8.1 First Order Methods

Unconstrained optimization

$$\min_x f(x) \tag{8.1}$$

assume x is convex and differentiable

Gradient descent

- Choose $x^{(0)}$
- Iterate $x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k)})$
- Stop sometime

Why?
(Taylor expansion)

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} H ||y - x||_2^2 \tag{8.2}$$

Gradient descent says replace with -

$$\frac{1}{2}\frac{1}{t} I ||y - x||_2^2 \tag{8.3}$$

Choose $y = X^+$ to maximize -

$$x^+ = x - t\nabla f(x) \tag{8.4}$$

What to use t_k for? Need t to make it work Fixed (for all iterations) only works if t is exactly right Usually does not work

Sequence

$$t_k \quad s.t. \sum_{k=1}^{\infty} t_k = \infty, \quad \sum_{k=1}^{\infty} t_k^2 < \infty \tag{8.5}$$

Back tracking line search
At each iteration choose best t
1. Set $0 < \beta < 1, 0 < \alpha < \frac{1}{2}$
t_{init}

2. At each k
While $f(x^k - t\nabla f(x^k)) > f(x^k) - \alpha t ||f(x^k)||_2^2$ (To get strong convex term)
set $t = \beta t$(shrink t)
$x^T = x - t\nabla f(x)$ approximation to fixed line search

Could solve (at each k)

$$t = \underset{s>=0}{\operatorname{argmin}} f(x^{(k)} - sf(x^{(k-1)})) \tag{8.6}$$

(this equation is usually not solvable)
Exact line search

Theorem 8.1. *If f is convex and differentiable*

$$\begin{aligned} ||\nabla f(x) - \nabla f(y)||_2 \leq L||x-y||_2 \\ (Lipschitz) \\ if, t(fixed) \leq \frac{1}{2} \end{aligned} \tag{8.7}$$

GD satisfies

$$f(x^{(k)}) - f^{(*)} <= \frac{||x^{(0)} - x^k||_2^2}{2tk} \tag{8.8}$$

convergence rate $O(\frac{1}{k})$

more accurate we want the slower we get. this was with week condition

Theorem 8.2. *If f strongly convex with previous conditions*

$$f(x^{(k)}) - f^{(*)} <= C^k \frac{L}{2}||x^{(0)} - x(k)||_2^2 \tag{8.9}$$

(if strong convexity, convergence is exponentially fast)

Example 8.3.

$$f(x) = \frac{1}{2}||b - Ax||_2^2 \tag{8.10}$$

if we want Lipschitz ∇

$$\nabla^2 f = A^T A \leq LI = \sigma_{max}^2(A)I \tag{8.11}$$

(as long as larger singular value)

if we want strong convexity.

$$\nabla^2 f \geq mI = \sigma_{min}^2(A)I \tag{8.12}$$

(minimum eigenvalue) need A *as full rank*

c depends on $\frac{1}{m}$, *if* $m < 0$ *then will be slow*

stop if $||\nabla f(x^{k-1})||_2$*is small*

Stochastic Gradient descent (SGD)
- Objective function as sum of individual function, GD becomes

$$
\begin{aligned}
\min_x \sum_{i=1}^{m} f_i(x) \\
x_{(k)} = x_{(k-1)} - tk \sum_{i=1}^{m} \nabla f_i(x^(k-1)) \\
x_{(k)} = x_{(k-1)} - tk \nabla f_{ik}(x^(k-1)) \\
i_k \in 1, ..., m
\end{aligned} \tag{8.13}
$$

Pick i_k?
usually $i_k = 1, ..., m, 1, ..., m,$ (cyclic option)
OR
$i_k \ unif(1, m)$
SGD works best when you are far from the optimum not when you are close to optimum

Good idea?
- Large datasets
- Curvature is flat
Norm is small but you may not be at the optimal

Subgradient descent (f not differentiable)

$$
\begin{aligned}
X^k = x^{k-1} - t_k g^{(k-1)} \\
g^{(k-1)} \in \delta f(x^{(k-1)})
\end{aligned} \tag{8.14}
$$

not a descent method
not guaranteed x^k is better than x^{k-1}, keep track of current value
t_k fixed or square summable ($\frac{t_0}{k}$)

