Back to Optimization with constraints

$$\min f(x)$$
$$\text{s.t } g_i(x) \leq 0 \quad i = 1..m$$
$$\& \quad h_j(x) = 0 \quad j = 1..l$$

Temporarily absorb h_j's as ineq constraints

We need $-h_j$ & h_j both convex $\therefore$ affine h_j

From midsem Question

① Define: $I_{g_i}(x) = 0$ if $g_i(x) \leq 0$ &
$= \infty$ o/w

$$\min_x f(x) + \sum_i I_{g_i}(x)$$:- Convex function But not diff.

Solve by either analysing optimality conditions in terms of subgradients or employ subgradient descent...

② Write it equivalently as a cone program (yet to be analysed)... MIDSEM

(3) Replace $I_{g_i}(x)$ with a more "graceful" penalty function

$$\min_{x} f(x) - \sum_{i=1}^{m} \lambda_i \log(-g_i(x))$$

iteratively decrease $\lambda_i \geq 0$

(4) Instead consider the Lagrangian fn

$$L(x,\lambda) = f(x) + \sum \lambda_i g_i(x)$$

We will briefly visit (1) & then (4)

& later (2) & (3)

(3) Recall gradient descent & Newton:

$$x^{k+1} = \min_{x} f(x^k) + \nabla^T f(x^k)(x - x^k) + \frac{t}{2}(x-x^k)^T M (x-x^k)$$

$M = I$ for gradient desc & $\nabla^2 f(x^k)$ for Newton

"Proximal" / "Mirror descent" / Projection algos treat problem of finding x^{k+1} as that of locating next iterate as close as possible to x^k

In the sense of an approximation or in the sense of minimizing constraint violation etc

Suppose!

$$\min f(x)$$
$$\text{s.t } g_i(x) \le 0$$

$$\min f(x) + \eta \max_i g_i(x)$$

(we let η iteratively tend to ∞)

You need to find the formulation of the constrained opt problem for which the subgradient can be discovered easily.

H/W Eg Lasso: $\min_x \|Ax-y\|_2^2 \longrightarrow$ Regression loss/error

$$\|x\|_1 \le \theta$$

$$\boxed{\begin{aligned} &\min\ f(x) \\ &\text{st}\ \ g_i(x) \le 0 \end{aligned}}$$

option 1 (0/1)

$$I_{g_i}(x) \to \begin{cases} 0 & \text{if } g_i(x) \le 0 \\ \infty & \text{o/w} \end{cases}$$

If g_i is convex, dmn I_{g_i} is convex & $I_{g_i}(x)$ is a convex fn

$$\partial I_{g_i}(x) = \{ d \in \mathbb{R}^n \mid I_{g_i}(y) \ge I_{g_i}(x) + d^T(y-x) \ \forall y \}$$

∞ if $g_i(y) > 0$ so no issues

0 if $g_i(y) \le 0$

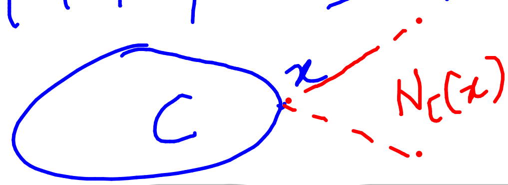
(if $g_i(x) \le 0$) $\{ d \in \mathbb{R}^n \mid 0 \ge d^T(y-x) \ \forall y \text{ s.t. } g_i(y) \le 0 \}$

$$= \underbrace{\{ d \in \mathbb{R}^n \mid d^T x \ge d^T y \ \forall y \text{ s.t. } g_i(y) \le 0 \}}$$

Normal cone $N_C(x)$ to C at point x.

