- Q-superlinear convergence:

$$\lim_{k\to\infty} \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|} = 0$$

- Q-sublinear convergence:

$$\lim_{k\to\infty} \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|} = 1$$

  - *e.g.* For Lipschitz continuity, v^k in gradient descent is Q-sublinear: $\lim_{k\to\infty} \frac{k}{k+1} = 1$

- Q-convergence of order p:

$$\forall k \geq \theta, \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|^p} \leq M$$

  - *e.g.* $p = 2$ for Q-quadratic, $p = 3$ for Q-cubic, *etc.*
  - M is called the asymptotic error constant

Illustrating Order Convergence

- Consider the two sequences $\mathbf{s}_1$ and $\mathbf{s}_2$.

  $\mathbf{s}_1 = \left[\frac{11}{2}, \frac{21}{4}, \frac{41}{8}, \ldots, 5 + \frac{1}{2^n}, \ldots\right]$

  $\mathbf{s}_2 = \left[\frac{11}{2}, \frac{41}{8}, \frac{641}{128}, \ldots, 5 + \frac{1}{2^{2^n-1}}, \ldots\right]$

  Both sequences converge to 5. However, it seems that the second converges faster to 5 than the first one.
- For $\mathbf{s}_1$, $s_1^* = 5$ and Q-convergence is of order $p = 1$ because:

$$\frac{\left\|s_1^{k+1} - s_1^*\right\|}{\left\|s_1^k - s_1^*\right\|^1} = \frac{\left\|\frac{1}{2^{k+1}}\right\|}{\left\|\frac{1}{2^k}\right\|} = \frac{1}{2} < 0.6 (= M)$$

- For $\mathbf{s}_2$, $s_2^* = 5$ and Q-convergence is of order $p = 2$ because:

$$\frac{\left\|s_2^{k+1} - s_2^*\right\|}{\left\|s_2^k - s_2^*\right\|^2} = \frac{\left\|\frac{1}{2^{2^{k+1}-1}}\right\|}{\left\|\frac{1}{2^{2^k-1}}\right\|^2} = \frac{1}{2} < 0.6 (= M)$$

An algorithm A is faster than algorithm B if
either it has a larger (p) order of convergence
or it has the same order but a lower value of M

- **Claim:** Q-convergences of the order p are special cases of Q-superlinear convergence
- $\forall k \geq \theta$, $\frac{\left\|s^{k+1}-s^*\right\|}{\left\|s^k-s^*\right\|^p} \leq M$

$$\implies \lim_{k\to\infty} \frac{\left\|s^{k+1} - s^*\right\|}{\left\|s^k - s^*\right\|} \leq \lim_{k\to\infty} M\left\|s^k - s^*\right\|^{p-1} = 0$$

- Therefore, irrespective of the value of M (as long as $M \geq 0$), order $p > 1$ implies Q-superlinear convergence

Question: Could we analyze Gradient descent more **specifically**?

- Assume backtracking line search
- Continue assuming Lipschitz continuity
  - Curvature is upper bounded: $\nabla^2 f(x) \preceq LI$
- Assume **strong convexity**
  - Curvature is lower bounded: $\nabla^2 f(x) \succeq mI$
  - For instance, we might not want to use gradient descent for a quadratic function (curvature is not accounted for)

There exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient. That is, with a good understanding of one, we can easily understand the other one. See `http://xingyuzhou.org/talks/Fenchel_duality.pdf` for a quick summary!

(Better) Convergence Using Strong Convexity

Important Aside: Second Order conditions for Convexity, Strong Convexity, Lipschitz Continuity of Gradient, Convex Conjugate, Fenchel Duality.

