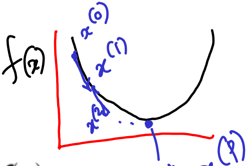
First Order Descent Methods

Instructor: Prof. Ganesh Ramakrishnan

General descent algorithm



for some p is ideal

$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$

- Let us say we want to minimize a function $f(x)$
- The general descent algorithm involves two steps:
  - Determining a good descent direction $\Delta x^{(k)}$, typically forced to have unit norm
  - Determining the step size using some line search technique $\rightarrow t^{(k)}$
- We want that $f(x^{(k+1)}) < f(x^{(k)})$ $\rightarrow$ imp for convex fns

For non-convex may not insist that $f(x^{k+1}) < f(x^k)$

- If the function f is convex, we must have $\nabla^\top f(x^{(k)})(x^{(k+1)} - x^{(k)}) < 0$
- That is, the descent direction $\Delta x^{(k)}$ must make an obtuse angle with the gradient vector $\nabla f(x^{(k)})$

$f(x^{k+1}) \geq f(x^k) + \underbrace{\nabla^T f(x^k) t^k \Delta x^k}$ & $f(x^{k+1}) < f(x^k)$

since $t^k > 0$ then $\nabla^T f(x^k) \Delta x^k < 0$ (is necessary)

General descent algorithm

- In descent for a convex function f, we must have:

$$f(x^{(k+1)}) \geq f(x^{(k)}) + \nabla^\top f(x^{(k)})(x^{(k+1)} - x^{(k)})$$

  Here, the LHS is the actual value and RHS is the linear approximation of $f(x^{(k+1)})$
- Since step size $t^{(k)} > 0$,
  $\nabla^\top f(x^{(k)})\Delta x^{(k)} < 0$ $\rightarrow$ only a necessary condition. Not sufficient! $\therefore$ You need to estimate $t^{(k)}$ to avoid overstepping
- Algorithm:
  1. Set a starting point $x^{(0)}$
  2. repeat
     1. Determine $\Delta x^{(k)}$
     2. Choose a step size $t^{(k)} > 0$ using line search
     3. Obtain $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$
     4. Set $k \leftarrow k + 1$

  until stopping criterion (such as $\left\|\nabla f(x^{(k+1)})\right\| < \epsilon$) is satisfied

Steepest descent

- The idea of steepest descent is to determine a descent direction such that for a unit step in that direction, the prediction of decrease in the objective is maximized → $\nabla^T f(x^k)\, \Delta x^k$ as negative as possible
- However, consider $\Delta x = \text{argmin}_v \underbrace{\begin{bmatrix} -5 & 10 & 15 \end{bmatrix}}_{\nabla f(x^k)} v$

$$\implies \Delta x = \begin{bmatrix} \infty \\ -\infty \\ -\infty \end{bmatrix}$$

which is unacceptable

for diff choices of norm, you get diff Δx^k

- Thus, there is a necessity to restrict the norm of v
- The choice of the descent direction can be stated as:

$$\Delta x = \underset{v}{\text{argmin}}\, \nabla^\top f(x) v$$

s.t. $||v|| = 1$

Various choices of the norm result in different solutions for Δx

- For 2-norm, $\Delta x = -\dfrac{\nabla f(x^{(k)})}{\left\|\nabla f(x^{(k)})\right\|_2}$

  (gradient descent)

- For 1-norm, $\Delta x = -\operatorname{sign}\left(\dfrac{\partial f(x^{(k)})}{\partial x_i^{(k)}}\right) e_i$, where e_i is the ith standard basis vector

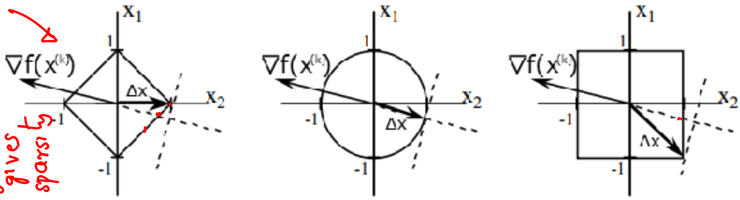
  (coordinate descent)

  $i = \underset{j}{\operatorname{argmax}} \left| \dfrac{\partial f(x^k)}{\partial x_j^k} \right|.$

- For ∞-norm, $\Delta x = -\operatorname{sign}(\nabla f(x^{(k)}))$

H/W what abt $\Delta x = \underset{\|v\|_p = 1}{\operatorname{argmin}}$ of $\nabla^T f(x^k) v$

Also gives further intuition why $\| \, \|_1$ gives sparsity



SDD in L1-norm   SDD in L2-norm   SDD in L∞ -norm

Steepest descent direction

Gradient Descent

Interpretation of gradient descent

$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{k})$ $\left(\therefore \Delta x^k = -\nabla f(x^k) \right)$