1 If $x \in \text{int}(\text{dmn } g_i)$ then $N_C(x) = \{0\}$ i.e no nontrivial descent possible 2 Otherwise

$$N_C(x) = \{ d \in \mathbb{R}^n \mid d^T x \ge d^T y \ \forall y \in C \}$$



option 2 (continuous)

Let $C_i = \{ x \mid g_i(x) \le 0 \}$ are convex sets & let

$$\text{dist}(x, C_i) = \min \{ \|x-u\| : u \in C \}$$

If C_i is closed, convex then $\exists$ unique $u^* \in C$ that minimizes $\|x-u\|$. Let us call $u^* = P_{C_i}(x)$ so that

$$\text{dist}(x, C_i) = \|x - P_{C_i}(x)\|$$

We are interested in $\hat{x}$ st $g_1(x) \le 0, \dots g_m(x) \le 0$

ie $\hat{x} \in C_1 \cap C_2 \dots \cap C_m$

Claim: (If $\hat{x}$ exists)

$$\min_{x \in \mathbb{R}^n} \underbrace{\max_{i=1..m} \text{dist}(x, C_i)}_{\text{call it } D(x)} = 0$$

$D(\hat{x}) = 0$

$$\nabla \text{dist}(x, C_i) = \frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|}$$

if $D(x) = \text{dist}(x, C_i) \ne 0$ then $\dfrac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \in \partial D(x)$

For Lasso, it can be shown that for every θ there exists a $\lambda \geq 0$ s.t following two problems are equivalent:

$$\min_x \|Ax-y\|_2^2 + \lambda\|x\|_1$$

① ... say soln is x^* & $\|x^*\|_1 = \beta$

$$\min_x \|Ax-y\|_2^2 \quad \text{s.t } \|x\|_1 \leq \theta$$

② ... say solution is $\hat{x}$

Solution to ② with $\theta = \beta = x^*$ is also x^*!

Solution to ① with λ as soln to $A^T(y-Ax) = \lambda g_{\hat{x}}$ is also $\hat{x}$!

$g_{\hat{x}} \in \partial\|\hat{x}\|_1$

H/W: Subgradient of $\|x\|_1 = f(x)$ $\quad x \in \mathbb{R}^n$

$$f(x) = \|x\|_1 = \max_{i=1\ldots N} \{ f_1(x), f_2(x) \ldots f_i(x) \ldots f_N(x) \}$$

$f_1(x) = s_1^T x$, $\quad f_2(x) = s_2^T x$, $\quad \ldots$, $\quad f_N(x) = s_N^T x$ $\qquad N = 2^n$

$$s_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \qquad s_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad s_N = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ -1 \end{bmatrix}$$

if No component of $x = 0$ then $s = \begin{bmatrix} \operatorname{sgn}(x_1) \\ \operatorname{sgn}(x_2) \\ \vdots \\ \operatorname{sgn}(x_n) \end{bmatrix}$

In general if $f(x) = s_1'^T x = s_2'^T x = \ldots = s_k'^T x$

then $\partial f(x) = \operatorname{conv}\{ s_1', s_2' \ldots s_k' \}$ $\quad \ldots \quad (\partial f(x))_i = \begin{cases} +1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ \theta(-1) + (1-\theta)(1) & \text{if } x_i = 0 \\ \theta \in [0,1] & \end{cases}$

i.e. $\partial f(x) = \{ d \mid \|d\|_\infty \le 1,\ d^T x = \|x\|_1 \}$

Claim: If $\bar{a}_j = 2\sum_{i=1}^{n}[A_{ij}]^2$ & $\bar{b}_j = 2\sum_{i=1}^{n} A_{ij}(y_i - x_{-j}^T A_{-j})$

Then:

$$x_j^* = \begin{cases} (\bar{b}_j + \lambda)/\bar{a}_j & \text{if } \bar{b}_j < -\lambda \\ 0 & \text{if } \bar{b}_j \in [-\lambda, \lambda] \\ (\bar{b}_j - \lambda)/\bar{a}_j & \text{if } \bar{b}_j > \lambda \end{cases}$$

So to satisfy this, Lasso iterates on x^k as follows: $x^{(0)} \to \bar{b}^{(0)} \to x^{(1)}$ . . . . . . until convergence . . . We can understand through following simplification where $A = I$

