|  |  |  |  |
| --- | --- | --- | --- |
| Optimization | Linear Regression: (empirical loss function) | | Criterion: fit existing data (**a**­1, b1­), ..., (**a**n­, bn) |
| SVM: for binary classification: F(**a**) = sgn(**x**T**a**),  For model to fit data: g(**a**i) = < 0 for i s.t. bi = -1  g(**a**i) = > 0 for i s.t. bi = 1  By using = c**x**, c = , g(**a**i) = ≤ -1 for i s.t. bi = -1,  g(**a**i) = ≥ 1 for i s.t. bi = 1  For criterion 2: want to maximise (margin of error) btw classes  minimizing minimizing  So SVM problem is: s.t. ≥ 1 i = 1,..., n | | Criterion 1: fit existing data (**a**­1, b1­), ..., (**a**n­, bn)  Criterion 2: largest margin between classes  Note = **a**T**x**  Can combine 2 statements into  ≥ 1 i = 1,..., n  And model is : F(**a**) = sgn(**x**T**a**) |
| Notation | **x** = [x1; x2; ...; xd] = | | **x** is column vector by default  For shorter vectors, can write [x;y], [x;y;z] |
| Linear Algebra | l2 norm2 =  measures the geomtric dist btw **x** and **y** | Inner product = **x**T**y** =  Cauchy-Schwarz Inequality: – ≤ **x**T**y** ≤ | |
| General Formulation | Minimize/Maximize f(**x**)  s..t gi(**x**) = 0, i = 1,..., m (equality constraint)  hi(**x**) ≤ 0, j = 1,..., p (inequality constraint) | | Or min f(**x**) s.t. **g**(**x**) = 0, **h**(**x**) ≤ 0  **h**(**x**) ≤ 0 : hi(**x**) ≤ 0, i = 1,...,p  Note max f(**x**) min - f(**x**) |
| Feasible Set | Solutions that satisfy the constraints: feasible. Set of soln = S = . So solve min f(**x**) s.t. **x** S  When there is no constraint, i.e. S = , this is an unconstrained problem. Otherwise: infeasible | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Optimization | | Mtds to solve optimization. Finding a lower bound. Stationary point method. Convex sets. Convex functions  In a minimization prob, min f(**x**), **x** S, **x**\* S is an optimal soln if for any **x** S, f(**x**\*) ≤ f(**x**).  **x**\* = . f(**x**\*) is called the optimal value. (possible that there is no optimal soln) | | | | | | |
| 1. Finding a lower bound | | Derive a lower bound for f(**x**) w **x** S  Show f(**x**\*) reaches the lower bound  Problems: not easy to derive lower bound, cannot be used for large scale problems, not systematic | | | E.g. min f(**x**) = -x1x2, s.t. x1 + x2 = 24, x1 ≥ 0, x2 ≥ 0  Observe x2 = 24 - x1. f(**x**) = -x1(24 - x1) = (x1 - 12)2 - 144  (x1 - 12)2 ≥ 0. f(**x**) ≥ -144 (lower bound)  , f(**x**\*) = -144 | | | |
| 2. Stationary pt mtd | | Problems: cannot be sure it works  May give infeasible solutions | Using f(**x**) = -x1(24 - x1), f'(x1) = 2x1 - 24, 0 ≤ x1 ≤ 24. f'(x1) = 0 x1 = 12  But if use f(**x**) = -x1x2, then , = [0;0] x1 = 0 = x2 | | | | | |
| Convex set | A set D is convex if for any 2 pts **x** and **y** in D, the line segment joining **x** and **y** also lies in D. i.e. **x**, **y** D **x** + (1-)**y** D  Notation: line segment [**x**, **y**] := {**x** + (1-)**y** D } | | | | | | Cauchy-Schwarz: **x**T**y** ≤ | |
| E.g. Half Space: Given **a** , b , H = {**x** |**a**T**x** ≤ b} is a convex set  **x** H **a**T**x** ≤ b. **y** H (1 - )**a**T**y** ≤ (1 - )b  **x** + (1-)**y** = **a**T(**x** + (1 - )**y**) ≤ b H is a convex set | | | | | |
| E.g. d-Ball, given a center **a** , has radius b > 0  B(**a**, b) = { **x** | ≤ b } is convex  **x** B ≤ b. **y** H (1 - ) ≤ (1 - )b        B is a convex set | | | | | |
| Proposition 3.1: If C1, C2, ..., Cm are convex sets in , then C = is also convex | | | | | |  | |
| E.g. Polyhedron or polyhedron set {**x** |**A**T**x** ≤ **b**} = D = {**x** ≤ b1, **x** ≤ b2, ..., **x** ≤ bn} = (intersection of half space) D is convex  **A** = , **b** = , **Ax** = | | | | | |
| Convex fn | Let D be a convex set. Consider a fn f: D R.  f is convex if f(**x** + (1-)**y**) ≤ f(**x**) + (1-)f(**y**), **x**, **y** D,  f is strictly convex if f(**x** + (1-)**y**) < f(**x**) + (1-)f(**y**), distinct **x**, **y** D,  f is concave if f(**x** + (1-)**y**) ≥ f(**x**) + (1-)f(**y**), **x**, **y** D,  f is strictly concave if f(**x** + (1-)**y**) > f(**x**) + (1-)f(**y**), distinct **x**, **y** D, | | | | | Diagram  Description automatically generated Chart  Description automatically generated  line segment lies above; below | | |
| Linear fn: f(**x**) = **a**T**x** + b, **a**, **x**  f(**x** + (1-)**y**) = **a**T(**x** + (1-)**y**) + b = (**a**T**x** + b) + (1-)(**a**T**y** + b) = f(**x**) + (1-)f(**y**)  f is convex. Note f is also concave | | | | |
| l2 norm to a pt a, f(**x**) =  f(**x** + (1-)**y**) = = ≤ + (1-) = f(**x**) + (1-)f(**y**) f is convex | | | | | triangle inequality: | | |
| Square fn: f(**x**) = **x**2 is convex (proof using f(**x** + (1-)**y**) - {f(**x**) + (1-)f(**y**)} ≤ 0) | | | | |  | | |
| Max of coordinates: f(**x**) = max{x1, ..., xn}  f(**x** + (1-)**y**) = max{**x1** + (1-)**y**1, ..., **x**n + (1-)**y**n}. Suppose **x**i + (1-)**y**i is the max, then f(**x** + (1-)**y**) = **x**i + (1-)**y**i  Since xi ≤ max{x1, ..., xn} = f(**x**), yi ≤ max{y1, ..., yn} = f(**y**). Then f(**x** + (1-)**y**) = **x**i + (1-)**y**i ≤ f(**x**) + (1-)f(**y**) f is convex | | | | | | | |