Theorem 8.4. *If f is Lipschitz, rate* $O(\frac{1}{\sqrt{k}})$
It is worst than gradient descent

Example 8.5. $\min_\beta \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta_i||$

subdifferential

$$
\begin{aligned}
g(\beta) = -X^T(y - X\beta) + \lambda\delta||\beta||_1 \\
-X^T(y - X\beta) + \lambda v \\
v_i = \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}
\end{aligned} \tag{8.15}
$$

So $sign(\beta) \in \delta||\beta||_1$

From KKT (stationarity)

$$
\begin{aligned}
\lambda v = X^T(y - X\beta) \\
X_i^T(y - X\beta) = \lambda v_i \\
\lambda sign(\beta_i), \quad iff \quad B_i \neq 0 \\
X_i^T(y - X\beta) <= \lambda, \quad iff \quad \beta_i = 0
\end{aligned} \tag{8.16}
$$

This gives LARS algorithm

Proximal gradient descent Decomposable f(x) = g(x) + h(x)

g is convex, differentiable (do GD on g only) h is convex only

$$x^+ = \underset{Z}{\operatorname{argmin}}\, g(x) + \nabla g(x)_T(z-x) + \frac{1}{2t}||z-x||_2^2 + h(2) \tag{8.17}$$

approximate the hessian,

$$\begin{aligned} &= \underset{Z}{\operatorname{argmin}} + \frac{1}{2t}||z-(x-t\nabla g(x))||_2^2 + h(2) \\ prox_t(x) &:= \underset{Z}{\operatorname{argmin}}\, \frac{1}{2t}||x-z||_2^2 + h(Z) \end{aligned} \tag{8.18}$$

only depends on h not y

$$x^(k) = prox_{tk}(x^(k-1) - t_k\nabla g(x^(k-1))) \tag{8.19}$$

Example 8.6. *LASSO (ISTA) - Iterative soft thresholding for L_1 norm*

$$\begin{aligned} \min_{\beta} &= \frac{1}{2}||y-X\beta||_2^2 - \lambda||\beta||_1 \\ prox_t(\beta) &= \underset{Z}{\operatorname{argmin}}\, \frac{1}{2t}||\beta - Z||_2^2 + \lambda||Z||_1 \\ &= S_{\lambda t}\beta \\ &(soft \quad thresholding) \\ S_\tau(\beta) &= \begin{cases} \beta_i - \tau & if\ \beta_i > \tau \\ 0 & if\ -\tau \le \beta_i \le \tau \\ \beta_i + \tau & if\ \beta_i < -\tau \end{cases} \\ \beta_+ &= S_{\lambda t}(\beta + tX^T(y - X\beta)) \end{aligned} \tag{8.20}$$

Projected gradient descent:

$$\begin{aligned} &min_{x\in C} f(x) \\ take, \quad g(x) &= f(x) \\ h(x) &= I_c(x) \\ prox(x) &= \underset{Z\in C}{\operatorname{argmin}}\, ||x-Z||_2^2 \\ x^+ &= P_c(x - t\nabla g(x)) \end{aligned} \tag{8.21}$$

Where P_c project onto C

All algorithms discusses here also have accelerator versions. Use information from previous updates

8.2 Second Order Methods

Good option if you can take 2 derivatives
1. Newton's method - f(x) 2x differentiation

$$x^+ = x - (\nabla_2 f(x))_{-1} \nabla f(x) \tag{8.22}$$

Hessian scales with the problem,

$$\nabla^2 f(x) \approx \frac{1}{t} I \tag{8.23}$$

may not converge

Damped Newton

$$X^+ = x - t(\nabla^2 f(x))^{-1} \nabla f(x) \tag{8.24}$$

use backtracking to get t

Theorem 8.7. *f convex 2x differentiation, Strongly convex*

$$\begin{aligned} \nabla f &\quad \textit{Lipschitz} \\ \nabla^2 f &\quad \textit{Lipschitz} \\ f(x) - f^* = &\begin{cases} f(x^0) - f^* - \gamma k & \textit{if } k \leq k_0 \\ C(\frac{1}{2})^(2^(k - k_0 + 1)) & \textit{if } k > k_0 \end{cases} \end{aligned} \tag{8.25}$$

8.2.1 Other Methods

Barrier Method
Primal Dual Interior Point method
BFGS and Quasi Newton