Convexity, Local and Global Optimality, etc.

Local and Global Minima, Gradients and Convexity

- Recall that for functions of single variable, at local extreme points, the tangent to the curve is a line with a constant component in the direction of the function and is therefore parallel to the *x*-axis.
  - If the function is differentiable at the extreme point, then the derivative must vanish.
- This idea can be extended to functions of multiple variables. The requirement in this case turns out to be that the tangent plane to the function at any extreme point must be parallel to the plane $z = 0$.
  - This can happen if and only if the gradient ∇F is parallel to the $z-$axis at the extreme point, or equivalently, the gradient to the function f must be the zero vector at every extreme point.

(Sub)Gradients and Optimality: Sufficient Condition

- For a convex *f*, 0 belonging to subdifferential of f at x

As a sufficient condition for minimum at x

(Sub)Gradients and Optimality: Sufficient Condition

- For a convex f,

$$f(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \Leftarrow 0 \in \partial f(\mathbf{x}^*)$$

- The reason: $\mathbf{h} = 0$ being a subgradient means that for all $\mathbf{y}$

f(y) >= f(x) + ...zero term...

(Sub)Gradients and Optimality: Sufficient Condition

- For a convex f,

$$f(\mathbf{x}^*) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \Leftarrow 0 \in \partial f(\mathbf{x}^*)$$

- The reason: $\mathbf{h} = 0$ being a subgradient means that for all $\mathbf{y}$

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + 0^T(\mathbf{y} - \mathbf{x}^*) = f(\mathbf{x}^*)$$

- The analogy to the differentiable case is: $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.
- Thus, for a convex function $f(\mathbf{x})$, if $\nabla f(\mathbf{x}) = 0$, then $\mathbf{x}$ must be a point of glolbal minimum.
- Is there a necessary condition for a differentiable (possibly non-convex) function having a (local or global) minimum at $\mathbf{x}$? (A little later)

Subgradients in Lasso: Sufficient Condition Test

We illustrate the sufficient condition again using Lasso as an example. Consider the simplified Lasso problem: y and \lambda are specified and fixed as inputs to the probler

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

Recall the subgradients of $f(\mathbf{x})$:

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

To talk of convexity, we woulc
like \lambda > 0

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.
A solution to this problem is

if yi > \lambda, xi = yi - \lambda
if yi < -\lambda, xi = yi + \lambda
otherwise xi = 0

Subgradients in Lasso: Sufficient Condition Test

We illustrate the sufficient condition again using Lasso as an example. Consider the simplified Lasso problem:

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

Recall the subgradients of $f(\mathbf{x})$:

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.
A solution to this problem is $\mathbf{x}^* = S_\lambda(\mathbf{y})$, where $S_\lambda(\mathbf{y})$ is the soft-thresholding operator:

$$S_\lambda(\mathbf{y}) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

Now let $\mathbf{x}^* = S_\lambda(\mathbf{y})$ and we can get $g = 0$. Why? If $y_i > \lambda$, we have $x_i^* - y_i = -\lambda + \lambda \cdot 1 = 0$. The case of $y_i < \lambda$ is similar. If $-\lambda \leq y_i \leq \lambda$, we have $x_i^* - y_i = -y_i + \lambda(\frac{y_i}{\lambda}) = 0$. Here, $s_i = \frac{y_i}{\lambda}$.