| Corollary 3.1: Let f1, ..., fk: D R be convex fns on a convex set D . Then f(**x**) = , where ai ≥ 0 i, is also a convex fn on D  If at least 1 of f­i is strictly convex on D, then f is strictly convex on D (i.e. LC of convex fns is convex) | | | | | | | |
| R/s btw sets & fns | | Proposition 3.5: Suppose D is convex. If f: D R is convex, then for any a , the sublevel set Sa := {**x** D| f(**x**) ≤ a} is convex  Level set: f(**x**) = a. Superlevel set: f(**x**) ≥ a | | | | | | |
| Half space: H = {**x** |**a**T**x** ≤ b}. Consider f(x) = **a**T**x**, a linear fn which is convex. H is a sublevel set, hence is convex | | d-Ball: B(**a**, b) = { **x** | ≤ b }. Consider f(**x**) = , which is convex. B is a sublevel set of f, hence is convex | | | | |
| S = {**x** |x2 ≥ 5, - 4x2 ≤ 7} = {x2 ≥ 5} { - 4x2 ≤ 7}  {x2 ≥ 5} = {xT[0,1] ≥ 5} = {xT[0,-1] ≤ -5} is a half space, hence is convex  { - 4x2 ≤ 7} is the sublevel set of { - 4x2} = LC of sq fn and linear fn, hence is convex  So S is convex | | | | | | Corollary 3.1 + Proposition 3.5  Proposition 3.1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Gradient, convexity and optimum | Suppose f: S is differentiable at **x**, then the gradient vector of f at **x** is the col vector  If f has cts gradient on S, we say it is C1 on S  Taylor's Theorem, 1st order: Suppose f: S is C1. Suppose [**x**, **y**] S, then **w** [**x**, **y**] s.t. f(**y**) = f(**x**) + (**y** - **x**)  Suppose **d** is non-zero, = 1, **d** is a direction vector. Then **x** + is moving from **x** in dir **d** and is rate of change in f when **x** moves from **w** along dir **d**.  f(**x** + ) = f(**x**) + = f(**x**) + . So **.** Hence = (since **w** **x** when )  At **x**, the vector is a normal vector to the level set. | | |
| f(**x**) = x1 - .  Linear fn: f(**x**) = **b**T**x**. **b** (can write out individual elems and verify) | | Quadratic fn: f(**x**) = . Suppose A is symmetric  **Ax** (list out elems and combine OR use defn to get = (**Ax**)T**d**) |
| Stationary pt and optimality | Thrm 3.1: Suppose that f(**x**) is C1 on an open convex set S in . Then  (a) the fn f is convex iff f(**x**) + (**y** - **x**) ≤ f(**y**) **x**, **y** S  (b) the fn f is strictly convex iff f(**x**) + (**y** - **x**) < f(**y**) **x** ≠ **y** S  To prove (a), use RHS to get defn of convex fn & use d = y - x in defn of + convex fn defn  Suppose f is C1, then optimal soln stationary pt (so can try to find stationary pts first). If f is convex stationary pts are optimal. If f is strictly convex, optimal soln is unique. | | |
| Corollary (Sufficiency): Suppose f(**x**) is C1 and convex on an open convex set S in . Suppose = **0**. Then **x**\* is an optimal soln for , **x** S. Suppose in addition that f(**x**) is strictly convex on S, then **x**\* is the unique optimal soln. | | |
| Thrm (Necessity): Suppose that f(**x**) is C1 on an open set S in and **x**\* is its min or max, then = 0 (note convexity not needed) | | |
| Positive definite matrices | If A is real symmetric, is an eigenvalue of A w eigenvector **v**i if A**v**i = **v**i | To find eigenvalues: find roots to det(I - A) = 0 | |
| Let A be a real symmetric d x d matrix.  (a) A is positive semidefinite (PSD) if **x**TA**x** ≥ 0, **x** . Notation: A 0  (b) A is positive definite (PD) if **x**TA**x** > 0, **x** ≠ 0 . Notation: A 0  (c) A is negative semidefinite if **x**TA**x** ≤ 0, **x** . i.e. -A is PSD  (d) A is negative definite if **x**TA**x** < 0, **x** ≠ 0 . i.e. -A is PD  (e) A is indefinite if A is neither positive nor negative semidefinite | Thrm 2.3: Let A be a real symmetric d x d matrix  (a) A is PSD iff every eigenvalue of A is nonnegative ( ≥ 0)  (b) A is PD iff every eigenvalue of A is positive ( > 0)  (c) A is NSD iff every eigenvalue of A is nonpositive ( ≤ 0)  (d) A is ND iff every eigenvalue of A is negative ( < 0)  (e) A is indefinite iff there is a +ve and -ve eigenvalue of A | |
| Let A . The kth principal minor of A is the determinant of the kth principal submatrix of A, i.e. = det[a11], = det ...  Thrm: Suppose A is a symmetric matrix, then A is PD iff > 0 for all k = 1, ..., d | | |
| Hessian and convexity | The Hessian of f at **x** is the d x d matrix,  If f has cts second order derivatives, we write f C2, then Hessiam matrix is symmetric  Thrm 2.5: Suppose f: S is C2. Suppose line segment [**x**, **y**] S, then **w** [**x**, **y**] s.t. f(**y**) = f(**x**) + (**y** - **x**) +  So if **x** is a stationary pt (i.e. = 0), and ≥ 0 (i.e. PSD), then **x** is the min  Thrm 3.2: Suppose f(**x**) is C2 on an open convex set D in .  - The fn f is convex on D iff the Hessian matrix is PSD at each **x** D  - If Hessian matrix is PD at each **x** D, then f is strictly convex on D  - The fn D is concave on D iff the Hessian matrix is NSD at each **x** D  - If Hessian matrix is ND at each **x** D, then f is strictly concave on D  - If Hessian matrix is indefinite at some pt **x** D, then f is not a convex nor a concave fn on D (i.e. **x** where have +ve & -ve ) | | |
