Definition 41 [Subgradient]: *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A vector $\mathbf{h} \in \Re^n$ is said to be a subgradient of f at the point $\mathbf{x} \in \mathcal{D}$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{y} \in \mathcal{D}$. The set of all such vectors is called the subdifferential of f at $\mathbf{x}$.

Theorem 76 *Let $f : \mathcal{D} \rightarrow \Re$ be a convex function defined on a convex set $\mathcal{D}$. A point $\mathbf{x} \in \mathcal{D}$ corresponds to a minimum if and only if*

$$\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geq 0$$

for all $\mathbf{y} \in \mathcal{D}$.

If $\nabla f(\mathbf{x})$ is nonzero, it defines a supporting hyperplane to $\mathcal{D}$ at the point $\mathbf{x}$. Theorem 77 implies that for a differentiable convex function defined on an open set, every critical point must be a point of (global) minimum.

Theorem 77 *Let $f : \mathcal{D} \rightarrow \Re$ be differentiable and convex on an open convex domain $\mathcal{D} \subseteq \Re^n$. Then $\mathbf{x}$ is a critical point of f if and only if it is a (global) minimum.*

Theorem 78 *Let $f : \mathcal{D} \rightarrow \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,*

1. *f is convex on $\mathcal{D}$ if and only if is its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) \geq 0 \tag{4.53}$$

2. *f is strictly convex on $\mathcal{D}$ if and only if is its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) > 0 \tag{4.54}$$

3. *f is uniformly or strongly convex on $\mathcal{D}$ if and only if is its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$,*

$$\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T(\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2 \tag{4.55}$$

for some constant $c > 0$.

Necessity: Suppose f is uniformly convex on $\mathcal{D}$. Then from theorem 75, we know that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$
f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) - \frac{1}{2}c||\mathbf{y}+\mathbf{x}||^2
$$
$$
f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x}-\mathbf{y}) - \frac{1}{2}c||\mathbf{x}+\mathbf{y}||^2
$$

Adding the two inequalities, we get (4.55). If f is convex, the inequalities hold with $c = 0$, yielding (4.54). If f is strictly convex, the inequalities will be strict, yielding (4.54).

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

$$
\phi(1) - \phi(0) = \phi'(t) \tag{4.56}
$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y}-\mathbf{x})$, (4.56) translates to

$$
f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y}-\mathbf{x}) \tag{4.57}
$$

Also, by definition of monotonicity of ∇f, (from (4.53)),

$$
\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x}) = \frac{1}{t}\left(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x})\right)^T (\mathbf{z}-\mathbf{x}) \geq 0 \tag{4.58}
$$

Combining (4.57) with (4.58), we get,

$$
\begin{aligned}
f(\mathbf{y}) - f(\mathbf{x}) &= \left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \\
&\geq \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})
\end{aligned} \tag{4.59}
$$

By theorem 75, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (4.58) inherited from strict monotonicity, and letting the strict inequality follow through to (4.59). For the case of strong convexity, from (4.55), we have

$$
\begin{aligned}
&\phi'(t) - \phi'(0) = \left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x}) \\
&= \frac{1}{t}\left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{z}-\mathbf{x}) \geq \frac{1}{t}c||\mathbf{z}-\mathbf{x}||^2 = ct||\mathbf{y}-\mathbf{x}||^2
\end{aligned} \tag{4.60}
$$

$$
\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)]dt \geq \frac{1}{2}c||\mathbf{y}-\mathbf{x}||^2 \tag{4.61}
$$

which translates to

$$
f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}c||\mathbf{y}-\mathbf{x}||^2
$$

What abt local maxima/minima & subgradient?

$\nabla f(x) = 0$ & f is convex then x is global min

What if $g_x = 0$?

$$f(y) \geq f(x) + g_x^T (y - x) \quad \forall y$$

if $g_x = 0$ then $f(y) \geq f(x) \Rightarrow x$ is pt of global min

Regularizer

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

$\lambda \geq 0$

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

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

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

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

Subgradient method

Descent: $\nabla^T f(x^k)\Delta x < 0$

subgradient method is simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

Instead of $\nabla f(x^k)$ you compute some subgradient $g^{(k)}$ at pt $x^{(k)}$

