Convex Optimization

Problem set 2

Due Monday November 4th

1 Gradent Decent without Linesearch

In this problem we will consider gradient descent with predetermined step sizes. That is, instead of determining $t^{(k)}$ by a linesearch method using the objective function, the current iterate $x^{(k)}$ and the descent direction $\triangle x^{(k)}$, it will be set to some pre-determined sequence.

1. For a strongly convex twice-continuously differentiable function $f(x)$ with bounded Hessian, $mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$, $\kappa = M/m$, consider gradient descent with a fixed step size $t^{(k)} = \frac{1}{M}$. Prove that with this step size, after

$$k = \frac{1}{\log\left(\frac{1}{1-\frac{1}{\kappa}}\right)} \log\left(\frac{f(x^{(0)}) - p^*}{\epsilon}\right) \tag{1}$$

   iterations, $x^{(k)}$ will be ϵ-suboptimal.

   How many gradient evaluations are performed to reach an ϵ-suboptimal solution? How many function evaluations?

2. The above choice of step-size requires knowing beforehand a bound on the Hessian. You will now show that the choice of a fixed (equal for all iterations) stepsize *must* depend on the function (or at least on the magnitude of its Hessian).

   For any t, show a twice-continuously differentiable strongly convex function $f(x)$ with bounded Hessian, and an initial point $x^{(0)}$ such that gradient descent with fixed stepsize $t^{(k)} = t$ starting at $x^{(0)}$ yields a sequence of iterates that does *not* converge to the optimum.

   Note that we requrie the Hessian be bounded, i.e. there exists some M s.t. $\nabla^2 f(x) \preccurlyeq MI$ for all x, but the bound M will of course depend on t.

   Hint: it is enough to consider scalar (one-dimensional) quadratic functions.

3. If we do not know in advance a bound on the Hessian, we can use a predetermined decreasing sequence of step-sizes $t^{(k)}$. Consider a strongly convex twice-differentiable function f :

$\mathbb{R}^n \to \mathbb{R}$ with bounded Hessian, $mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$. Prove that for any predetermined sequence of step-sizes obeying

$$\lim_{k\to\infty} t^{(k)} = 0 \quad \text{and} \quad \sum_{k=0}^{\infty} t^{(k)} = \infty, \tag{2}$$

gradient descent on $f(\cdot)$ with steps of sizes $t^{(k)}$ converges to the optimal value: $\lim_{k\to\infty} f(x^{(k)}) = \inf_x f(x)$.

Hint: first establish that for large enough k, $f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{2}t^{(k)} \left\|\nabla f(x^{(k)})\right\|^2$. Next bound the norm of the gradient in terms of the suboptimality. Now establish that for any ϵ, we will reach an ϵ-suboptimal point after a finite number of iterations.

Note that the choice of step-sizes does *not* depend on the function or on the magnitude of the Hessian, and so we can run the method without prior knowledge of a bound on the Hessian. However, the number of iterations required to reach an ϵ-suboptimal solution will depend on the magnitude of the Hessian.

4. **[Extra Credit]** Can you bound the number of iterations needed to reach an ϵ-suboptimal solution using the method above? The bound will depend on the Hessian of the function, and on the specific choice of stepsize sequence. Can you suggest a predetermined sequence of stepsizes, that does not depend on the function to be optimized, and results in an iteration bound very similar to (1)? This will estbalish that at least from a theoretical worst-case perspective, the line-search is not really necesairy and we can get similar gurantees even with predetermined stepsizes.

2 Newton's Method

1. In this question we will formaly prove the affine invarience of Newton's method. Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and an affine transform $y \in \mathbb{R}^m \mapsto Ay + b$ where $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Define $g(y) = f(Ay + b)$.

   (a) For $x = Ay + b$, let $\triangle x$ and $\triangle y$ be the Newton steps for $f(x)$ and $g(y)$ respectively. Prove that $\triangle x$ = $A\triangle y$.