| E.g. | S = {**x** : > 1/3, + 2 + ≤ 10}  Consider f(**x**) = + 2x1x2 + , ,  = 6, = 12 - 4 = 4(3 - 1)  Let D = {**x**: > 1/3} (convex set as half space). Then and both > 0 is PD f(**x**) is strictly convex (thrm 3.2)  By Proposition 3.5: S10 = {**x** : f(**x**) ≤ 10, **x** D} is a convex set | | |
| Summary | Solve . 1) Find stationary pt **x**\* of f. 2) If f is convex, **x**\* is a minimizer. 3) If 0 **x**, f is convex. 4) To check PSD: eigenvalues/principal minors | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Local Optimum | | A point **x**\* S is a local minimzer of f(**x**) if an > 0 s.t. f(**x**) ≥ f(**x**\*) **x** S  = {**x**: < }, i.e. the ball centered at **x**\* w radius  A point **x**\* S is a global minimizer of f(**x**) if an > 0 s.t. f(**x**) ≥ f(**x**\*) **x** S. (global min is also a local min)  A stationary pt **x**\* of f which is neither a local minimizer nor a local maximizer is a saddle point | | | |
| Thrm 4.2: Let X be an open subset of Rn. Suppose f: X R is C1 in X. If **x**\* X is a local minimizer of f on X, then **x**\* is a stationary pt, i.e. = **0**. In addition, if f is C2, then must be PSD | | | |
| Thrm 4.3: Let X be an open subset of Rn. Suppose f: X R is C2. If **x**\* X is a stationary pt, i.e. = **0**, and is PD, then **x**\* is a local minimizer (if is ND, then **x**\* is a local max) | | | |
| So given a stationary pt **x**\*, check Hessian : 1) PD/all eigenvalues > 0 **x**\* is local min. 2) ND/all eigenvalues < 0 **x**\* is local max  3) Indefinite/eigenvalues > 0 & < 0 **x**\* is a saddle pt. 4) PSD or NSD inconclusive | | | |
| page18image35460176Solve .  Case 1: f is convex  - local min = global min  Case 2: f is not convex  - global min also local min  - they are stationary points w PSD Hessian  - assuming they exist, we can go through all local mins  1) Find stationary pt **x**\* of f. 2) If f is convex, **x**\* is a minimizer. 3) If 0 **x**, f is convex. 4) To check PSD: eigenvalues/principal minors | | | |
| E.g. | | f(x) = x3. f'(x) = 3x2. f''(x) = 6x. f'(x) = 0 x\* = 0. f''(x\*) = 0, is only PSD.  For any > 0, - (since ball centered at 0), f(-) = < f(0) 0 is not a local min  Also,  , f() = > f(0) 0 is not a local max  0 is neither a local min or local max 0 is a saddle pt | | | |
| Solve min f(**x**) = , assuming there is a soln  . x\* = (0,1) and x\* = (, 1). f(0, 1) = -3. f(, 1) ≈ -11.543 < -3 (, 1) is the optimal soln | | | |
| Bounded set and Weierstrass theorem (NOT TESTED) | | Let S be a nonempty set. The set S is said to be bounded if there is a positive num M s.t. ≤ M **x** S. If S is both bounded and closed, it is compact.  - A set S is bounded iff there is a positive num M s.t. |xi| ≤ M, i = 1,2,...,n **x** S  - Every seq in a compact set S has a converging subsequence w its limit in S  Weierstrass extreme-value Thrm 2.6: A cts fn on a nonempty compact set S has a global max pt and a global min pt in S  A cts fn f: is coercive if  Thrm 4.1: A coercive cts fn on a nonempty set S has a global min pt in S | | | |
| Iterative Algos | To find stationary pt **x**\*: (**x**\*) = **0**  Mtd 1: Generate iterates **x**1, **x**2, ..., **x**n **x**\*  Mtd 2: Generate decreasing sets C1 C2 C3 ... Cn **x**\* | | Stop algo at: 1) Budget: stop at k = T iteration, T is given  2) Tolerance: stop at ≤ Tol  3) Improvement: stop if ≤ Tol or f(**x**k) - f(**x**k+1) ≤ Tol  Or use a combination of any | | |
| To assess algo: 1) Complexity (how many iteration needed). 2) Implementation (easy to code). 3) Update cost (cost of running 1 step of algo) | | | | |
| Bisection mtd (Mtd 2) | | Suppose f' exists and is cts. Want to find f'(x) = 0. Suppose f'(a) and f'(b) have diff signs  Then there is an x (a, b) s.t. f'(x) = 0. So we decrease intervals w x\* inside  Evaluate f' at midpoint c = (a+b)/2. If f'(c) = 0: job done.  Else: pick half interval w boundary value of opp signs  Input: a0, b0, f', Tol. Output: ct so f'(ct) ≤ Tol  Interval length at next iteration is half. Accuracy of using midpoint: at least half of interval  |ak - bk| = 2-k|a0 - b0|. OR log|ak - bk| = -k log 2 + log|a0 - b0|  Linear convergence to x\*: log error decreases linearly | | | t = 0  while |f'(ct)| > Tol do  ct = 0.5at + 0.5bt;  at+1 = at, bt+1 = bt;  if f'(ct)f'(at) < 0 then  bt+1 = ct;  else  at+1 = ct;  t = t+1 |
| Golden Ratio (Mtd 2) | | Suppose f is unimodal on [a,b]. So f is decreasing on [a, **x**\*] and increasing on [b, **x**\*]  Consider 2 intermediate points at a time. Try to keep **x**\* inside  Choose 2 pts, < . If f() < f(): **x**\* [a1, ]. If () ≥ f(): **x**\* [, b1].  Choice of and can be arbitrary, but using golden ratio save computation cost  We want . So .  Solving, = 0.382. 1 - = 0.618 = golden ratio (case for f() > f(); same for other)  Intervals shrink = 0.618 of previous interval.  