Proximal Operator and Sufficient Condition Test

- Recap: $d(\mathbf{x},\mathcal{C})$ returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x},\mathcal{C}) = \inf_{y\in\mathcal{C}} ||\mathbf{x}-\mathbf{y}||^2$. Then $d(\mathbf{x},\mathcal{C})$ is a convex function.
- Recap: $\underset{y\in\mathcal{C}}{\operatorname{argmin}} ||\mathbf{x}-\mathbf{y}||$ is a special case of the proximal operator: $prox_f(\mathbf{x}) = \underset{y}{\operatorname{argmin}}\, PROX_f(\mathbf{x},\mathbf{y})$ of a convex function $f(\mathbf{x})$. Here, $PROX_f(\mathbf{x},\mathbf{y}) = f(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||^2$ The special case is when $f(\mathbf{y})$ is the indicator function $I_C(\mathbf{y})$ introduced earlier to eliminate the contraints of an optimization problem.
  - Recall that $\partial I_C(\mathbf{y}) = N_C(\mathbf{y}) = \{\mathbf{h}\in\Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z}\in C\}$
  - For the special case $f(\mathbf{y}) = I_C(\mathbf{y})$, the subdifferential $\partial PROX_f(\mathbf{x}.\mathbf{y}) = \partial f(\mathbf{y}) + \mathbf{y} - \mathbf{x} = \{\mathbf{h}-\mathbf{x}\in\Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z}\in C\}$
  - As per sufficient condition for minimum for this special case, $prox_f(\mathbf{x}) = \underset{y\in\mathcal{C}}{\operatorname{argmin}} ||\mathbf{x}-\mathbf{y}||$
- We will invoke this when we discuss the **proximal gradient descent** algorithm

Local Extrema: Necessary Condition

Definition

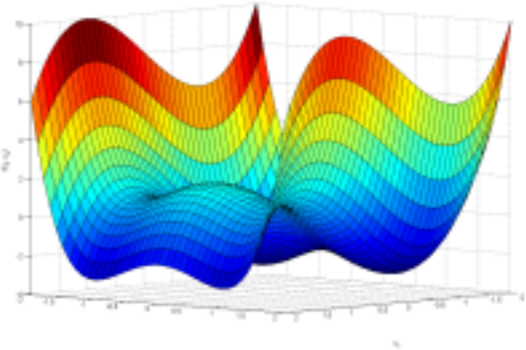
[Recap: Local maximum]: *A function f of n variables has a local maximum at* $\mathbf{x}^0$ *if* $\exists \epsilon > 0$ *such that* $\forall \ ||\mathbf{x} - \mathbf{x}^0|| < \epsilon$. $f(\mathbf{x}) \leq f(\mathbf{x}^0)$. *In other words,* $f(\mathbf{x}) \leq f(\mathbf{x}^0)$ *whenever* $\mathbf{x}$ *lies in some circular disk around* $\mathbf{x}^0$.

Definition

[Recap: Local minimum]: *A function f of n variables has a local minimum at* $\mathbf{x}^0$ *if* $\exists \epsilon > 0$ *such that* $\forall \ ||\mathbf{x} - \mathbf{x}^0|| < \epsilon$. $f(\mathbf{x}) \geq f(\mathbf{x}^0)$. *In other words,* $f(\mathbf{x}) \geq f(\mathbf{x}^0)$ *whenever* $\mathbf{x}$ *lies in some circular disk around* $\mathbf{x}^0$.

Recap: Local Extrema

Figure below shows the plot of $f(x_1, x_2) = 3x_1^2 - x_1^3 - 2x_2^2 + x_2^4$. As can be seen in the plot, the function has several local maxima and minima.



Local Extrema: Necessary Condition through Fermat's Theorem

A theorem fundamental to determining the locally extreme values of functions of multiple variables.

Claim

If $f(\mathbf{x})$ defined on a domain $\mathcal{D} \subseteq \Re^n$ has a local maximum or minimum at $\mathbf{x}^*$ and if the first-order partial derivatives exist at $\mathbf{x}^*$, then $f_{x_i}(\mathbf{x}^*) = 0$ for all $1 \leq i \leq n$.

Proof:

We already saw the intuition that the rate of increase or decrease of the function at x^* should be 0 in all directions

Local Extrema: Necessary Condition through Fermat's Theorem

A theorem fundamental to determining the locally extreme values of functions of multiple variables.