Second Order Conditions for Convexity

Theorem

A twice differential function $f : \mathcal{D} \to \Re$ *for a nonempty open convex set* $\mathcal{D}$

1. *is convex if and only if its domain is convex and its Hessian matrix is positive semidefinite at each point in* $\mathcal{D}$*. That is* $\nabla^2 f(\mathbf{x}) \succeq 0 \quad \forall\ \mathbf{x} \in \mathcal{D}$
2. *is strictly convex if its domain is convex and its Hessian matrix is positive definite at each point in* $\mathcal{D}$*. That is* $\nabla^2 f(\mathbf{x}) \succ 0 \quad \forall\ \mathbf{x} \in \mathcal{D}$
3. *is uniformly convex if and only if its domain is convex and its Hessian matrix is uniformly positive definite at each point in* $\mathcal{D}$*. That is, for any* $\mathbf{v} \in \Re^n$ *and any* $\mathbf{x} \in \mathcal{D}$*, there exists a* $c > 0$ *such that* $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq c||\mathbf{v}||^2$ Also known as strong convexity

c and m are used interchangeably as the strong convexity factor/constant
Strong convexity of m ==> Atleast m curvature
Lipschitz continuous gradient of L ==> Atmost L curvature

Proof of Second Order Conditions for Convexity

In other words

$$\nabla^2 f(\mathbf{x}) \succeq cI_{n\times n}$$

where $I_{n\times n}$ is the $n \times n$ identity matrix and $\succeq$ corresponds to the positive semidefinite inequality. That is, the function f is strongly convex *iff* $\nabla^2 f(\mathbf{x}) - cI_{n\times n}$ is positive semidefinite, for all $\mathbf{x} \in \mathcal{D}$ and for some constant $c > 0$, which corresponds to the positive minimum curvature of f.

PROOF: We will prove only the first statement; the other two statements are proved in a similar manner.

Necessity: Suppose f is a convex function, and consider a point $\mathbf{x} \in \mathcal{D}$. We will prove that for any $\mathbf{h} \in \Re^n$, $\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$. Since f is convex, we have

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + t\nabla^T f(\mathbf{x})\mathbf{h} \tag{48}$$

Consider the function $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ defined on the domain $\mathcal{D}_\phi = [0, 1]$.

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h})\frac{dx_i}{dt} = \mathbf{h}^T.\nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T\nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h})\frac{dx_i}{dt} = \mathbf{h}^T.\nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

$$f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{x}) + t\mathbf{h}^T \nabla f(\mathbf{x}) + t^2\frac{1}{2}h^T\nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3)$$

Proof of Second Order Conditions for Convexity (contd.)

In conjunction with (48), the above equation implies that

$$\frac{t^2}{2} h^T \nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3) \geq 0$$

Dividing by t^2 and taking limits as $t \to 0$, we get

$$h^T \nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$$

For necessary condition, take limits

Proof of Second Order Conditions for Convexity (contd.)

Sufficiency: Suppose that the Hessian matrix is positive semidefinite at each point $\mathbf{x} \in \mathcal{D}$. Consider the same function $\phi(t)$ defined above with $\mathbf{h} = \mathbf{y} - \mathbf{x}$ for $\mathbf{y}, \mathbf{x} \in \mathcal{D}$. Applying Taylor's theorem with $n = 2$ and $a = 0$, we obtain,

$$\phi(1) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(c)$$

for some $c \in (0, 1)$. Writing this equation in terms of f gives

$$f(\mathbf{x}) = f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{y})$$

where $\mathbf{z} = \mathbf{y} + c(\mathbf{x} - \mathbf{y})$. Since $\mathcal{D}$ is convex, $\mathbf{z} \in \mathcal{D}$. Thus, $\nabla^2 f(\mathbf{z}) \succeq 0$. It follows that

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

By a previous result, the function f is convex. $\square$

Lipschitz Continuity vs. Strong Convexity

- Lipschitz continuity of gradient (references to ∇^2 assume double differentiability)

$$\nabla^2 f(x) \preceq LI$$

$$\left\|\nabla f(x) - \nabla f(y)\right\| \leq L\|x - y\|$$

$$f(y) \leq f(x) + \nabla^\top f(x)(y - x) + \frac{L}{2}\|y - x\|^2$$

- Strong convexity: Curvature should be **atleast somewhat** positive

$$\nabla^2 f(x) \succeq mI$$

$$f(y) \geq f(x) + \nabla^\top f(x)(y - x) + \frac{m}{2}\|y - x\|^2$$

  - $m = 0$ corresponds to (sufficient condition for) normal convexity.
  - Later: For example, augmented Lagrangian is used to introduce strong convexity

Conjugate Functions

- Recall from Lecture 14 the (Young's) inequality for scalars $h, x \in \Re$ and for $p, q \in \Re^+$ such that for $\frac{1}{p} + \frac{1}{q} = 1$:

Conjugate Functions

- Recall from Lecture 14 the (Young's) inequality for scalars $h, x \in \Re$ and for $p, q \in \Re^+$ such that for $\frac{1}{p} + \frac{1}{q} = 1$: $hx \leq \frac{x^p}{p} + \frac{h^q}{q}$
- In other words: $\frac{h^q}{q} \geq hx - \frac{x^p}{p}$
- The RHS $hx - \frac{x^p}{p}$ viewed as a function of x, is maximized at point x at which

Conjugate Functions

- Recall from Lecture 14 the (Young's) inequality for scalars $h, x \in \Re$ and for $p, q \in \Re^+$ such that for $\frac{1}{p} + \frac{1}{q} = 1$: $hx \leq \frac{x^p}{p} + \frac{h^q}{q}$
- In other words: $\frac{h^q}{q} \geq hx - \frac{x^p}{p}$
- The RHS $hx - \frac{x^p}{p}$ viewed as a function of x, is maximized at point x at which $\frac{d\frac{x^p}{p}}{x} = h$, that is at $x^{p-1} = h$
- Note that, under this condition, $h^q =$

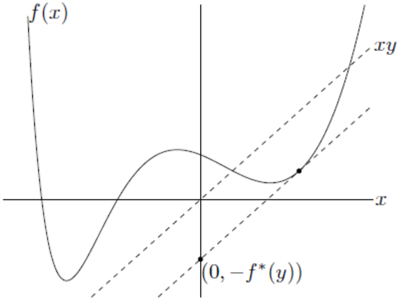
Conjugate Functions

- Recall from Lecture 14 the (Young's) inequality for scalars $h, x \in \Re$ and for $p, q \in \Re^+$ such that for $\frac{1}{p} + \frac{1}{q} = 1$: $hx \leq \frac{x^p}{p} + \frac{h^q}{q}$
- In other words: $\frac{h^q}{q} \geq hx - \frac{x^p}{p}$
- The RHS $hx - \frac{x^p}{p}$ viewed as a function of x, is maximized at point x at which $\frac{d\frac{x^p}{p}}{x} = h$, that is at $x^{p-1} = h$
- Note that, under this condition, $h^q = x^{q(p-1)} = x^p$ (since $\frac{1}{p} + \frac{1}{q} = 1$) and the inequality becomes an equality
- That is, if $f(x) = \frac{x^p}{p}$ and $f^*(h) = \frac{h^q}{q}$ then

Conjugate Functions

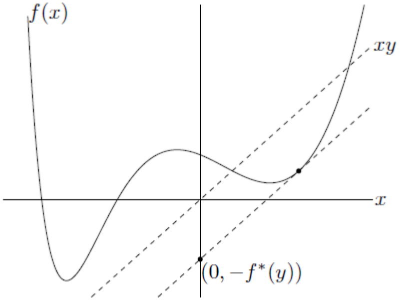
- Recall from Lecture 14 the (Young's) inequality for scalars $h, x \in \Re$ and for $p, q \in \Re^+$ such that for $\frac{1}{p} + \frac{1}{q} = 1$: $hx \leq \frac{x^p}{p} + \frac{h^q}{q}$
- In other words: $\frac{h^q}{q} \geq hx - \frac{x^p}{p}$
- The RHS $hx - \frac{x^p}{p}$ viewed as a function of x, is maximized at point x at which $\frac{d\frac{x^p}{p}}{x} = h$, that is at $x^{p-1} = h$
- Note that, under this condition, $h^q = x^{q(p-1)} = x^p$ (since $\frac{1}{p} + \frac{1}{q} = 1$) and the inequality becomes an equality
- That is, if $f(x) = \frac{x^p}{p}$ and $f^*(h) = \frac{h^q}{q}$ then $f^*(h) \geq hx - f(x)$ and equality is attained when $f'(x) = h$

Conjugate Functions



- That is, if $f(x) = \frac{x^p}{p}$ and $f^*(h) = \frac{h^q}{q}$ then f*(h) >= hx - f(x)