General steps: 1) Find left and right boundary, so f is unimodal inside  2) Find left and right intervals of proportion  3) Choose the one with larger "interior" boundary value  4) Repeat 2 and 3 until the interval length is lower than Tol  Length of kth iteration = |ak - bk| = k|a0 - b0| = 0.618k of original len  OR log|ak - bk| = klog + log|a0 - b0|So linear convergence to x\*. | | Input: a0, b0, f, Tol  Output: (at + bt)/2 as an approximate soln  t = 0;  while bt - at > TOL do  = 0.618ai + 0.382bi;  = 0.382ai + 0.618bi;  [ai+1, bi+1] = [ai, bi];  if f() > f() then  bi+1 = ;  else  ai+1 = ;  t = i+1  Note f need not be convex  Accuracy: 0.618k+1 \* original length ≤ error  OR 0.5\*0.618k \* original length ≤ error | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Univariate Newton's mtd (Mtd 1) | Suppose f' and f'' are available. Then  Thrm: Suppose |x0 - x\*| is sufficiently small and f is C3 with f''(x\*) ≠ 0, then for a constant M, |xk+1 - x\*| ≤ M|xk - x\*|2, k (proof in lecture 5)  Therefore, error of xk+1 is square of error of xk quadratic convergence  And  So digit of accuracy doubles w each iteration (if error of 1st = 10-2, then error of 2nd = 10-4, error of 3rd = 10-8, ...)  Trying to solve x where f'(x) = 0 | Taylor's expansion: f'(x\*) ≈ f'(xk) + f''(xk)(x\* - xk)  x\* is the stationary pt we want to find, so f'(x\*) = 0  Then, x\* ≈ xk - f'(xk)/f''(xk)  Chart  Description automatically generated with medium confidence  tan = f''(xk) = f'(xk)/(xk - xk+1)  Derivation for multivariate similar  Note quadratic convergence only when |x0 - x\*| is small | | |
| For multivariate, suppose and are both available.  Then  Trying to solve |
| To see how algo is performing, can plot (if we don't know optimal pt, i.e. **x**\*, then just use final iterate **x**T as approximate to **x**\*)  Or can plot f(**x**k) against k | | | |
| Line Search | Multivariate NLP: min f(**x**), **x**  Univariate NLP: tk = , g(t) := f(**x**k + t**v**k)  Need to find **v**k and tk (aka stepsize). Then let **x**k+1 = **x**k + tk**v**k (i.e. search along line) | | tk can be: 1) optimal step size: t = arg min f(**x**k + t**v**k)  2) Using some prefixed tk.  3) Backtrack with Armijo rule | |
| Backtrack | Armijo rule uses maximises t but in practical in terms of fn value decrease  (derivation in lect 5). 1) Start with t =  2) Check if reduction is met: f(**x**k + t**v**k) ≤ f(**x**k) + t  OR 2) = g(t) ≤ g(0) + t | | 3) If yes: use t for soln  4) If no: replace t with (where | |
| Newton + Line Search | Standard Newton's mtd:  Newton + Line Search: 1) Use **v**k =  2) Find tk (Standard Newton: tk = 1) | | 3) Let  4) Repeat step 2-3 until  This is called "Newton with X", where X = fixed step size/ backtracking/ optimal step size | |
| Gradient Descent / Steepest descent | Suppose we let **x**k+1 = **x**k + t**v**k w a small t and = 1  How to choose **v**k s.t. f(**x**k+1) is minimized? In practice, take **v**k =  (derivation in lect 5 using Cauchy Schwarz Inequality)  Mtd is called "Gradient descent with X", where X = fixed step size/ backtracking/ optimal step size | | | Input: **x**0, , tk, Tol  Output: **x**k so  k = 0; while do  **v**k = ; find tk based on some strategy;  **x**k+1 = **x**k + tk**v**k; k = k+1; |
| Complexity analysis of gradient descent | A C1 fn f is strongly convex in a convex D if with c > 0, f(**y**) ≥ f(**x**) + (**y** - **x**) + **x**, **y** D.  Equivalent to  Proposition: f C2 is strongly convex in D iff Hf(**x**) - cI (identity matrix) is PSD for all **x** D  is M-Lipschitz if . (If all eigenvalues of Hf(**x**) are below M, is M-Lipschitz  Thrm 6.2: Consider applying gradient descent w fixed step size, i.e. **x**k+1 = **x**k – t. If f is c-strongly convex, is M-Lipschitz and t ≤ , then  Thrm 6.2: Consider applying gradient descent w optimal step size, i.e. **x**k+1 = **x**k – tk, tk = . If f is c-strongly convex, is M-Lipschitz and t ≤ , then  Linear convergence: , which is slower than Newton's method | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Overview | Large dimensions: Coordinate Descent (Jacobi and Gauss-Seidel's mtd) | | Large data size: SGD |
| Coordinate descend | Note if f is C1, (**x**) = **0** f(**x**) = 0, i  Fixed point: xk+1 = xk (x no change after update) | If **x**\* is an optimal soln, then ... | |
| Jacobi's Mtd | First iterate: x0 = [].  g1(x1) = f(x1, ,...,). = . g2(x2) = f(, ,...,). = ... gn(xn) = f(, ,...,xn). =  x1 = []. So need to use inner loop to update each xi / weights. Outer loop update from x0 to x1 to .... | | |
| Gauss-Seidel Mtd | First iterate: x0 = [].  g1(x1) = f(x1, ,...,). = . g2(x2) = f(, ,...,). = ... gn(xn) = f(, ,...,xn). =  x1 = []. Double loop also. Main idea: use most recent value to calculate next xi. | | |
| Coordinate GD | Sometimes cannot solve minimization exactly. Instead use  Check if xk+1 meets criteria, else find next iterate | | |
| Extra | Jacobi can use parallel computing, Gauss-Seidel cannot. But GS should converge faster | |  |
| Overview | In data science, trying to find model to describe relation b = g(**x**, **a**) + . Linear Regression: b = **x**T**a** + . SVM: b = sgn(**x**T**a**)  **x**: weights/params we are trying to find. **z** = (**a**, b): data sample. **v**: input variable. b: output label. : possible noise.  Loss function at data point (**a**, b): F(**x**, **z**) = L(g(**x**, **a**), b). Linear regression use l2 loss: L(y) =  Data follows a dist (**a**, b) , want to minimize f(**x**) = E**z**F(**x**, **z**) = E**a**,bL(g(**x**, **a**), b) (expectation of loss fn over all data points/population loss) | | |