Claim

If $f(\mathbf{x})$ defined on a domain $\mathcal{D} \subseteq \Re^n$ has a local maximum or minimum at $\mathbf{x}^*$ and if the first-order partial derivatives exist at $\mathbf{x}^*$, then $f_{x_i}(\mathbf{x}^*) = 0$ for all $1 \leq i \leq n$.

Proof: The idea behind this result can be stated as follows. The tangent hyperplane to the function at any extreme point must be parallel to the plane $z = 0$. This can happen if and only if the gradient $\nabla F = [\nabla^T f, \ -1]^T$ is parallel to the $z-$axis at the extreme point. Or equivalently, the gradient to the function f must be the zero vector at every extreme point, *i.e.*, $f_{x_i}(\mathbf{x}^*) = 0$ for $1 \leq i \leq n$.

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $\underline{g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)}$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then

H/w

gi is f at point of optimum x^* along all components except along the i^th component

Basically observing behaviour of f at point of optimum along individual dimensions

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector,

||e_i|| = 1
|e_i^T (x - x^*)| <= ||e_i|| ||x-x^*||

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector, $|\mathbf{e}_i^T(\mathbf{x} - \mathbf{x}^*)| = |x_i - x_i^*| \leq \|\mathbf{x} - \mathbf{x}^*\|\|\mathbf{e}_i\| = \|\mathbf{x} - \mathbf{x}^*\|$.
4. Thus, the existence of an open ball $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ characterizing the minimum in $\Re^n$ also guarantees

the existence of an open ball around xi^* in R which corresponds to the minimum of g_i(xi^*) = f(x^*)

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector, $|\mathbf{e}_i^T(\mathbf{x} - \mathbf{x}^*)| = |x_i - x_i^*| \leq \|\mathbf{x} - \mathbf{x}^*\|\|\mathbf{e}_i\| = \|\mathbf{x} - \mathbf{x}^*\|$.
4. Thus, the existence of an open ball $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ characterizing the minimum in $\Re^n$ also guarantees the existence of an open ball (projected ball corresponding to a projected norm) $\{x_i | \|x_i - x_i^*\| < \epsilon\}$ around x_i^* in $\Re$.
5. Therefore each function $g_i(x_i)$ must have a local extremum at x_i^*. Which, by an earlier result (derived for differentiable functions of single argument) implies that

g_i'(xi^*) = 0

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector, $|\mathbf{e}_i^T(\mathbf{x} - \mathbf{x}^*)| = |x_i - x_i^*| \leq \|\mathbf{x} - \mathbf{x}^*\| \|\mathbf{e}_i\| = \|\mathbf{x} - \mathbf{x}^*\|$.
4. Thus, the existence of an open ball $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ characterizing the minimum in $\Re^n$ also guarantees the existence of an open ball (projected ball corresponding to a projected norm) $\{x_i | \|x_i - x_i^*\| < \epsilon\}$ around x_i^* in $\Re$.
5. Therefore each function $g_i(x_i)$ must have a local extremum at x_i^*. Which, by an earlier result (derived for differentiable functions of single argument) implies that $g_i^{'}(x_i^*) = 0$
6. Now $g_i^{'}(x_i^*) = f_{x_i}(\mathbf{x}^*)$ and hence the gradient of f must vanish at x^*

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector, $|\mathbf{e}_i^T(\mathbf{x} - \mathbf{x}^*)| = |x_i - x_i^*| \leq \|\mathbf{x} - \mathbf{x}^*\| \|\mathbf{e}_i\| = \|\mathbf{x} - \mathbf{x}^*\|$.
4. Thus, the existence of an open ball $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ characterizing the minimum in $\Re^n$ also guarantees the existence of an open ball (projected ball corresponding to a projected norm) $\{x_i | \|x_i - x_i^*\| < \epsilon\}$ around x_i^* in $\Re$.
5. Therefore each function $g_i(x_i)$ must have a local extremum at x_i^*. Which, by an earlier result (derived for differentiable functions of single argument) implies that $g_i^{'}(x_i^*) = 0$
6. Now $g_i^{'}(x_i^*) = f_{x_i}(\mathbf{x}^*)$ and hence $f_{x_i}(\mathbf{x}^*) = 0$ that is