   (b) Prove that for any $t > 0$, the exit condition for backtracking linesearch on $f(x)$ in direction $\triangle x$ will hold if and only if the exit condition holds for $g(y)$ in direction $\triangle y$.

   (c) Consider running Newton's method on $g(\cdot)$ starting at some $y^{(0)}$ and on $f(\cdot)$ starting at $x^{(0)} = Ay^{(0)} + b$. Use the above to prove that the sequences of iterates obeys $x^{(k)} = Ay^{(k)} + b$ and $f(x^{(k)}) = g(y^{(k)})$.

   (d) Prove that Newton's decrement for $f(\cdot)$ at x is equal to Newton's decrement for $g(\cdot)$ at y, and so the stopping conditions are also identical.

(e) Now consider a function $h(x) = cf(x)$ for some scalar $c > 0$. Prove that the Newton search directions $\triangle x$ and step sizes are the same when optimizing $h(\cdot)$ and $f(\cdot)$. Conclude that the sequence of Newton iterates is the same when optimizing $h(\cdot)$ or $f(\cdot)$.

(f) Will the Newton decrement and the stopping condition also be the same for $h(\cdot)$ and $f(\cdot)$?

2. We defined self-concordance of a scalar function using the condition $|f'''(t)| \leq 2(f''(t))^{3/2}$. The constant 2 in this definition is arbitrary, and this definition depends on the scaling of the function $f(t)$. In this problem, we will consider r-self concordant functions.

   For $r > 0$, we say that a convex scalar function is r-self concordant iff for all t:

$$|f'''(t)| \leq r(f''(t))^{3/2}$$

   We say $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is r-self concordant if $t \mapsto f(x + tv)$ is r-self-concordant for all $x, v \in \mathbb{R}^n$.

   (a) First, establish that r-self-concordancy is afine invarient. That is, if $f(x)$ is r-self concordant, then $g(y) = f(Ay + b)$ is also r-self concordant.

   (b) Show that $f(t) = -\log t$ is self-concordant under the simpler definition, but $f(t) = -\frac{1}{2}\log t$ is not. For what value of r is it r-self concordant?

   (c) What functions are r-self concordant for all $r > 0$?

   (d) If a function $f(\cdot)$ is r-self concordant, for what scaling constant $c > 0$ is the functions $\tilde{f}(x) = cf(x)$ self-concordant under the simpler definition?

   (e) Consider unconstrained minimization of a strictly convex function $f(\cdot)$ that is r-self concordant. Provide a bound on the suboptimality $f(x) - p^*$ in terms of r and the Newton decrement $\lambda(x)$, for small enough values of $\lambda(x)$ (be sure to specify for what values of $\lambda(x)$ the bound is valid). (See Section 9.6.3 of Boyd and Vandenberghe. Hint: how do the suboptimality and the Newton decrement scale when the function is scaled?).

   (f) Provide a bound on the number of iterations needed to reach an ϵ-suboptimal solution using Newton's method for a r-self concordant function. (Hint: Establish that the Newton iterations, that is both the search direction and the line are invarient to scaling

3 Steepest Descent

In this problem, we will define the direction of Steepest Descent with respect to a norm, and investigate it for several specific norms. For a norm $\|\cdot\|$ on $\mathbb{R}^n$, a direction of steepest descent of

$f : \mathbb{R}^n \to \mathbb{R}$ at x with respect to $\|\cdot\|$ is given by:

$$\begin{aligned}\triangle x &\overset{\text{def}}{=} \arg\min_{\|v\|=1} \nabla f(x)^T v = \arg\min_{\|v\|=1} \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t} \\ &\approx \lim_{t\to 0} \arg\min_{\|v\|=1} \frac{f(x+tv)-f(x)}{t} = \lim_{t\to 0} \arg\min_{\|v\|=t} f(x+v)\end{aligned}$$

where the $\approx$ should be interpreted only as an intuitive correspondence, and so the second line only as a rough intuition (e.g. the $\arg\min$ might have multiple minimizers, making the limit not properly defined. It is even possible that some directions of steepest descent, as defined by the first line, are not the limit point for any sequence of minimizers in the second line). But roughly speaking, $\triangle x$ is the direction v in which the greatest decrease in $f(x)$ can be achieved subject to an infinitesimally small constraint on $\|v\|$.