| E.g. | Consider LR problem, b = **a**T**x**\* + (true model). ~ N(0,1). If **a** ~ N(0, ∑), what is the pop loss w L2 loss?  Note E(**a**) = 0, Var(**a**) = E(**a**2) - [E(**a**)]2 = E(**aa**T) = ∑. E() = 0. E() = 1  Pop loss = E[L(g(**x**,**a**), b)] = E[] = E[(**a**T**x** - b)2] = E[(**a**T**x** - **a**T**x**\* - )2] = E[(**x** -**x**\*)T**aa**T(**x** - **x**\*) - 2**a**T(**x** - **x**\*) + ] =  (**x** -**x**\*)TE(**aa**T)(**x** - **x**\*) - 2E[**a**T(**x** - **x**\*)]E[] + E[] = (**x** -**x**\*)T∑(**x** - **x**\*) - 0 + 1 = (**x** -**x**\*)T∑(**x** - **x**\*) + 1 | | |
| Stochastic GD (SGD) | GD: **x**k+1 = **x**k + tk. = E**z**[F(**x**k, **z**)] = E**z**[F(**x**k, **z**)]. Target loss function: f(**x**) = E**z**[F(**x**, **z**)] = E**a**,b[L(g(**x**, **a**), b)]  If we have N data points **z**i, ≈ . When N is large, computation is expensive, O(N) for each update of **x** | | |
| SGD: draw a random **z**i from . = E[F(**x**k, **z**i)] = E[F(**x**k, **z**i)]  If draw **z**i from existing data, i is uniform from {1,...,N}. **x**k+1 = **x**k + tk . O(1) for each update of **x**  Ei[F(**x**k, **z**i)]= = . is an unbiased estimator | | |
| Step size must be fixed deterministic num: tk = t0, (0.5, 1]. Where = ∞ and < ∞  Stopping rule: and have to go through all data points.  Use average-SGD: for better result, as normal SGD might have oscillation around optimal point  Mini batch SGD: **x**k+1 = **x**k - , where M << N | | |
|  | Thrm: Suppose f is c-strongly convex, is bounded, tk = , then E[] = O(1/n)  1/n convergence. Takes decreasing step size | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Constrained Optimization | | Minimize f(**x**) **x**  s.t. gi(**x**) = 0, i = 1,..., m (equality constraint)  hj(**x**) ≤ 0, j = 1,..., p (inequality constraint) | f, gi, hj are all cts  The feasible set: S = |
| Regular point | Let **x**\* S. An inequality constraint hj(**x**) ≤ 0 is active at **x**\* if hj(**x**\*) = 0. Otherwise, it is inactive at **x**\*  For equality constraints, they will always be active  E.g. 0 ≤ x1 ≤ 4 h1(**x**) = -x1 ≤ 0. h2(**x**) = x1 ≤ 4. Sub in **x**. If hj(**x**) = 0, then hj is active | | |
| (LICQ): Let **x**\* S be a feasible point of (NLP). Let J(**x**\*) = be the index set of active constraints at **x**\*. Suppose that the set of gradient vectors is linearly independent (LI). Then **x**\* is a regular point or the regularity condition holds at **x**\*  These are known as constraint qualification at **x**\*. We only consider LI constraint qualification (LICQ) | | |
| LICQ: If only equality constraints are active, we check to be LI  If **x**\* is in the interior of S, only equality constraints are active (i.e. hj(**x**\*) < 0, so only gi active)  If only 1 constraint is active, check c is active, check c ≠ 0  If only 2 constraints c1, c2 are active, check c1, c2 are not multiple of each other  If m constraints are active, check [c1, ..., cm] have rank m | | |
| min f(**x**) = x1x2x3 s.t. g1(**x**) = x1 - x2 + 5x3 - 26 = 0, h1(**x**) = x1 + 2 - 4 ≤ 0, h2(**x**) = (x1 - 3)3 - 4 + x3 ≤ 0, h3(**x**) = + - x3 - 4 ≤ 0  Is **x** = [2;1;5] a regular point? g1(**x**) = h1(**x**) = h2(**x**) = 0 (active). h3(**x**) < 0 (inactive)  (**x**) = [1;-1;5]. (**x**) = [1;6;0]. (**x**) = [3;-8;1]. So **x** is feasible, now need check gradients of active constraints are LI  i.e. If a(**x**) + b(**x**) + c(**x**) = 0 and a = b = c = 0, then gradients are LI. Solve Simultaneous eqns to get a = b = c = 0. **x** is regular pt | | |
| KKT conditions | A pt **x**\* satisfies the KKT 1st order conditions if **x**\* is a regular point. And there are , ..., and , ..., s.t. , J(**x**\*) : index set of active inequality constraints at **x**\*  **x**\* is a KKT point if it satisfies KKT 1st order condition. | | |
| Lagrange multipliers: Multipliers for equality (, ..., ), Multipliers for inequality () and | | |
| Complementary Slackness condition: (i.e. if constraint is slack, hj(**x**\*) < 0, cannot be slack = 0, and vice versa) | | |
| KKT necessary condition. Thrm 12.1: Suppose f, gi, hj : are C1 on feasible set S. Suppose **x**\* S is regular. If **x**\* is a local min, then **x**\* is a KKT point | | |
| 2nd order KKT condition | | A point **x**\* satisfies the KKT 2nd order conditions, if **x**\* is a KKT pt, H­L(**x**\*) is a PD matrix, where HL(**x**\*) = Hf(**x**\*) + , where , are the Lagrangian multipliers (exactly same as in 1st order condition) | |
| Thrm 12.2: Let f, gi, hj : are C2 on feasible set S. Suppose **x**\* S is a KKT pt, i.e. there exist and s.t. . Let HL(**x**\*) as defined previously. Suppose HL(**x**\*) is PD, then **x**\* is a local minimizer of f on S | |
| General 2nd order KKT condition (not tested): A pt **x**\* satisfies the KKT 2nd order necessary conditions if **x**\* is a KKT pt. Let T(**x**\*) be the tangent space at **x**\*. T(**x**\*) = . For all **y** T(**x**\*), **y**THL(**x**\*)**y** ≥ 0. | |
