sherman-Morrison formula to directly update the inverse of the Jacobian matrix:

$$H_{t+1}^{-1} = H_t^{-1} + \frac{\Delta x_t - H_t^{-1} \Delta g^t}{\|\Delta g^t\|^2} {\Delta g^t}^{\mathrm{T}} \tag{8}$$

Sherman-Morrison-Woodbury formula

The SMW formula describes the inverse of $A + uv^T$ when there is already a inverse for A.

Definition 15.5. Suppose A is an invertible square matrix and u, v are column vectors and suppose that $1 + v^T A^{-1} u \neq 0$.

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}. \tag{9}$$

where, uv^T is the outer product of u and v.

If the inverse of A is already known, the formula provides a numerically cheap way to compute the inverse of A corrected by the matrix uv^T (can be visualized as a perturbation or a rank-1 update). The computation is relatively cheap as $(A + uv^T)^{-1}$ does not have to be computed from scratch which takes $O(n^3)$, but can be computed by correcting A^{-1} in $O(n^2)$ time.

Refer to `http://www.cs.cornell.edu/~bindel/class/cs6210-f09/lec12.pdf` for derivation of this formula through block Gaussian elimination

References

Stephen Boyd and Lieven Vandenberghe *Convex optimization.* Cambridge university press, 2004.

SMW formula, `https://en.wikipedia.org/wiki/Sherman%E2%80%93Morrison_formula`,.

Schatten Norm, `https://en.wikipedia.org/wiki/Schatten_norm`,.

Broyden Method, `https://en.wikipedia.org/wiki/Broyden%27s_method`,.