Eg: $\min_x \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1$ → Regularizer $\left(\underset{x}{\operatorname{argmin}} \|y - x\|^2 + \lambda\|x\|_1 = x^*\right)$

$\lambda \geq 0$

I will suggest a soln by setting "some" $g_x = 0$

$$x_i^* = \begin{cases} -\lambda + y_i & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ \lambda + y_i & \text{if } y_i < -\lambda \end{cases}$$

Higher $\lambda \Rightarrow$ more x_i's are zeros

lots of zeros esp if several $|y_i| \leq \lambda$ .. sparsity

Why should this be imp for minimization? 2 ways of answering

① $g_x = \frac{1}{2}\nabla(\|y - x\|^2) + \lambda\partial\|x\|_1$

$= (x - y) + \lambda \begin{bmatrix} \text{sign}(x_1) \\ \vdots \\ \text{sign}(x_n) \end{bmatrix}$

② $\min_{x_i} \frac{1}{2}(y_i - x_i)^2 + \lambda|x_i|$

for each i $\quad g_{x_i} = (x_i - y_i) + \lambda\,\text{sign}(x_i)$

In either case ① or ②, setting $g_x = 0$ or $g_{x_i} = 0$ for each i, & checking that ✱ satisfies this equation,

Another example

Maximum eigenvalue of a symmetric matrix

$$f(x) = \lambda_{max}(A(x)) \quad .. \quad A(x) = A_0 + x_1 A_1 + \dots + x_n A_n$$

$$\& \ A_i \in S^m$$

$$f(x) = \lambda_{max}(A(x)) = \sup_{\|y\|_2 = 1} y^T A(x) y$$

Index set $\mathcal{I}$ over fns

each function is affine in x for fixed y & has gradient $\nabla f_y(x) = (y^T A_1 y, \dots y^T A_n y)$

http://en.wikipedia.org/wiki/Rayleigh_quotient

Active fns $y^T A(x) y$ are the ones for which y is (normalised) eigenvector for max eigenvalue λ_{max} of $A(x)$

$$\therefore g_x = (y^T A_1 y, \dots y^T A_n y)$$

Now visiting... 4

$$L(x,\lambda,\mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

for

$$\min f(x)$$

s.t. $g_i(x) \le 0 \quad i = 1..m$

$h_j(x) = 0 \quad j = 1..l$

$\min L(x,\lambda,\mu)$ --- you should ideally have $\lambda_i \ge 0$ to penalize $g_i(x) > 0$

$$\min_{\substack{x \\ g_i(x)\le 0 \\ h_j(x)=0}} f(x) \ge \min_{\substack{x \\ g_i(x)\le 0,\ \lambda_i \ge 0 \\ h_j(x)=0}} f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

$$\ge \min_{x,\ \lambda_i \ge 0} L(x,\lambda,\mu)$$

$$\min_x f(x) \geq \max_{\lambda_i \geq 0,\ \mu_j} \min_x L(x,\lambda,\mu)$$

s.t. $g_i(x) \leq 0$

$h_j(x) = 0$

Pushes up the lower bound from previous inequality.

<u>Importance</u>:

1. maximizing $L^*(\lambda,\mu)$ subject to just $\lambda_i \geq 0$ could be easier & provide a lower bound for original objective, provided $L^*(\lambda,\mu)$ has a manageable form.
2. $L^*(\lambda,\mu)$ is min over affine fns $L(x,\lambda,\mu)$ indexed by x ... $\therefore L^*(\lambda,\mu)$ is concave fn of λ & μ

max over concave has all "nice" properties that min over convex has

(3) $L^*(\lambda,\mu)$ is concave for all f, g_i & h_j choices. HOWEVER if f is convex, g_i's are convex & h_j's affine, most often the lower bound $\left(\max_{\substack{\lambda_i \geq 0 \\ \mu_j}} L^*(\lambda,\mu)\right)$ turns out to be the exact solution..