| KKT necessary condition. Thrm 12.1: Suppose f, gi, hj : are C1 on feasible set S. Suppose **x**\* S is regular. If **x**\* is a local min, then **x**\* is a KKT point. If f, gi, hj are also C2, then **x**\* also satisfies the KKT 2nd order necessary condition. | |
| KKT sufficient condition. Thrm 12.2. Let f, gi, hj : be C2 fns. Suppose **x**\* S is a KKT pt, i.e. . there exist and s.t. . Let HL(**x**\*) as defined previously. **y**THL(**x**\*)**y** > 0, **y** T(**x**\*) - {0}. Then **x**\* is a strict local min of f on S | |
| E.g. | | min f(**x**) = (x1 - 4)2 + (x2 - 6)2. h1, ..., h4 inactive. h5(**x**) = 3x1 + 2x2 - 18 active. **x**\* = [34/13; 66/13] is feasible  For **x**\* to be regular, {h5} is LI, i.e. h5 ≠ 0. h5 = [3; 2]  For **x**\* to be KKT pt, (only consider active inequality constraint, since inactive multipliers = 0).  ) = [-36/13; -24/13]. Solving, = 12/13 > 0. So **x**\* is a KKT pt  For **x**\* to be local min, HL(**x**\*) must be PD. HL = Hf + (12/13) \* (inactive multipliers = 0). Solving HL = [2,0; 0,2] is PD, **x**\* is local min | |
| Find KKT points. 1) feasible, 2) regular, 3) KKT 1st order. Start with 3). Consider cases when h(**x**) active or inactive.  If have multiple KKT points, and asked to find global min, just check f(**x**) for all KKT pts. | |
| Summary | | |  |  |  | | --- | --- | --- | | Local min | Unconstrained | Constrained | | Necessary cond | **x**\* is a local min **x**\* is a critical/stationary pt  If f is C2, then Hf(**x**\*) must be PSD | **x**\* is a local min and regular **x**\* is a KKT pt  If f, gi, hj are C2, then HL(**x**\*) must be PSD on T(**x**\*) | | Sufficient cond | **x**\* is a stationary pt and H(**x**\*) is PD **x**\* is a local min | **x**\* is a KKT pt and HL(**x**\*) PD on T(**x**\*) **x**\* is a local min | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Primal Problem (P) | (P): min f(**x**) s.t. gi(**x**) = 0, i = 1,..., m, hj(**x**) ≤ 0, j = 1,...,p, **x** X, where X  Can impose additional constraints, include some constraints to simplify discussion | | Directly finding KKT pts can be difficult as multipliers are hard to find |
| Penalty for inequality constraints | Instead of min f(**x**) s.t. h1(**x**) ≤ 0. Consider min f(**x**) + h1(**x**)  ≥ 0 is the strength of penalty. If h1(**x**) > 0, adding h1(**x**) penalizes the loss. For sufficiently large , equivalent to requiring h1(**x**) ≤ 0  If h1(**x**) ≤ 0, . If h1(**x**) > 0,  Combine both cases, | | |
| Penalty for equality constraint | Instead of min f(**x**) s.t. g1(**x**) = 0. Consider min f(**x**) + g1(**x**)  If g1(**x**) < 0, with < 0 penalizes. If g1(**x**) > 0, with > 0 penalizes. For sufficiently large ± , equivalent to g1(**x**) = 0  i.e., | | |
| Lagrangian fn | (P): min f(**x**) s.t. gi(**x**) = 0, i = 1,..., m, hj(**x**) ≤ 0, j = 1,...,p, **x** X  Lagrangian fn: L(**x**,, ) = f(**x**) + | | Rewrite as  Can swap order of min and max. Doesn't matter |
| Lagrangian dual problem (D) | Lagrangian dual fn: , . (D): Lagrangian dual problem:  are called Lagrangian dual variables/Lagrangian multipliers  Proposition 6.1: Suppose the dual fn , is finite for all , with ≥ 0. Then is a concave fn  Then max , can be solved by finding stationary points. But sometimes may not be C1 | | |
| E.g. | min f(**x**) = s.t. h1(**x**) = -x1 -x2 + 4 ≤ 0. X = . Lagrangian fn: L(**x**, ) = f(**x**) + h1(**x**) = -x1 -x2 + 4  Dual fn: = = = when (x1 = x2 = )  Dual problem: = = 8 when = 4. So min f(**x**) = 8, x1 = x2 = 4/2 = 2 | | |
| E.g. | Same but now X = . From dual fn above, we want x1 = x2 =  x1 = if 0 ≤ ≤ 1 or 0 ≤ ≤ 2. x2 = if 1 ≤ ≤ 5 or 2 ≤ ≤ 10.  If 0 ≤ ≤ 2: x1 = /2. But x2 cannot be /2. But we know L(**x**, ) is increasing for x2 on [1,5]. So minimizer x2 = 1.  So = 0 + = + 1  If 2 ≤ ≤ 10: x2 = . x1 is optimised at x1 = 1. So = = + 1  If 10 ≤ : x1 optimised at x1 = 1. x2 optimised at x2 = 5. So = 26 - 2  . Dual problem: max . When , is decreasing. So max = = 6  When , . max when = 6, = 10 | | |
| Weak duality | Thrm 13.1: Consider primal problem (P). Let **x** be a feasible soln to (P) and , be a feasible soln to (D). Then f(**x**) ≥ ,  Corollary 6.1: If **x**\* is feasible to (P) and , is feasible to (D) s.t. f(**x**\*) = , , then **x**\* is an optimal soln to (P) and (, ) is an optimal soln to (D)  Duality Gap: min{f(**x**): **x**  S} – max{, : ≥ 0}. Can be used as stopping criteria for algos  For constrained NLP: solve Dual problem to get , . If given **x**\*. Check if f(**x**\*) = , **x**\* is global optimal | | |
| Linear optimization | min f(**x**) = **c**T**x** s.t. h­­j(**x**) = **x** – bj ≤ 0, j = 1,..., m. The feasible set: **Ax** ≤ **b** is a polygon  Proposition: Suppose the feasible set is a bounded polygon, 1 soln to the Linear Optimization problem is in the set V = , where the union is taken over all feasible combinations of ji, and d is the dimension.  Note vertices of polygon lies in V, i.e. optimizer is at one of the vertices (aka extreme points) | | |