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

not a descent method, so we keep track of best point so far

Earlier: $\Delta x^k = -\nabla f(x^k) = \operatorname{argmin}_{\|v\|_2=1} v^T \nabla f(x^k)$

$$f_{\mathrm{best}}^{(k)} = \min_{i=1,\ldots,k} f(x^{(i)})$$

We know: $f(x^{k+1}) \geq f(x^k) + g^{(k)T}(x^{k+1} - x^k)$

Step size rules

step sizes are fixed ahead of time

- *constant step size:* $\alpha_k = \alpha$ (constant)
- *constant step length:* $\alpha_k = \gamma / \|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$)
- *square summable but not summable:* step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

- *nonsummable diminishing:* step sizes satisfy

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Assumptions

- $f^\star = \inf_x f(x) > -\infty$, with $f(x^\star) = f^\star$
- $\|g\|_2 \leq G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on f)
- $\|x^{(1)} - x^\star\|_2 \leq R$

these assumptions are stronger than needed, just to simplify proofs

Stopping criterion

- terminating when $\dfrac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i} \leq \epsilon$ is really, really, slow

- optimal choice of α_i to achieve $\dfrac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i} \leq \epsilon$ for smallest k:

  $$\alpha_i = (R/G)/\sqrt{k}, \quad i = 1, \ldots, k$$

  number of steps required: $k = (RG/\epsilon)^2$

- the truth: there really isn't a good stopping criterion for the subgradient method . . .

Example: Piecewise linear minimization

$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

to find a subgradient of f: find index j for which

$$a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

and take $g = a_j$

subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k a_j$

Speeding up subgradient methods

- subgradient methods are very slow
- often convergence can be improved by keeping memory of past steps

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

(heavy ball method)

other ideas: localization methods, conjugate directions, . . .

Back to optimization with constraints

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

include: $h_j(x) \le 0$ & $-h_j(x) \le 0$

option 1 (0|1) $\rightarrow$

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

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 $\hat{x}$ s.t. $g_1(x) \le 0, \dots g_m(x) \le 0$

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 $\dfrac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \in \partial D(x)$

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

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

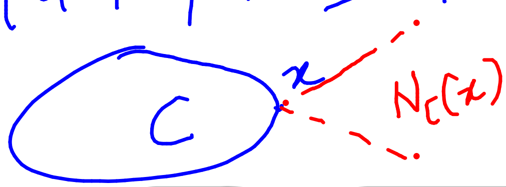
∞ if $g_i(y) > 0$ so no issues

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 $N_C(x)$ for convex set C at pt x is $\{d \in \mathbb{R}^n \mid d^T x \ge d^T y \ \forall y \in C\}$



$$\boxed{\begin{aligned} \min_{x}\ & f(x) \\ \text{s.t}\ & g_i(x) \le 0 \\ & h_j(x) = 0 \end{aligned}} \equiv \boxed{\begin{aligned} \min_{x}\ & f(x) \\ \text{s.t}\ & g_i(x) \le 0 \\ & h_j(x) \le 0 \\ & -h_j(x) \le 0 \end{aligned}}$$

$h_j(x)$ & $-h_j(x)$ are both convex $\Rightarrow$ $h_j(x)$ is affine ie

$$[h(x)] = Ax + b = 0$$

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \quad h_j(x) = a_j^T x + b_j$$

Equivalent convex problems

two problems are (informally) **equivalent** if the solution of one is readily obtained from the solution of the other, and vice-versa

some common transformations that preserve convexity:

- **eliminating equality constraints**

To solve analytically
OR
To apply descent methods

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b \end{array}$$

is equivalent to

$\nabla f_0(z) = F \nabla f_0(x)$
$\nabla f_i(z) = F \nabla f_i(x)$

$$\begin{array}{ll} \text{minimize (over } z) & f_0(Fz + x_0) \\ \text{subject to} & f_i(Fz + x_0) \leq 0, \quad i = 1, \dots, m \end{array}$$

where F and x_0 are such that

Q: What if we want to invoke steepest descent with ∞ norm or 1 norm in $\|v\|_p = 1$?

$$Ax = b \quad \Longleftrightarrow \quad x = Fz + x_0 \text{ for some } z$$

- **introducing equality constraints**

$$\begin{array}{ll} \text{minimize} & f_0(A_0x + b_0) \\ \text{subject to} & f_i(A_ix + b_i) \leq 0, \quad i = 1, \dots, m \end{array}$$

is equivalent to

$$\begin{array}{ll} \text{minimize (over } x, y_i) & f_0(y_0) \\ \text{subject to} & f_i(y_i) \leq 0, \quad i = 1, \dots, m \\ & y_i = A_ix + b_i, \quad i = 0, 1, \dots, m \end{array}$$