[Eg: Support Vector m/cs have dual easier to solve very often & dual gives exact solution as primal]

Duality gap: $f(x^*) - L^*(\lambda^*,\mu^*)$

Soln: x^*

PRIMAL: $\min_x f(x)$
s.t. $g_i(x) \leq 0$
$h_j(x) = 0$

$\geq$ (equality under convexity++)

DUAL: $\max_{\substack{\lambda_i \geq 0 \\ \mu_j}} L^*(\lambda,\mu)$

Soln: λ_i^*, μ_j^*

(4) Convergence criterion: Say Primal $(f(x^*))$ = Dual $(L^*(\lambda^*,\mu^*))$
$f(x^k) - L^*(\lambda^k,\mu^k)$ can be used as measure of distance from optimal soln where duality gap $= 0$

Recall: $f(x^*) \geq L^*(\lambda,\mu) \ \forall \lambda,\mu$

$$\min_{x} f(x) \le \max_{\substack{\lambda_i \ge 0 \\ \mu \in \mathbb{R}}} L^*(\lambda, \mu)$$

$$\text{s.t. } g_i(x) \le 0$$

$$h_j(x) = 0$$

Q1: Did we require f, g_i's & h_j's to be convex or affine? Ans: No

Q2: Is L^* concave irrespective of f, g_i's & h_j's? Note: $L(x, \lambda, \mu)$ is affine in λ, μ

$$L^* = \min_{x} \underbrace{L(x, \lambda, \mu)}_{L_x(\lambda, \mu)}$$

min of affine fns is concave

Next we provide insight into equality between primal & dual solutions through min-max theorem & saddle point theorem

Claim 1: Min-max / inf-sup inequality

$$\sup_y \inf_x f(x,y) \leq \inf_x \sup_y f(x,y)$$

Proof: $\forall x \left[f(x,y) \leq \sup_y f(x,y) \right] \Rightarrow \left[\inf_x f(x,y) \leq \inf_x \sup_y f(x,y) \right]$

$\forall y$

$$\Downarrow$$

$$\sup_y \inf_x f(x,y) \leq \inf_x \sup_y f(x,y)$$

Claim 2

A saddle point of a function $f : \mathcal{X} \times \mathcal{Y} \to \Re \cup \{\pm \inf\}$ is a pair $(\overline{x}, \overline{y}) \in \mathcal{X} \times \mathcal{Y}$ satisfying

$$\sup_y f(\overline{x}, y) \leq f(\overline{x}, \overline{y}) \leq \inf_x f(x, \overline{y})$$

Show that if $f(x,y)$ has a saddle point $(\overline{x}, \overline{y})$ then

$$\sup_y \inf_x f(x,y) = \inf_x \sup_y f(x,y)$$

Proof: By the min-max / inf-sup inequality

$$\sup_y \inf_x f(x,y) \leq \inf_x \sup_y f(x,y) \rightarrow (1)$$

Now, if f has a saddle point $(\bar{x}, \bar{y})$ then

By defn of saddle pt

$$\inf_x \sup_y f(x,y) \leq \sup_y f(\bar{x}, y) \leq f(\bar{x}, \bar{y}) \leq \inf_x f(x, \bar{y})$$

$$\leq \sup_y \inf_x f(x,y)$$

Thus

$$\inf_x \sup_y f(x,y) \le \sup_y \inf_x f(x,y) \rightarrow (2)$$

By (1) & (2), we have min-max equality!

$$\sup_y \inf_x f(x,y) = \inf_x \sup_y f(x,y) = f(\bar{x},\bar{y})$$

Illustration of saddle point at $(0,0)$ for $f(x_1,x_2) = x_1^2 - x_2^2$ on pages 242 & 243 of http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf

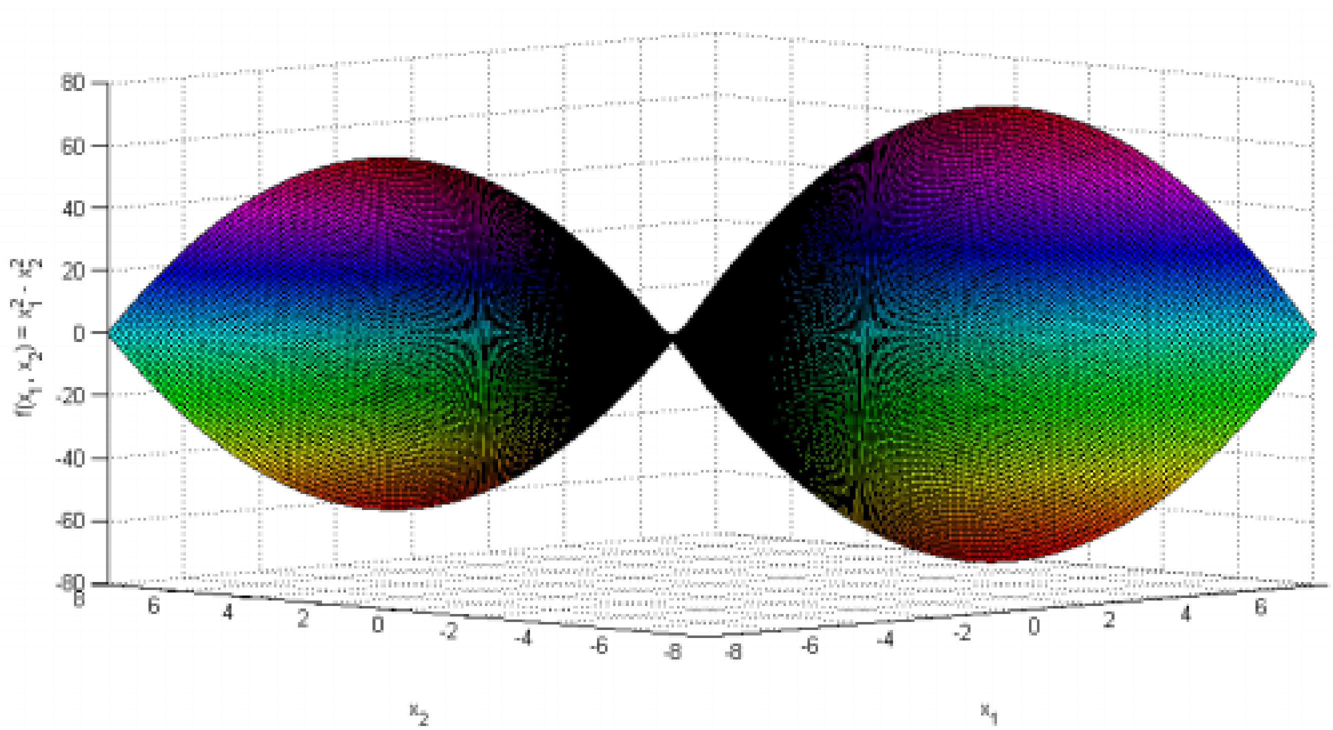


Figure 4.19: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, which