| Basically L(**x**, ) = (**c**T + **A**)**x** - **b**.  1) **c**T + **A** ≠ **0**. Let **x** = -a(**c** + **A**). L(**x**, ) = -a- **b**. Let a +∞. L(**x**, ) – ∞ 2) **c**T + **A** = **0**. L(**x**, ) = - **b**  Dual problem: | | |
| E.g. | min f(**x**) = x1 + 2x2 s.t.  h1(**x**) = -x1 ≤ 0, h2(**x**) = -x2 ≤ 0, h3(**x**) = x1 + x2 - 2 ≤ 0  Consider vertices of polygon  f(0,0) = 0, f(0,2) = 4, f(2,0) = 2  So min f(**x**) = 0 | L(**x**, ) = x1 + 2x2 + h1 + h2 + h3 = (1 - + )x1 + (2 - + )x2 - 2  If (1 - + ) ≠ 0 or (2 - + ) ≠ 0, then = -∞  If (1 - + ) = (2 - + ) = 0, then = - 2  (1 - + ) = (2 - + ) = 0  = 0, = 1, = 1. = 0 = f(0,0) | |

|  |  |  |  |
| --- | --- | --- | --- |
| Flow chart | Diagram  Description automatically generated | Diagram  Description automatically generated | |
| Convex Optimization & KKT cond | Proposition: The feasible set of (P) is convex: min f(**x**) s.t. gi(**x**) := **x** - bi = 0, i = 1,...,m; hj(**x**) ≤ 0, j = 1,...,p  Thrm 12.3: Suppose f, hj are convex C1 fns and gi(**x**) := **x** - bi. If **x**\* S is a KKT pt, then **x**\* is a global minimizer of f on S  Necessity would require all optimal soln to be regular. Note (CP) = convex optimization problem | | |
| E.g. | Check f, hj convex -> find Hessian -> find eigenvalues of Hessian -> eigenvalues > 0 = PD = convex.  Note affine fns = linear fn + constant (e.g. x1 + x2 + 3) always convex.  Conclude (P) is a (CP). So next check x\* is a KKT point (LC of gradient of active constraint = 0) -> global min | | |
| min f(**x**) = x2 - log(1+x1), h1(**x**) = -x1 ≤ 0, h2(**x**) = -x2 ≤ 0, h3(**x**) = 2x1 + x2 - 3 ≤ 0. Solve optimization problem.  Check (P) is CP: find eigenvalue of Hessian of f(**x**). hi are affine fns, so convex.  Find KKT pt: 0 = =  So = 0 and 1 - + = 0. Note ≥ 0. So 2 > 0 and = 1 + > 0. So > 0, > 0  i.e. h2 and h3 are active, i.e. -x2 = 0 and 2x1 + x2 - 3 = 0. So = 1.5, = 0  Note h1(**x**\*) = -1.5 < 0. So = 0. So subbing back, = 6/5 ≥ 0, = 1/5 ≥ 0  So **x**\* is a KKT pt, and a optimal soln | | |
| Strong duality | Thrm 13.2: Suppose in the primal problem: X is convex, f, hj are convex, gi are affine. And there is **x**0 X s.t. h(**x**0) < 0 and g(**x**0) = 0. And 0 is interior of g(X) = {g(x): x X}. Then duality gap is 0: min{f(**x**): g(**x**) = 0, h(**x**) ≤ 0, **x**  X} = max{, : ≥ 0, }.  i.e. solving primal problem equivalent to solving dual problem (note for primal, **x** X converted to hi, but for dual, just ignore **x** X) | | |
| Projection on a convex set | Thrm 5.2: Suppose S is a closed convex set. Then for every **y**, there is a unique minimizer (**y**) of the problem min{}  To find projection: 1) Solve constrained optimization. 2) Use graphical interpretation find an ansatz, and check ansatz using KKT.  3) **x**\* = (**y**) iff ≤ 0 for all **x** S. Note if **y** S, then (**y**) = **y** | | |
| E.g. | For **y** R2, find (**y**) for S = { ≤ 1}. Want to min f(**x**) = s.t. h(**x**) = - ≤ 0  Using KKT mtd: = 0. **x** - **y** + **x** = 0. So **x** = **y**/(1 + )  Case 1) = 0. Then **x** = **y** which is feasible only if h(**x**) = h(**y**) < 0, i.e. ≤ 1  Case 2) > 0. Then h(**x**) = 0, i.e. = , i.e. = . Then **x** = | | |
| Half space: For **y** Rd, find (**y**) for S = {**a**T**x** + b ≤ 0}. Want to min f(**x**) = s.t. h(**x**) = **a**T**x** + b ≤ 0  0 = = **x** - **y** + **a**. So **x** = **y** - **a**  Case 1) > 0. Then h(**x**) = **a**T**x** + b = **a**T(**y** - **a**) + b = 0. Then = . Since > 0, then **a**T**y** + b > 0, i.e. **y** S  Case 2) = 0. Then **x** = **y** which is feasible if **y** S | | |
| SVM | mina s.t. ≥ 1, for all i = 1,...,n. S = {**a**T**x** + 1 ≤ 0}. Taking **y** = 0, (since margin line on both side of origin)  From earlier e.g., = . Then (**y**) = **y** - . .  max equivalent to min | | |
| Projected Gradient Descent | 1. Let **v**k = -f(**x**k)  2. Let g(t) = f(**xk** + t**v**k) and find tk = mint ≥ 0 g(t)  3. Let **x**' = **x**k + tk**v**k  4. Let **x**k+1 = (**x**') | | Project G.D. w fixed step sizes  1. Let **v**k = -f(**x**k)  2. Let **x**' = **x**k + tk**v**k  3. Let **x**k+1 = (**x**') |
| Thrm 5.2: For any **x** S, **y**, | | |
| E.g. | min f(**x**) = x1 + x2, s.t. ≤ 1. Start from **x**0 = **0**. Find **x**1 using project G.D. w step size t0 = 1  **x**' = **x**0 - t0f = **0** - [1;1] = [-1; -1]. **x**1 = (**x**') = = [ ; ] | | |
| min f(**x**) = , s.t. 0 ≤ x1, x2 ≤ 1. Start w **x**0 = **0**, find **x**1 using projected G.D. if A = [2,0 ; 0,1], b = [1;2]  f(**x**) = ... = - 2x1 + 1 + /2 - 2x2 + 2. f(**x**) = [2x1 - 2; x2 - 2]. **v**0 = - f(**x**0) = [2; 2]  g(t) = f(x0 + tv0) = f([2t; 2t]) = 6t2 - 8t + 3. g'(t) = 12t - 8 = 0. t\* = 2/3  **x**' = [4/3; 4/3]. **x**1 = (**x**') = [1; 1] | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Penalty mtd w inequality constraints | Consider min f(**x**) s.t. hj(**x**) ≤ 0, j = 1,...,p. The strictly feasible region S- := {x : hj(x) < 0, j}. Add a penalty B, P(**x**, ) = f(**x**) + B(**x**) which penalizes soln outside of S-  A barrier fn: B(x) = , where = – log() (logarithmic barrier). (which is too strong, so need decr strength)  Define P(**x**; ) = f(**x**) + B(**x**). Try to find xk = arg min P(**x**; ). If is finite, xk is in S-. So xk is always feasible. As decr to 0, we create a central path. | | | | |