- **introducing slack variables for linear inequalities**

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & a_i^Tx \leq b_i, \quad i = 1, \dots, m \end{array}$$

is equivalent to

$$\begin{array}{ll} \text{minimize (over } x, s) & f_0(x) \\ \text{subject to} & a_i^Tx + s_i = b_i, \quad i = 1, \dots, m \\ & s_i \geq 0, \quad i = 1, \dots m \end{array}$$

Convex f, g_i

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

option 1: $F(x) = f(x) + \sum_i I_{g_i}(x)$

option 2: $F(x) = f(x) + \max_i \min_{u : g_i(u) \leq 0} \|x - u\|_2$

$$\min_x F(x)$$

Either obtain solution by setting $g_F(x) = 0$ & solving for x OR applying a descent algorithm

Convex f, g_i

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

option 1 $\rightarrow$

$$F(x) = f(x) + \underbrace{\sum_i I_{g_i}(x)}_{\text{subgradient = normal cone (on bndry)}}$$

option 2 $\rightarrow$

$$F(x) = f(x) + \max_i \underbrace{\min_{u : g_i(u) \le 0} \|x - u\|_2}_{\frac{(x - P_{g_i}(x))}{\|x - P_{g_i}(x)\|}}$$

$$\min_x F(x)$$

Either obtain solution by setting $g_F(x) = 0$ & solving for x OR applying a descent algorithm.

option 3 $\rightarrow$

$$F_t(x) = f(x) + \left(-\frac{1}{t}\right) \sum_i \log(-g_i(x))$$

$$\underbrace{x^*(t) = \operatorname*{argmin}_x F_t(x)}_{\text{Barrier method}}$$

It turns out that analysing Barrier method (or analysing convergence of prox/projected gradient descent) becomes meaningful if we understand conditions for optimality for constrained opt. ..

option 4 $\rightarrow$

$$F_k(x) = f_{Q_k}(x) + \sum_i I_{g_i}(x)$$

$$\underbrace{x^{k+1} = \operatorname*{argmin}_x F_k(x)}_{\text{Projected gradient method}}$$

Recall: $f_{Q_k}(x)$ = Quadratic approx to f around x^k

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

$$\therefore x^{k+1} = \operatorname*{argmin}_x \frac{1}{2t} \left\| x - \left(x^k - t \nabla f(x^k)\right) \right\|^2 + \sum_i I_{g_i}(x)$$

$$= \operatorname*{argmin}_{x : g_i(x) \le 0} \|x - \hat{x}^{k+1}\|^2$$

$$= P_{C_1 \cap C_2 \ldots \cap C_m}(\hat{x}^{k+1})$$

More generally, the 4th option:

called projected gradient descent

Proximal gradient descent.

$$\min \underbrace{f(x)}_{} + \underbrace{r(x)}_{}$$

A = features
x = feature wts
in Lasso

f is differentiable
$\|Ax-y\|^2$

r is not differentiable
eg: $\lambda\|x\|_1$

Iteratively solve: $x^{(0)}$

$$x^{(k+1)} = \min_x f(x^{(k)}) + \nabla f(x^{(k)})(x - x^k) + \frac{1}{2t}\|x - x^k\|^2 + r(x)$$

until convergence

For our problem: $x^{(k+1)} = \min_x \|x-y\|^2 \ldots \|x\|_1$

H/w: complete & reduce to known problem

Recall: $\min_x \|y-x\|^2 + \lambda\|x\|_1$ had a closed form optimal soln:

$$x^*_i = \begin{cases} y_i + \lambda & \text{if } y_i < -\lambda \\ -y_i + \lambda & \text{if } y_i > \lambda \\ 0 & \text{o/w} \end{cases}$$

Obtained by setting a subgradient to 0.