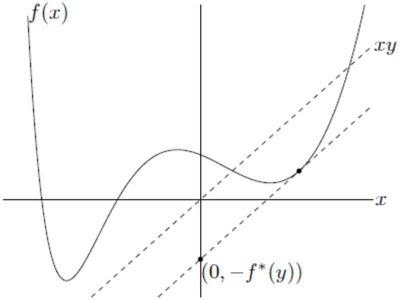
Conjugate Functions



- That is, if $f(x) = \frac{x^p}{p}$ and $f^*(h) = \frac{h^q}{q}$ then $f^*(h) \geq hx - f(x)$ and equality is attained when $f'(x) = h$. These observations can be generalized:

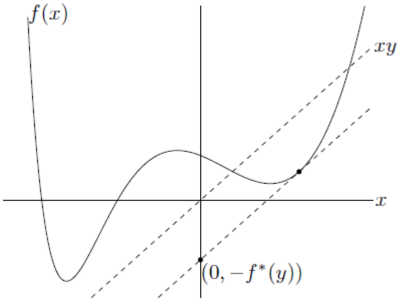
  f*(h) = supremum over x of hx-f(x)
  and
  hx <= f(x) + f*(h) otherwise

Conjugate Functions



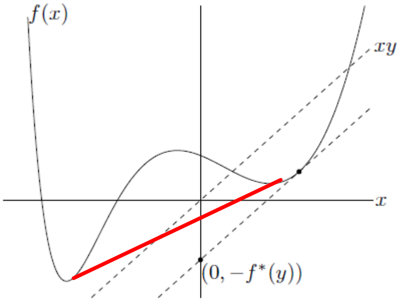
- That is, if $f(x) = \frac{x^p}{p}$ and $f^*(h) = \frac{h^q}{q}$ then $f^*(h) \geq hx - f(x)$ and equality is attained when $f'(x) = h$. These observations can be generalized:
- Conjugate Function of $f : \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x} \in \mathcal{D}} (\mathbf{h}^T\mathbf{x} - f(\mathbf{x}))$
- Fenchel inequality: $\mathbf{h}^T\mathbf{x} \leq f(\mathbf{x}) + f^*(\mathbf{h})$ or $f^*(\mathbf{h}) \geq \mathbf{h}^T\mathbf{x} - f(\mathbf{x})$
- The conjugate function $f^*(y)$ is the maximum gap between the linear function yx and $f(x)$, as shown by the dashed line in the figure. If f is differentiable, this occurs at a point x where $f'(x) = h$

Conjugate and Conjugate of the Conjugate



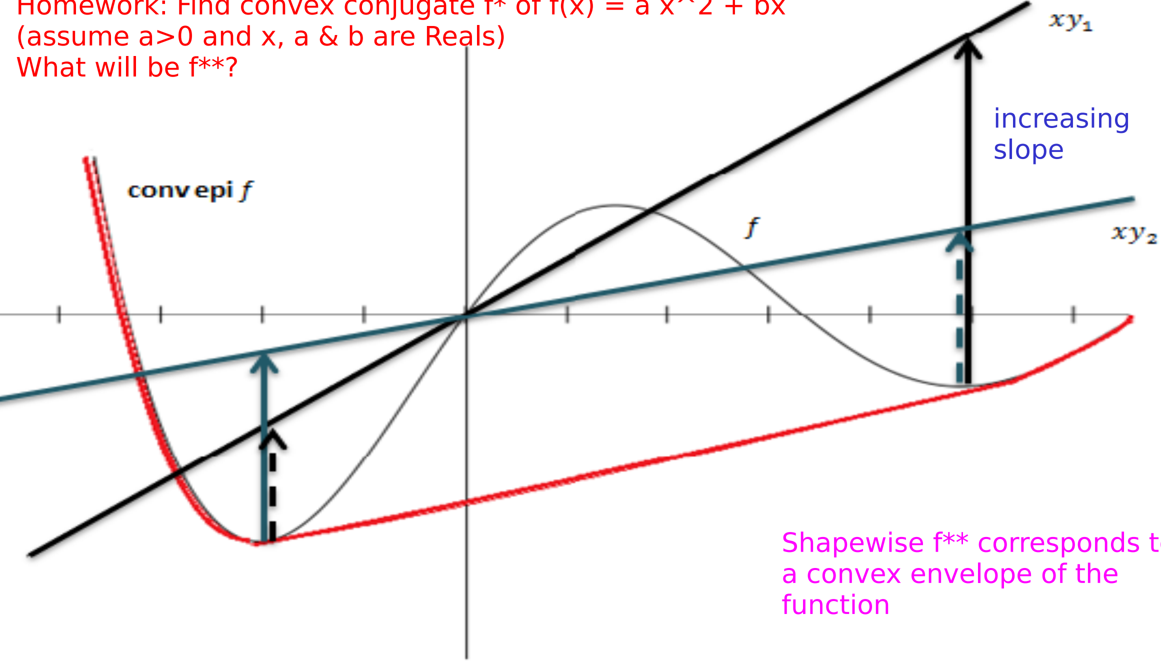
- Conjugate Function of $f: \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x} \in \mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- Even if f is not convex (and closed): f* = pointwise supremum of affine functions