A steepest descent method uses the direction of steepest descent as a descent direction at each iteration, taking a step in this direction with a stepsize chosen by some linesearch method.

1. Show that the direction of steepest descent for the Euclidean ℓ_2 norm is $\triangle x = -\nabla f(x)$, and thus steepest descent with respect to the Euclidean norm is just gradient descent.

2. For $H \succ 0$, consider the norm $\|x\| = x^T H x$. What is the direction of steepest descent with respect to this norm?

3. What are the directions of steepest descent with respect to the ℓ_1 norm $\|x\| = \sum_i |x_i|$? Explain how to find these directions, and think of how a steepest descent method w.r.t. the ℓ_1 norm would proceed.

4. What are the directions of steepest descent with respect to the ℓ_∞ norm $\|x\| = \max_i |x_i|$? Explain how to find these directions, and think of how a steepest descent method w.r.t. the ℓ_∞ norm would proceed.

4 Conjugate Direction Methods

Recall that $v^{(1)}, \ldots, v^{(k)} \in \mathbb{R}^n$ are H-conjugate iff for every $i \neq j$ we have $\left(v^{(i)}\right)^T H v^{(j)} = 0$. That is, $\tilde{v}^{(i)} = H^{1/2} v^{(i)}$ are orthogonal.

4.1 Conjugate Direction Minimization of a Quadratic Objective

Let $f(x) = \frac{1}{2} x^T H x - b^T x$, with H positive semi-definite, be a convex quadratic objective. Let $\Delta x^{(0)}, \ldots, \Delta x^{(n-1)}$ be non-zero H-conjugate directions. Consider iterative minimization along these directions, starting from some $x^{(0)}$:

```
1. For $i = 0$ to $n-1$
2.     $t^{(i)} \leftarrow \arg\min_t f\left(x^{(i)} + t\Delta x^{(i)}\right)$
3.     $x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$
```

4.1.1

Prove that:

$$t^{(i)} = \frac{\left(\Delta x^{(i)}\right)^T \left(Hx^{(i)} - b\right)}{\left(\Delta x^{(i)}\right)^T H\Delta x^{(i)}}$$

4.1.2

The principal result about conjugate directions is that the current point $x^{(k)}$ at each step k of the method above minimizes the quadratic objective $f(x)$ over the k-dimensional affine subspace spanned by $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$. That is:

$$x^{(k)} = \arg\min_{x \in M^k} f(x) \tag{3}$$

where

$$M^k = \left\{ x \mid x = x^0 + \sum_{i=0}^{k-1} \beta_i \Delta x^{(i)}, \beta_i \in \mathbb{R} \right\}$$

Prove equation (3):

1. Show that for all $i < k$: $\nabla f\left(x^{(k)}\right)^T \Delta x^{(i)} = \nabla f\left(x^{(i+1)}\right)^T \Delta x^{(i)}$. (Hint: write $x^{(k)}$ in terms of $x^{(i+1)}$, $t^{(i+1)}, \ldots, t^{(k-1)}$ and $\Delta x^{(i+1)}, \ldots, \Delta x^{(k-1)}$)
2. Show that $\nabla f\left(x^{(i+1)}\right)^T \Delta x^{(i)} = 0$. Conclude that $\nabla f\left(x^{(k)}\right)^T \Delta x^{(i)} = 0$ for $i < k$. (Hint: Consider the derivative of $f\left(x^{(i)} + t\Delta x^{(i)}\right)$ with respect to t).
3. Prove equation (3) by considering the derivatives of $x^0 + \sum_{i=0}^{k-1} \beta_i \Delta x^{(i)}$ with respect to β_i.