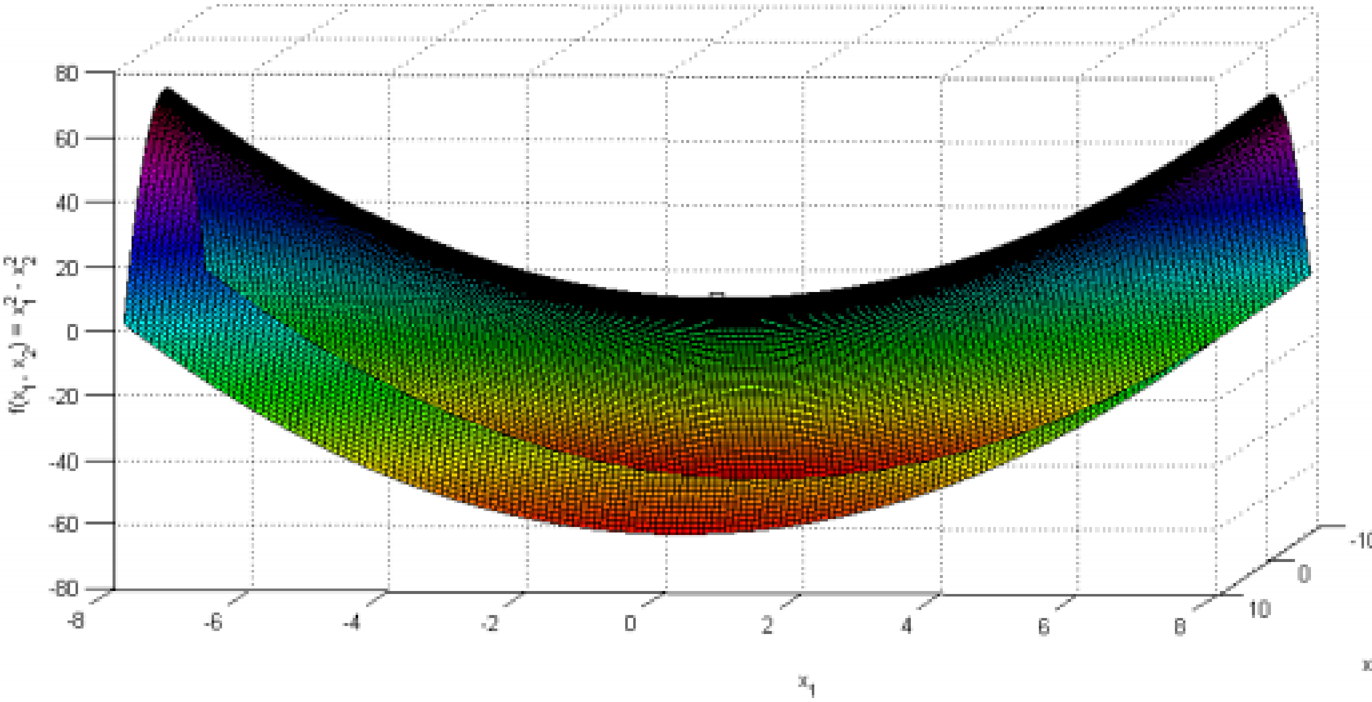


Figure 4.20: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, when vi
the x_1 axis is concave up.