| E.g. | min x s.t. -x + 1 ≤ 0. P(x, ) = f(x) - log(-h(x)) = x - log(x - 1). P'(x, ) = 1 - /(x-1). P'' > 0. For P'(x, ) = 0, x = + 1  = arg min P(x, ) = + 1. As 0, 1 (optimal pt) | | | | |
| Double loop structure | For stopping criteria: use duality gap  Consider grad at xk. 0 = P(xk; ) = f(xk) +  Consider Lagrangian w > 0. L(x, ) = f(x) +  We see xk is the local minimizer of L(x, ) | | | If we assume it is the global min, then = f(xk) + = f(xk) + = f(xk) + =  f(xk) – p, p = num of constraints  The duality gap is given by f(xk) - = p. Stop if p < | |
| Penalty mtd w equality constraints | Consider min f(**x**) s.t. gi(**x**) ≤ 0, i = 1,...,m. Let Q(x; ), the quadratic penalty fn, be Q(x; ) = f(x) +  When 0, the penalty for violating gi(x) = 0 is very heavy. | | | | |
| E.g. | min = f(x) s.t. g(x) = x1 + x2 - 1 = 0. Q(x, ) = f(x) + g(x)2  = arg min Q(x, ). Solve Q(x, ) = 0 to find x1, x2. Let 0, then = [x1; x2] tends to optimal soln | | | | |
| Algo | Logarithmic barrier mtd (inequality)  Input: **x**0, , f, h, , h, Tol  Output: **x**k s.t. h(**x**k) < 0, p < Tol  k = 0;  while p > Tol do:  **x** = **x**k;  while > Tol do:  | **v** = ; (which is our ­–P(x, )  | t = arg mint f(**x** + t**v**);  | **x** = **x** + t**v**;  **x**k+1 = **x**;  =  k = k+1; | | | Quadratic penalty mtd (equality)  Input: **x**0, , f, g, , g, Tol  Output: **x**k s.t. , ≤ Tol  k = 0;  while > Tol do:  **x** = **x**k;  while ≥ Tol do:  | **v** = ; (which is our ­–Q(x, )  | t = arg mint f(**x** + t**v**);  | **x** = **x** + t**v**;  **x**k+1 = **x**;  =  k = k+1; | |
| Penalties mtd in ML | SVM: (**a**1, b1), ..., (**a**N, bN). bi = -1 or 1  min**x**,b s.t. bi( + c) ≥ 1, for all i = 1,...,N | Penalised version: Consider slack variable = max{1 - bi( + c), 0}  min**x**,b s.t. ≥ 0, + bi( + c) ≥ 1, for all i = 1,...,N | | | |
| Regulariza-tion for ML | Consider model b = g(**x**, **a**) w data (**a**i, bi). Let F(**x**, **z**) = l(b, F(**x**, **a**))  min f(**x**) = E[F(**x**, **z**) + R(**x**)] ≈ + R(**x**)  F(**x**, **z**i) measures data misfit for ith data point  R(**x**) measures model simplicity. R aka regularization | | R(**x**) = (Tikhonov/ridge/L2 regularization): Smooth convex fn  R(**x**) = (Lasso/L1 regularization): cts and covex, favours sparsity  R(**x**) = = #{i: xi ≠ 0} (cardinality): non-cts and non-convex | | |
| E.g. | Ridge regression: f(**x**) = =  f(**x**) = AT(A**x** - **b**) + 2**x** = 0. Then **x** = (ATA + m)-1AT**b**  If = 0, then **x**\* = (ATA)-1AT**b**. But ATA might not be invertible (if rank of ATA < d, since rank (A) ≤ min(m,d). If m < d, then ATA cfm not invertible)  If ≠ 0, ATA + m is always invertible (since all its eigenvalues ≠ 0) | | | | = 2AT(Ax - b)  Lasso regression: f(**x**) = |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sub-gradient | | Let f: (-∞, ∞) be a convex fn. A vector **v** is a subgradient of f at **x** if f(**y**) ≥ f(**x**) +  The set of all subgradients at **x** = subdifferential of f at **x**, denoted as f(**x**) | | | | |
| If f is cts and convex, then is attained at **x**\* iff **0** f(**x\***) | | f is a convex set. If f is differentiable at **x**, then f(**x**) = {f(**x**)} | | |
| The subdifferential of f + g = (f + g)(**x**) {u + v: u f(**x**), v g(**x**)} | | | If f is differentiable at **x**, then (f + g)(**x**) = {f(**x**) + v: v g(**x**)} | |
| Given a set S, C = conv(S) is the convex hull of S, if C is the smallest convex set containing S | | | | |
| If S = {v1, ..., vn}, then conv(S) = | | | | |
| Suppose f(**x**) = max{f1(**x**), ..., fm(**x**)}, where fi are all convex and C1 fns. Suppose f(**x**\*) = f1(**x**\*) = ... = fj(**x**\*). Then f(**x\***) = conv({f1(**x**\*), ..., fj(**x**\*)}). | | | | Note that |g(**x**)| = max{-g(**x**), g(**x**)}  g(**x**) > / < / = 0, |g(**x**)| = {g(**x**)} / {g(**x**)} / conv{g(**x**), g(**x**)} |
| E.g. | f(x) = |x-1|. When x < 1, f(x) = {-1}. x > 1, f(x) = {1}. At x = 1, f(x) = conv{-1, 1} = [-1,1] which contains 0. So x\* = 1. | | | | | |
| Find subgradient of f(**x**) = |x1 + x2| + |x1| = g(x) + h(x). At (1,0), (0,1), (1,-1), (0,0), which is a global min  g(x) = max{x1 + x2, -x1 - x2} = max{g1(x), -g1(x)}. h(x) = max{x1, -x1} = max{h1(x), -h1(x)}  At (1,0), g(x) = x1 + x2. g(x) = {[1;1]}. h(x) = x1. h(x) = {[1;0]}. f(x) = {[2;1]}  At (0,1), g(x) = {[1;1]}. h(x) = h1(x) = - h1(x) = 0. So h(x) = conv{[1;0], [-1;0]}. So f(x) = conv{[2;1], [0;1]}. [0;0] is not in f(x) since x2 = 1  At (1,-1), g(x) = conv{[1;1], [-1;-1]}, h(x) = {[1;0]}. f(x) = conv{[2;1], [0;-1]}. Suppose [0;0] = [2;1] + (1-)[0;-1]. Then = 0 and = 0.5 (reject)  At (0,0), ... f(x) = conv{[2;1], [0;1], [0;-1], [-2;-1]}. For [2;1] + [0;1] + [0;-1] + [-2;-1] = 0, . ,  So 0 f(x) and (0,0) is minimizer. | | | | | |
| Subgradient Descent Mtd | | | Same as SGD/GD, just use **v**k = any element of subgradient f(**x**k) | | | |