Local Extrema: Fermat's Theorem

To formally prove this result,

1. Consider the function $g_i(x_i) = f(x_1^*, x_2^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*)$.
2. If f has a local minimum (maximum) at $\mathbf{x}^*$, then there exists an open ball $B_\epsilon = \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ such that for all $\mathbf{x} \in B_\epsilon$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ $(f(\mathbf{x}^*) \geq f(\mathbf{x}))$
3. Consider the norm to be the Eucledian norm $\|.\|_2$. By Cauchy Shwarz inequality, for a unit norm vector $\mathbf{e}_i = [0..1..0]$ with a 1 only in the i^{th} index in the vector, $|\mathbf{e}_i^T(\mathbf{x} - \mathbf{x}^*)| = |x_i - x_i^*| \leq \|\mathbf{x} - \mathbf{x}^*\| \|\mathbf{e}_i\| = \|\mathbf{x} - \mathbf{x}^*\|$.
4. Thus, the existence of an open ball $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ around $\mathbf{x}^*$ characterizing the minimum in $\Re^n$ also guarantees the existence of an open ball (projected ball corresponding to a projected norm) $\{x_i | \|x_i - x_i^*\| < \epsilon\}$ around x_i^* in $\Re$.
5. Therefore each function $g_i(x_i)$ must have a local extremum at x_i^*. Which, by an earlier result (derived for differentiable functions of single argument) implies that $g_i^{'}(x_i^*) = 0$
6. Now $g_i^{'}(x_i^*) = f_{x_i}(\mathbf{x}^*)$ and hence $f_{x_i}(\mathbf{x}^*) = 0$ that is $\nabla f(\mathbf{x}^*) = 0$. □

A Related Property: Convexity by Restricting to Line

A useful technique for verifying the convexity of a function is to investigate its convexity, by restricting the function to a line and checking for the convexity of a function of single variable.

Theorem

A function $f : \mathcal{D} \rightarrow \Re$ is (strictly) convex if and only if the function $\phi : \mathcal{D}_\phi \rightarrow \Re$ defined below, is (strictly) convex in t for every $\mathbf{x} \in \Re^n$ and for every $\mathbf{h} \in \Re^n$

$$\phi(t) = f(\mathbf{x} + t\mathbf{h})$$

with the domain of ϕ given by $\mathcal{D}_\phi = \left\{t \middle| \mathbf{x} + t\mathbf{h} \in \mathcal{D}\right\}$.

Thus, we have see that

- If a function has a local optimum at $\mathbf{x}^*$, it as a local optimum along each component x_i^* of $\mathbf{x}^*$ (that is each canonical direction and in general, local optimum along each line)
- If a function is convex in $\mathbf{x}$, it will be convex in each component x_i of $\mathbf{x}$

(that is each canonical direction and in general along each line)

A Related Property: Convexity by Restricting to Line (contd.)

Proof: We will prove the necessity and sufficiency of the convexity of ϕ for a convex function *f*. The proof for necessity and sufficiency of the strict convexity of ϕ for a strictly convex *f* is very similar and is left as an exercise.

Proof of Necessity: Assume that *f* is convex. And we need to prove that $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is also convex. Let $t_1, t_2 \in \mathcal{D}_\phi$ and $\theta \in [0,1]$. Then,

$$\underline{\phi(\theta t_1 + (1-\theta)t_2)} = f\big(\theta(\mathbf{x} + t_1\mathbf{h}) + (1-\theta)(\mathbf{x} + t_2\mathbf{h})\big)$$
$$\leq \theta f\big((\mathbf{x} + t_1\mathbf{h})\big) + (1-\theta) f\big((\mathbf{x} + t_2\mathbf{h})\big) = \underline{\theta\phi(t_1) + (1-\theta)\phi(t_2)} \qquad (23)$$

Thus, ϕ is convex.