Conjugate and Conjugate of the Conjugate



- Conjugate Function of $f : \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x} \in \mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- Even if f is not convex (and closed): f^* is convex (since it is pointwise suprememum of affine functions) and closed
- How about $f^{**}(\mathbf{x})$? f** is the convex envelope of f

Homework: Find convex conjugate f* of f(x) = a x^2 + bx
(assume a>0 and x, a & b are Reals)
What will be f**?



Shapewise f** corresponds to a convex envelope of the function

Conjugate Functions, Strong Convexity and Lipschitz Continuity

- Conjugate Function of $f : \mathcal{D} \rightarrow \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x} \in \mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- Fenchel inequality: $\mathbf{h}^T\mathbf{x} \leq f(\mathbf{x}) + f^*(\mathbf{h})$
- Eg:

Conjugate Functions, Strong Convexity and Lipschitz Continuity

- Conjugate Function of $f : \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x}\in\mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- Fenchel inequality: $\mathbf{h}^T\mathbf{x} \leq f(\mathbf{x}) + f^*(\mathbf{h})$
- Eg: $f(\mathbf{x}) = \frac{x^p}{p}$ and $f^*(\mathbf{h}) = \frac{h^q}{q}$ for $\frac{1}{p} + \frac{1}{q} = 1$
- $\nabla f^*(\mathbf{h}) = \operatorname{argmax}_{\mathbf{x}\in\mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$

Conjugate Functions, Strong Convexity and Lipschitz Continuity

- Conjugate Function of $f : \mathcal{D} \to \Re$: $f^*(\mathbf{h}) = \sup_{\mathbf{x}\in\mathcal{D}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- Fenchel inequality: $\mathbf{h}^T\mathbf{x} \leq f(\mathbf{x}) + f^*(\mathbf{h})$
- Eg: $f(\mathbf{x}) = \frac{x^p}{p}$ and $f^*(\mathbf{h}) = \frac{h^q}{q}$ for $\frac{1}{p} + \frac{1}{q} = 1$
- $\nabla f^*(\mathbf{h}) = \underset{\mathbf{x}\in\mathcal{D}}{\text{argmax}} \left(\mathbf{h}^T\mathbf{x} - f(\mathbf{x})\right)$
- If f is closed and strongly convex with parameter m, then f^* has a Lipschitz continuous gradient with parameter $1/m$. convex f atleast m curved => Lipshitz f* atmost 1/m curved
- If f is convex and has a Lipschitz continuous gradient with parameter L, then f^* is strongly convex with parameter $1/L$ Lipschitz gradient f atmost L curved => convex f* atleast 1/L curved

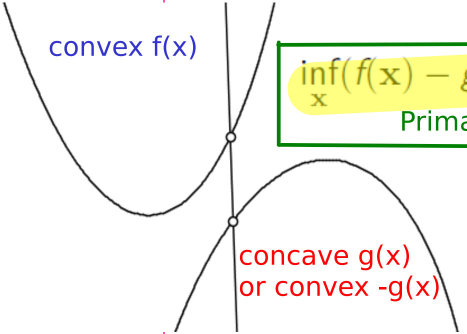
There exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient.