4.2 Generating Conjugate Directions

Let $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$ be H-conjugate and d a non-zero vector which is not spanned by $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$. Let

$$\Delta x^{(k)} = d - \sum_{i=0}^{k-1} \frac{d^T H \Delta x^{(i)}}{\left(\Delta x^{(i)}\right)^T H\Delta x^{(i)}} \Delta x^{(i)} \tag{4}$$

4.2.1

Prove that $\Delta x^{(0)}, \ldots, \Delta x^{(k)}$ are H-conjugate and that they span the same subspace as $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}, d$.

4.3 The Conjugate Gradient Method for a Quadratic Function

In the conjugate gradient method for a quadratic function $f(x) = \frac{1}{2}x'Hx - b'x$, each iteration starts with the negative gradient $d = -\nabla f(x)$ and applies equation (4) to obtain only the part of d that is conjugate to all previous directions:

```
For $i = 0$ to $n-1$
    $d^{(i)} = -\nabla f\left(x^{(i)}\right)$
    If $d^{(i)} = 0$ then terminate
    Calculate $\Delta x^{(i)}$ using equation (4)
    $t^{(i)} = \frac{\left(\Delta x^{(i)}\right)^T \left(Hx^{(i)} - b\right)}{\left(\Delta x^{(i)}\right)^T H \Delta x^{(i)}}$
    $x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$
```

4.3.1

Explain why after running the above method, if the method does not terminate early, than $x^{(n)}$ is an optimal point. If the method does terminate early, the last iterate is an optimal point.

4.3.2

The key to the conjugate gradient method is that the calculation of the direction $\Delta x^{(i)}$ can be greatly simplified. In particular, we have:

$$\Delta x^{(k)} = d^{(k)} + \beta^{(k)} \Delta x^{(k-1)} \tag{5}$$

with

$$\beta^{(k)} = \frac{\left(d^{(k)}\right)^T d^{(k)}}{d^{(k-1)} d^{(k-1)}} \tag{6}$$

Prove equation (5):

1. Prove that $d^{(k)}$ is orthogonal to $\Delta x^{(0)}, \ldots, \Delta x^{(k-1)}$ and hence also to $d^{(0)}, \ldots, d^{(k-1)}$. (Hint: Use the partial optimality property given in equation (3)).
2. Show that $t^{(i)} H \Delta x^{(i)} = d^{(i)} - d^{(i+1)}$. (Hint: expand the gradients and consider the update rule for $x^{(i+1)}$).
3. Using the above relation and the orthogonality of $d^{(0)}, \ldots, d^{(k)}$, evaluate $\left(d^{(i)}\right)^T H \Delta x^{(j)}$ for $j < i$. (Hint: For all but one value of j, this will be zero).
4. Similarly, evaluate $\left(\Delta x^{(j)}\right)^T H \Delta x^{(j)}$.
5. Substitute the above two relations into equation (4) and obtain equation (5), with $\beta^{(k)}$ expressed in terms of $d^{(k)}$, $d^{(k-1)}$ and $\Delta x^{(k-1)}$. Now, show that $\beta^{(k)}$ can be calculated as in equation (6) by expanding $\Delta x^{(k-1)}$ using equation (5), the orthogonality of $d^{(k)}$ and $d^{(k-1)}$ and the orthogonality of $\Delta x^{(k-2)}$ and $d^{(k)} - d^{(k-1)}$.

This concludes the proof of equations (5) and (6). We will actually prefer a slightly different form of equation (6):

$$\beta^{(k)} = \frac{\left(d^{(k)}\right)^T \left(d^{(k)} - d^{(k-1)}\right)}{d^{(k-1)} d^{(k-1)}} \tag{7}$$

6. Show that equation (7) is also valid and equivalent to equation (6) (when minimizing a quadratic function with exact line search).

Each iteration of the method therefore requires only vector-vector operations with computational cost $O(n)$, once the gradient has been computed. For a quadratic function, the most expansive operation is therefore computing the gradient which takes time $O(n^2)$.