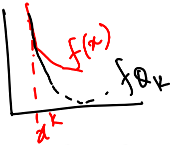
- Consider the optimization problem

$t^{(k)}$ absorbs $\dfrac{1}{\|\nabla f(x^k)\|}$

$$x^* = \arg\min_{x \in \mathbf{R}^n} f(x)$$

- The idea behind gradient descent is that you start with a $x^0 \in \mathbf{R}^n$, and $\forall k = 0, 1, 2, \ldots,$

$$x^{k+1} = x^k + t^k \Delta x^k$$



- x^{k+1} can be treated as a solution to a quadratic approximation of f around x^k

the optimal soln

convex

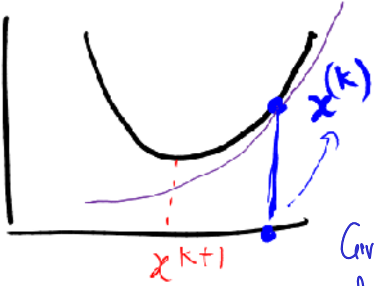
$$f_{Q_k}(x) = f(x^k) + \nabla^T f(x^k)(x - x^k) + \alpha \|x - x^k\|^2$$

min should be at x^{k+1}

- At each iteration, we can consider the quadratic approximation

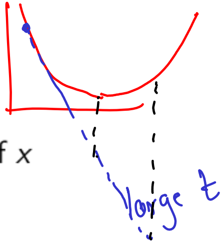
$$f_{Q_k}(x^{k+1}) = f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{1}{\underbrace{2t}_{\alpha}} \left\| x^{k+1} - x^k \right\|^2$$

- Equating $\nabla f_{Q_k}(x^{k+1}) = 0$
$\implies \nabla f(x^k) + \frac{1}{t}(x^{k+1} - x^k) = 0$
$\implies x^{k+1} = x^k - t \nabla f(x^k)$



Gives An interpretation of gradient descent

Finding the step size t



- If t is too large, we get diverging updates of x
- If t is too small, we get a very slow descent
- We need to find a t that is *just right*
- We discuss two ways of finding t:
  1. Exact line search
  2. Backtracking line search

$$t^k = \underset{t \geq 0}{\operatorname{argmin}}\; f(x^k + t\Delta x^k)$$

For any descent algo

β & α are provided
$\alpha = 1/2$

We already to make $\nabla^T f(x^k)\Delta x^k$ as negative as possible.
$t=1$ & $t \to \beta t$ $(\beta \in (0,1))$ until $f(x^{k+1}) < f(x^k) - t\alpha \nabla^T f(x^k) \Delta x^k$

Exact line search

$$t^{k+1} = \underset{t}{\operatorname{argmin}}\, f\left(x^k - t\nabla f(x^k)\right)$$

$$= \underset{t}{\operatorname{argmin}}\, \phi(t)$$

- This method gives the most optimal step size in the given descent direction $\nabla f(x^k)$
- It ensures that $f(x^{k+1}) \leq f(x^k)$
- If f is itself quadratic, it gives an optimal solution to the minimization of f (since the quadratic approximation f_Q would become exact and no longer approximate)

Backtracking line search

Q: Why is it that if t violates (a) then $\beta^{r} t$ has better chances of satisfying (a)?

- The algorithm
  - Choose a $\beta \in (0, 1)$
  - Start with $t = 1$
  - While $f\left(x^k - t\nabla f(x^k)\right) > f(x^k) - \frac{t}{2}\left\|\nabla f(x^k)\right\|^2$, do
    - Update $t \leftarrow \beta t$

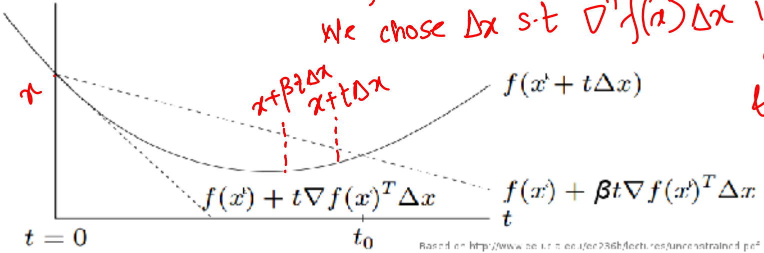
For grad descent $\Delta x^k = -\nabla f(x^k)$
& $\alpha = 1/2$

Repeat until
$f(x^k - t\nabla f(x^k)) \leq f(x^k) - \frac{t}{2}\|\nabla f(x^k)\|^2 \rightarrow$ (a)

Interpretation of backtracking line search

$\nabla^T f(x)\Delta x < 0$ is necessary & not suff

We chose Δx s.t $\nabla^T f(x)\Delta x$ is as negative as possible & might have overstepped



Based on http://www.ee.ucla.edu/ee236b/lectures/unconstrained.pdf

- Δx = direction of descent = $-\nabla f(x^k)$ for gradient descent
- A different way of understanding the varying step size with β: Multiplying t by β causes the interpolation to tilt as indicated in the figure

Assumptions for proving the convergence of gradient descent

- $f : \mathbf{R}^n \to \mathbf{R}$ is convex and differentiable
- f is Lipschitz continuous

- **Claim:** If $t^k \le \frac{1}{L}$, then

$$f(x^k) - f(x^*) \le \frac{\left\|x^0 - x^*\right\|^2}{2tk}$$

  - The gap between the optimal solution and the solution at the $\mathbf{kth}$ step is going to decrease with increasing step size t
  - $O(\frac{1}{k})$ rate or linear convergence