Fenchel Duality, Strong Convexity and Lipschitz Continuity

- Let f be a closed convex function on $\Re^n$ and let g be a closed concave function on $\Re^n$. Then, under some general conditions:

$$\inf_{\mathbf{x}}(f(\mathbf{x}) - g(\mathbf{x})) = \sup_{\mathbf{h}}(g^*(\mathbf{h}) - f^*(\mathbf{h}))$$

  where f^* is the convex conjugate of f and g^* is the concave conjugate of g
- Thus, there exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient. That is, with a good understanding of one, we can easily understand the other one. See `http://xingyuzhou.org/talks/Fenchel_duality.pdf` for a quick summary!

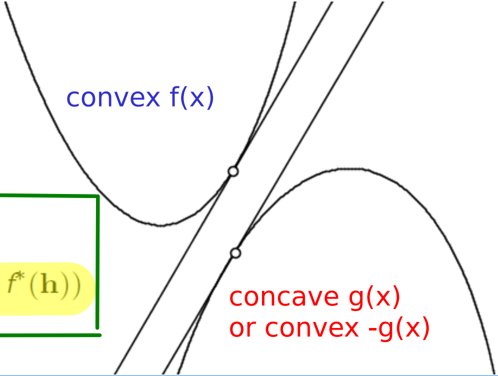


$$\inf_{\mathbf{x}} (f(\mathbf{x}) - g(\mathbf{x}))$$

Primal: Find x that gives smalles gap between f and g

Dual: Find slope h that gives largest gap between g* and f*

$$\sup_{\mathbf{h}} (g^*(\mathbf{h}) - f^*(\mathbf{h}))$$



Lipschitz Continuity vs. Strong Convexity: Example

- Consider the linear regression loss function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|^2$
- $\nabla f(\mathbf{x}) = -A^T(\mathbf{y} - A\mathbf{x})$
- $\nabla^2 f(\mathbf{x}) = A^T A$
- One can show that

Max and min eigenvalues of A^T A
characterize strong convexity and
Lipschitz continuity respective

Lipschitz Continuity vs. Strong Convexity: Example

- Consider the linear regression loss function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|^2$
- $\nabla f(\mathbf{x}) = -A^T(\mathbf{y} - A\mathbf{x})$
- $\nabla^2 f(\mathbf{x}) = A^T A$
- One can show that
  - $\nabla^2 f(\mathbf{x}) = A^T A \preceq LI$ where $L = \sigma_{max}$ is the largest eigenvalue of $A^T A$
  - $\nabla^2 f(\mathbf{x}) = A^T A \succeq mI$ where $m = \sigma_{min}$ is the smallest eigenvalue of $A^T A$

L/m puts some bound on the condition number of the Hessian

End of Important Aside: Second Order conditions for Convexity, Strong Convexity, Lipschitz Continuity of Gradient, Convex Conjugate, Fenchel Duality.

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|^2$
  $\geq$ minimum value of RHS wrt y

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|^2$
  $\geq$ minimum value the RHS can take as a function of y
- Minimum value of RHS
  $\nabla f(\mathbf{x}) + m\mathbf{y} - m\mathbf{x} = 0$
  $\implies y = x - \frac{1}{m}\nabla f(\mathbf{x})$
- Thus,

Using Strong Convexity: Revisiting Convergence Analysis

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|^2$
  $\geq$ minimum value the RHS can take as a function of y
- Minimum value of RHS
  $\nabla f(\mathbf{x}) + m\mathbf{y} - m\mathbf{x} = 0$
  $\implies y = x - \frac{1}{m}\nabla f(\mathbf{x})$
- Thus,
  $$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})\left(-\frac{1}{m}\nabla f(\mathbf{x})\right) + \frac{m}{2}\left\|-\frac{1}{m}\nabla f(\mathbf{x})\right\|^2$$
  $$\implies f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$$
  - Here, LHS is independent of $\mathbf{x}$, and RHS is independent of $\mathbf{y}$
  - Thus the inequality holds also for $\mathbf{y} = \mathbf{x}^*$ (point of minimum of $f(\mathbf{x})$)

Using Strong Convexity: Revisiting Convergence Analysis (contd.)

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|^2$$

- If $\left\|\nabla f(\mathbf{x})\right\|$ is small, the point is nearly optimal
  - If $\left\|\nabla f(\mathbf{x})\right\| \leq \sqrt{2m\epsilon}$, then:
    $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon$
  - As the gradient $\left\|\nabla f(\mathbf{x})\right\|$ approaches 0, we get closer to the optimal solution $\mathbf{x}^*$