A Related Property: Convexity by Restricting to Line (contd.)

Proof of Sufficiency: Assume that for every $\mathbf{h} \in \Re^n$ and every $\mathbf{x} \in \Re^n$, $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ is convex. We will prove that f is convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$. Take, $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{h} = \mathbf{x}_2 - \mathbf{x}_1$. We know that $\phi(t) = f\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)$ is convex, with $\phi(1) = f(\mathbf{x}_2)$ and $\phi(0) = f(\mathbf{x}_1)$. Therefore, for any $\theta \in [0, 1]$

$$\begin{aligned} f\big(\theta\mathbf{x}_2 + (1-\theta)\mathbf{x}_1\big) &= \phi(\theta) \\ \leq \theta\underline{\phi(1)} + (1-\theta)\underline{\phi(0)} \leq \theta \underline{f(\mathbf{x}_2)} &+ (1-\theta)\underline{f(\mathbf{x}_1)} \end{aligned} \tag{24}$$

This implies that f is convex.

Local Extrema: Illustration

Applying the previous result to the function $f(x_1, x_2) = 9 - x_1^2 - x_2^2$, we require that at any extreme point $\underline{f_{x_1} = -2x_1 = 0} \Rightarrow x_1 = 0$ and $\underline{f_{x_2} = -2x_2 = 0 \Rightarrow} x_2 = 0$. Thus, f indeed attains its maximum at the point $(0, 0)$ as shown in Figure 2.

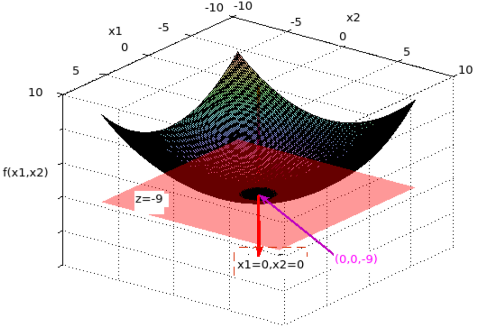


Figure 2:

Critical Point

Definition

[Critical point]: *A point* $\mathbf{x}^*$ *is called a critical point of a function* $f(\mathbf{x})$ *defined on* $\mathcal{D} \subseteq \Re^n$ *if*

1. *If* $f_{x_i}(\mathbf{x}^*) = 0$, *for* $1 \leq i \leq n$.
2. *OR* $f_{x_i}(\mathbf{x}^*)$ *fails to exist for any* $1 \leq i \leq n$.

Critical Point

A procedure for computing all critical points of a function f is:

1. Compute f_{x_i} for $1 \leq i \leq n$.
2. Determine if there are any points where any one of f_{x_i} fails to exist. Add such points (if any) to the list of critical points.
3. Solve the system of equations $f_{x_i} = 0$ simultaneously. Add the solution points to the list of saddle points.

Verify that the point of min we found for Lasso as per sufficient condition is also a critical point...[If any x_i = 0 or if gradient = 0] [H/W]

Critical Point

As an example, for the function $f(x_1, x_2) = |x_1|$, f_{x_1} does not exist for $(0, s)$ for any $s \in \Re$ and all of them are critical points. Figure 3 shows the corresponding $3-$D plot.

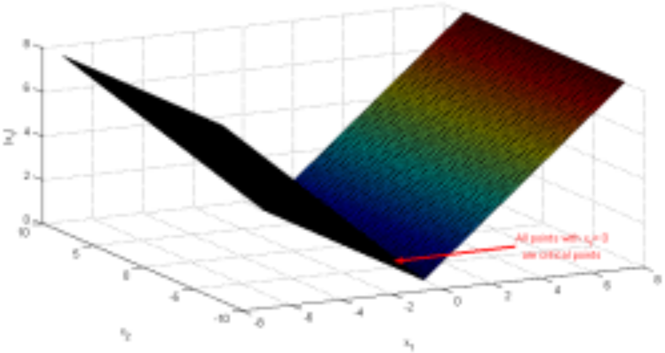


Figure 3:

Saddle Point

Is the converse of the foregoing result true? That is, if you find an $\mathbf{x}^*$ that satisifes $f_{x_i}(\mathbf{x}^*) =$ for all $1 \leq i \leq n$, is it necessary that $\mathbf{x}^*$ is an extreme point? The answer is no. In fact, points that violate the converse of this result are called saddle points.