5 Optional: Quasi-Newton Methods

You are not require to do this problem. It provides additional detail about BFGS

In quasi-Newton methods the descent direction is given by:

$$\Delta x^{(k)} = -D^{(k)} \nabla f\left(x^{(k)}\right)$$

In the exact Newton method, the matrix $D^{(k)}$ is the inverse Hessian. Quasi-Newton methods avoid calculating the Hessian and inverting it by updating an approximation of the inverse Hessian using the change in the gradients. For a quadratic function, the change in gradient is described by:

$$q^{(k)} = \left(\nabla^2 f\right) p^{(k)}$$

where $p^{(k)} = x^{(k+1)} - x^{(k)}$ and $q^{(k)} = \nabla f\left(x^{(k+1)}\right) - \nabla f\left(x^{(k)}\right)$. We therefore seek an approximation $D^{(k)}$ to the inverse Hessian that approximately satisfies:

$$p^{(k)} \approx D q^{(k)}$$

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method updates $D^{(k)}$ by making the smallest change, under some specific weighted norm, that agrees with the latest change in the gradient:

$$D^{(k+1)} = \arg \min_{p^{(k)} = D q^{(k)}} \left\| W^{1/2} \left(D - D^{(k)}\right) W^{\frac{1}{2}} \right\|_F \tag{8}$$

where $\|A\|_F = \sqrt{\sum_{ij} A_{ij}^2}$ is the Frobenius norm and W is any matrix such that $q^{(k)} = W p^{(k)}$.

5.1

Show that the solution of equation (8) is given by:

$$D^{(k+1)} = D^{(k)} + \frac{p^{(k)} \left(p^{(k)}\right)^T}{\left(p^{(k)}\right)^T q^{(k)}} - \frac{D^{(k)} q^{(k)} \left(q^{(k)}\right)^T D^{(k)}}{\left(q^{(k)}\right)^T D^{(k)} q^{(k)}} + \tau^{(k)} v^{(k)} \left(v^{(k)}\right)^T \tag{9}$$

where $\tau^{(k)} = \left(q^{(k)}\right)^T D^{(k)} q^{(k)}$, and:

$$v^{(k)} = \frac{p^{(k)}}{\left(p^{(k)}\right)^T q^{(k)}} - \frac{D^{(k)} q^{(k)}}{\tau^{(k)}}$$

The BFGS method is therefore given by (ignoring the stopping condition):

```
1. Start from some $x^{(0)}$ and an initial $D^{(0)}$
2. For $i \in \{0, 1, 2, \ldots\}$
3.     $\Delta x^{(i)} \leftarrow -D^{(i)} \nabla f\left(x^{(i)}\right)$
4.     $t^{(i)} \leftarrow \arg\min_t f\left(x^{(i)} + t\Delta x^{(i)}\right)$
5.     $x^{(i+1)} \leftarrow x^{(i)} + t^{(i)} \Delta x^{(i)}$
6.     Calculate $D^{(i+1)}$ according to equation (9)
```

5.2

We now consider applying BFGS to a quadratic objective $f(x) = \frac{1}{2}x'Hx - b'x$ with $x \in \mathbb{R}^n$ and H positive definite.

5.2.1

Show that for all $i < k \leq n$ we have $D^{(k)} q^{(i)} = p^{(i)}$. That is, for a quadratic objective, the approximate inverse Hessian matches all the changes in the gradient so far. Conclude that $D^{(n)} = H^{-1}$, i.e. after n iterations the correct Hessian is recovered.

5.2.2

Show that $\Delta x^{(0)}, \ldots, \Delta x^{(n-1)}$ are H-conjugate.

5.2.3

Show that with $D^{(0)} = I$, the sequence of iterates $x^{(i)}$ generated by BFGS is identical to those generated by the conjugate gradient method described above. It is important to note that this holds only for a quadratic objective, and when exact line search is used. For non-quadratic objectives, or when approximate line search is used, the two methods typically differ.