Projected gradient descent is prox gradient descent when $r(x) = I_\Omega(x)$

NECESSARY CONDITIONS FOR CONSTRAINED OPTIMALITY (pages 284-287 of . . .
http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf)

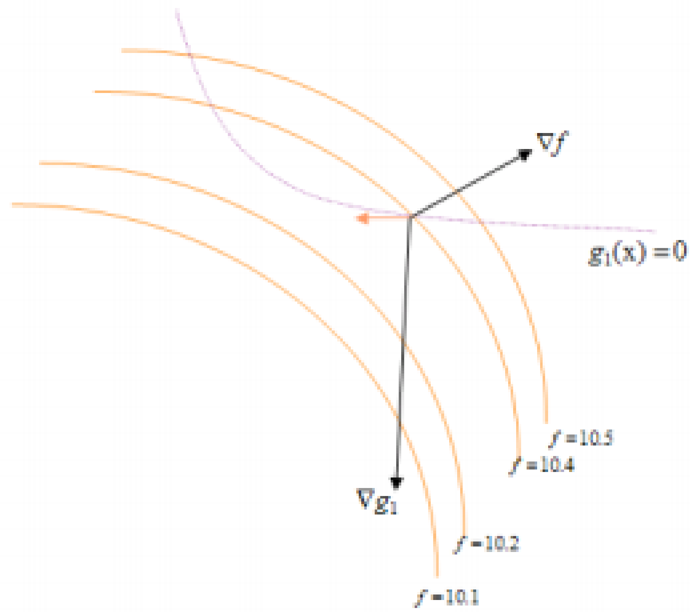


Figure 4.39: At any non-optimal and non-saddle point of the equality constrained problem, the gradient of the constraint will not be parallel to that of the function.

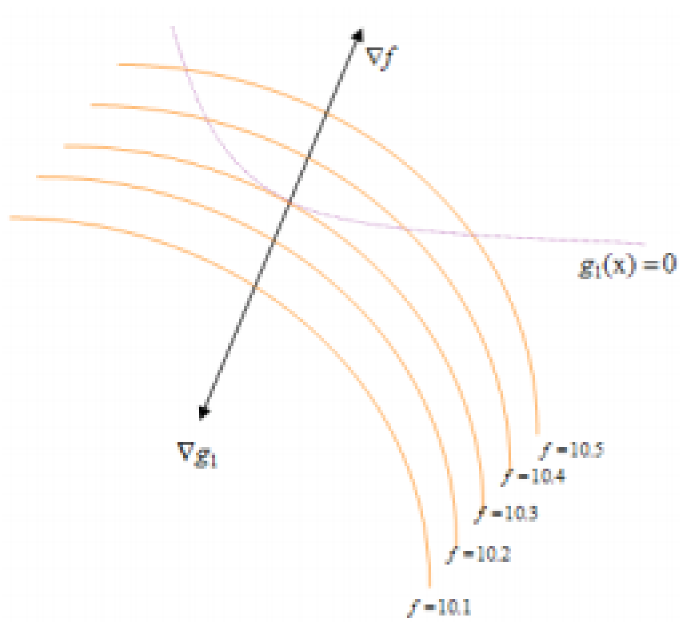


Figure 4.40: At the equality constrained optimum, the gradient of the constraint must be parallel to that of the function.

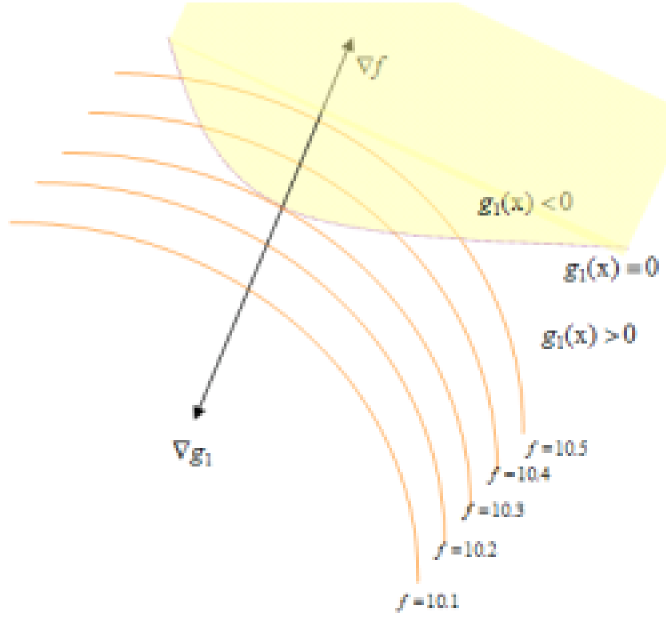


Figure 4.41: At the inequality constrained optimum, the gradient of the constraint must be parallel to that of the function.