Definition

[Saddle point]: *A point $\mathbf{x}^*$ is called a saddle point of a function $f(\mathbf{x})$ defined on $\mathcal{D} \subseteq \Re^n$ if $\mathbf{x}^*$ is a critical point of f but $\mathbf{x}^*$ does not correspond to a local maximum or minimum of the function.*

We saw the example of a saddle point in Figure **??**, for the case $n = 1$. The *inflection point* for a function of single variable, that was discussed earlier, is the analogue of the saddle point for a function of multiple variables.

Saddle Point

An example for $n = 2$ is the hyperbolic paraboloid[2] $f(x_1, x_2) = x_1^2 - x_2^2$, the graph of which is shown in Figure 4. The hyperbolic paraboloid has a saddle point at $(0, 0)$.

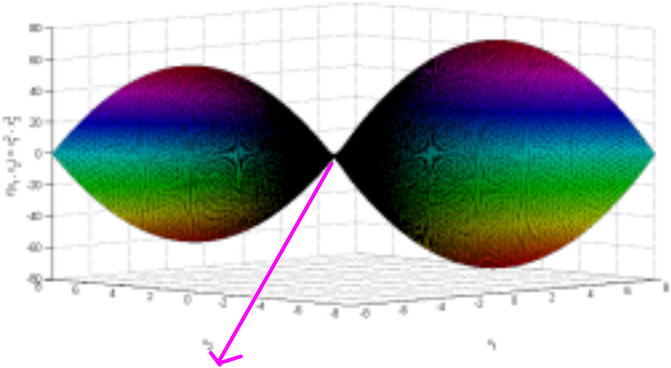


Figure 4:

[2] The hyperbolic paraboloid is shaped like a *saddle* and can have a critical point called the saddle point.

Saddle Point

The hyperbolic paraboloid opens up on x_1-axis (Figure 5):

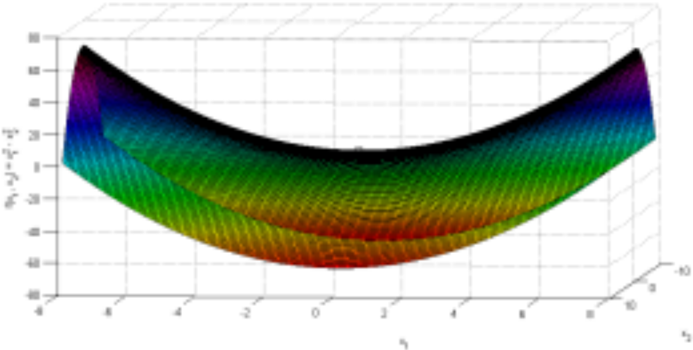
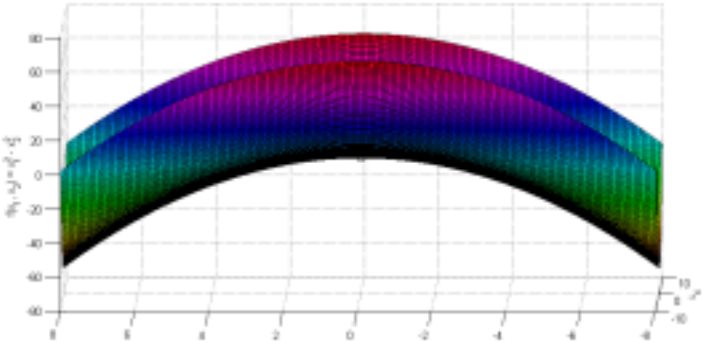


Figure 5:

Saddle Point

The hyperbolic paraboloid opens down on x_2-axis (Figure 6):



Breadthwise view of a horse saddle

Figure 6:

Extreme Points

- To get working on figuring out how to find the maximum and minimum of a function, we will take some examples. Let us find the critical points of $f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1 - 6x_2 + 14$ and classify the critical point.