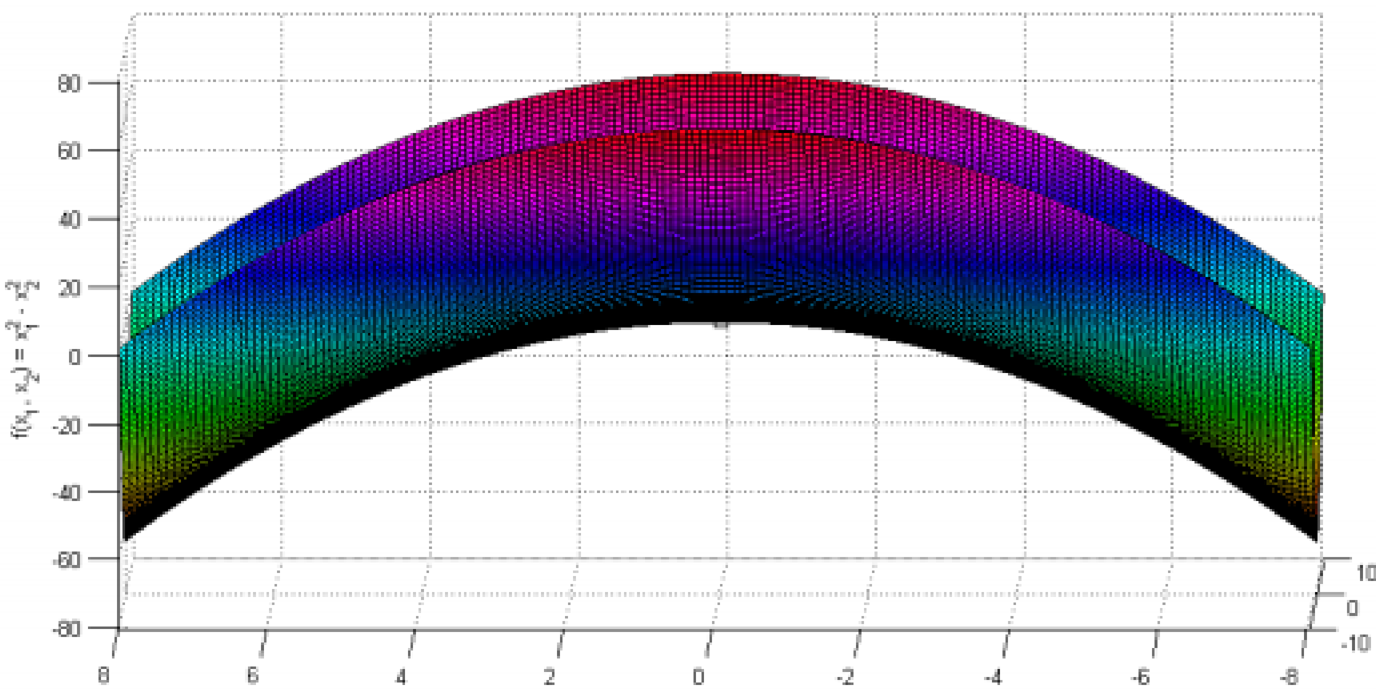


Figure 4.21: The hyperbolic paraboloid $f(x_1, x_2) = x_1^2 - x_2^2$, when vi
the x_2 axis is concave down.

$$\begin{cases} \min \ f(x) \\ \text{s.t } g_i(x) \le 0 \quad i = 1 \dots m \\ h_j(x) = 0 \quad j = 1 \dots k \end{cases}$$

We will generalize the inequalities & equalities

$$\min_x f(x) \quad \text{s.t } g_i(x) \le 0, \ h_j(x) = 0 \quad \geq \min_x \max_{\lambda,\mu} \underbrace{f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{k} \mu_j h_j(x)}_{L(x,\lambda,\mu)}$$

$$\text{s.t } g_i(x) \le 0, \quad h_j(x) = 0, \quad \lambda_i \geq 0 \quad \mu_j \in \mathbb{R}$$

By min-max thm: $\rightarrow$

$$\geq \min_x \max_{\lambda,\mu} L(x,\lambda,\mu) \qquad \lambda_i \geq 0, \ \mu_j \in \mathbb{R}$$

By saddle pt thm: If $L(x,\lambda,\mu)$ has a saddle pt $(\bar{x},\bar{\lambda},\bar{\mu})$ then equality holds

under strong duality

$\lambda_i^* g_i(x^*) = 0 \quad \forall i$

$\lambda_i^* \geq 0$

$\mu_j^* h_j(x^*) = 0 \quad \forall j$

$$\geq \max_{\lambda,\mu} \underbrace{\min_x L(x,\lambda,\mu)}_{L^*(\lambda,\mu) \text{ or lagrange dual fn.}} \qquad \lambda_i \geq 0, \ \mu_j \in \mathbb{R}$$

General weak duality result

$$= \underbrace{\max_{\lambda,\mu,\lambda\geq 0} L^*(\lambda,\mu)}_{\text{Dual opt problem}}$$

Karush Kuhn Tucker conditions (KKT conditions)

1. These conditions can be used to determine a saddle point $(\bar{x}, \bar{\lambda}, \bar{\mu})$ for $L(x, \lambda, \mu)$
2. We will first graphically motivate the KKT conditions
3. Next we will prove their necessity & sufficiency for optimality under convexity