Extreme Points

- To get working on figuring out how to find the maximum and minimum of a function, we will take some examples. Let us find the critical points of $f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1 - 6x_2 + 14$ and classify the critical point.
- This function is a polyonomial function and is differentiable everywhere. It is a paraboloid that is shifted away from origin. To find its critical points, we will solve $f_{x_1} = 2x_1 - 2 = 0$ and $f_{x_2} = 2x_2 - 6 = 0$, which when solved simultaneously, yield a single critical point $(1, 3)$.
- For a simple example like this, the function f can be rewritten as $f(x_1, x_2) = \underline{(x_1 - 1)^2 + (x_2 - 3)^2 + 4}$, which implies that $f(x_1, x_2) \geq 4 = f(1, 3)$. Therefore, $(1, 3)$ is indeed a local minimum (in fact a global minimum) of $f(x_1, x_2)$.

Descent Algorithms for Optimizing Unconstrained Problems

Techniques relevant for most (convex) optimization problems that do not yield themselves to closed form solutions. We will start with unconstrained minimization.

Goal: (especially of first order descent algos) To achieve a 0 (sub) gradient

$$\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$$

For analysis: (often Lipschitz continuity is also required in analysis)

- Assume that f is convex and differentiable and that it attains a finite optimal value p^*.
- Minimization techniques produce a sequence of points $\mathbf{x}^{(k)} \in \mathcal{D}$, $k = 0, 1, \ldots$ such that $f\left(\mathbf{x}^{(k)}\right) \to p^*$ as $k \to \infty$ or, $\nabla f\left(\mathbf{x}^{(k)}\right) \to \mathbf{0}$ as $k \to \infty$.
- Iterative techniques for optimization, further require a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$ and sometimes that $epi(f)$ is closed. The $epi(f)$ can be inferred to be closed either if $\mathcal{D} = \Re^n$ or $f(\mathbf{x}) \to \infty$ as $\mathbf{x} \to \partial\mathcal{D}$. The function $f(x) = \frac{1}{x}$ for $x > 0$ is an example of a function whose $epi(f)$ is not closed. (Often a function with closed epigraph is itself called closed)

Descent Algorithms [Analysis later, intuitions first]

- Descent methods for minimization have been in use since the last 70 years or more.
- General idea: Next iterate $\mathbf{x}^{(k+1)}$ is the current iterate $\mathbf{x}^{(k)}$ added with a descent or search direction $\Delta\mathbf{x}^{(k)}$ (a unit vector), which is multiplied by a scale factor $t^{(k)}$, called the step length.

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

- The incremental step is determined while aiming that $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$
- We assume that we are dealing with the **extended value extension** $\widetilde{f}$ of the convex function $f : \mathcal{D} \rightarrow \Re$, with $\mathcal{D} \subseteq \Re^n$ which returns ∞ for any point outside its domain. However, if we do so, we need to make sure that the initial point indeed lies in the domain $\mathcal{D}$.

Definition

$$\widetilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{D} \\ \infty & \text{if } \mathbf{x} \notin \mathcal{D} \end{cases} \tag{25}$$

Descent Algorithms

- A single iteration of the general descent algorithm consists of two main steps, *viz.*,
  1. determining a good descent direction $\Delta \mathbf{x}^{(k)}$, which is typically forced to have unit norm and
  2. determining the step size using some line search technique.
- If the function f is convex, from the necessary and sufficient condition for convexity restated here for reference:

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

- We require that $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$ and since $t^{(k)} > 0$, we must have

Homework