Constrained optimality and duality

Antonio Frangioni

Department of Computer Science
University of Pisa
www.di.unipi.it/~frangio
frangio@di.unipi.it

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

Outline

Constrained optimization

- Finally back to the full (P) $f_* = \min\{\, f(x) \;:\; x \in X \,\}$
- "$x \in X$" $\equiv$ constraints, whence constrained optimization (doh!)
- Explicitly describe X, as opposed to "hide it in $\imath_X$"
- $x \in X$ feasible solution, $x \notin X$ unfeasible solution (doh!)
- Global optimum in general hard to find (unless everything convex)
- Usual weaker notion: x_* is local optimum if it solves

$$\min\{\, f(x) \;:\; x \in \mathcal{B}(x_*, \varepsilon) \cap X \,\} \qquad \text{for some } \varepsilon > 0$$

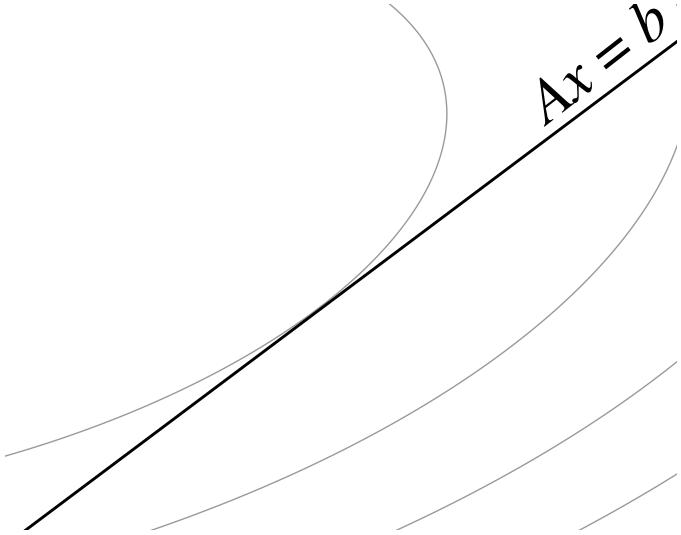
- As usual, strict local optimum if $f(x) < f(y) \;\; \forall\, y \in \mathcal{B}(x_*, \varepsilon) \cap X$
- $x_* \in \mathit{int}\, X \implies$ local minimum $\equiv$ local optimum
- Constrained (local) optimality conditions $\neq$ only on ∂X
- Need to "describe how ∂X looks around x_*"

Outline

Linear equality constraints

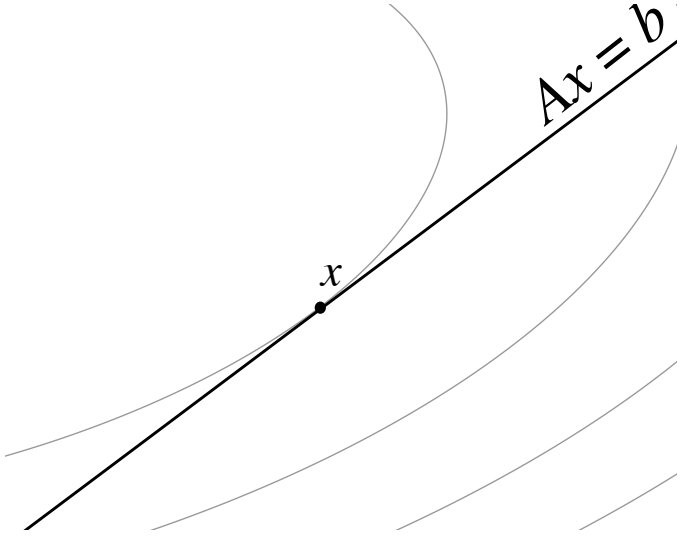
- Simple case: Linear equality constraints (P) $\min\{ f(x) : Ax = b \}$
  $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = m < n$, rows of A linearly independent (**why?**)
- $x \in X \equiv x \in \partial X \wedge$ "∂X looks the same everywhere"
- (P) $\equiv$ unconstrained problem: $A = [A_B, A_N]$, $x = [x_B, x_N]$
  $det(A_B) \neq 0 \Longrightarrow Ax = b \equiv x_B = A_B^{-1}(b - A_N x_N) \Longrightarrow$
  $$D = \begin{bmatrix} -A_B^{-1} A_N \\ I \end{bmatrix}, \; d = \begin{bmatrix} A_B^{-1} b \\ 0 \end{bmatrix}, \; (P) \equiv \min_{w \in \mathbb{R}^{n-m}} \{ r(w) = f(Dw + d) \}$$
- m linear constraints kill m degrees of freedom $\equiv m$ variables
- Reduced gradient $\nabla r(w) = D^T \nabla f(Dw + d)$: solve $\nabla r(w^*) = 0$
- Note: $D^T A^T = 0$ (**check**) $\Longrightarrow z = \mu A \Longrightarrow D^T z = 0$
- "Poorman's KKT conditions": $Ax = b \wedge \exists \mu$ s.t. $\mu A = \nabla f(x) \Longrightarrow$
  x is a "stationary point" for (P) $(x \equiv w \wedge \nabla r(w) = 0)$
- μ first example of dual variables: to prove x optimal you have to find μ
- f convex $\Longrightarrow$ (P-KKT) sufficient for optimality (**why?**)

Linear equality constraints, geometrically



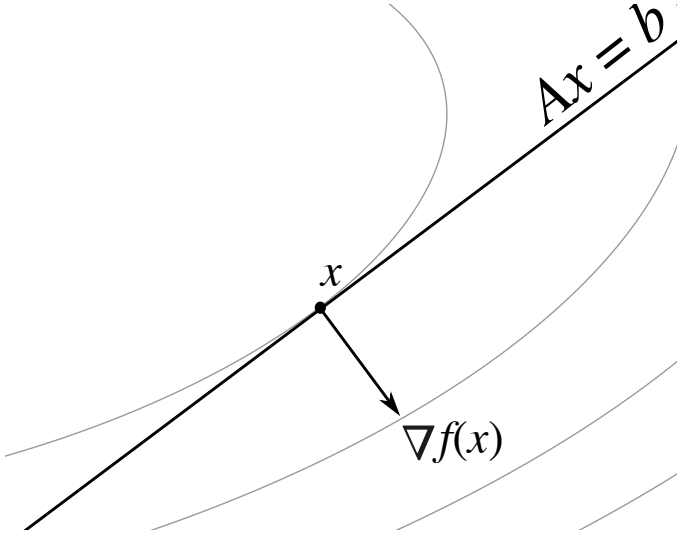
- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$

Linear equality constraints, geometrically



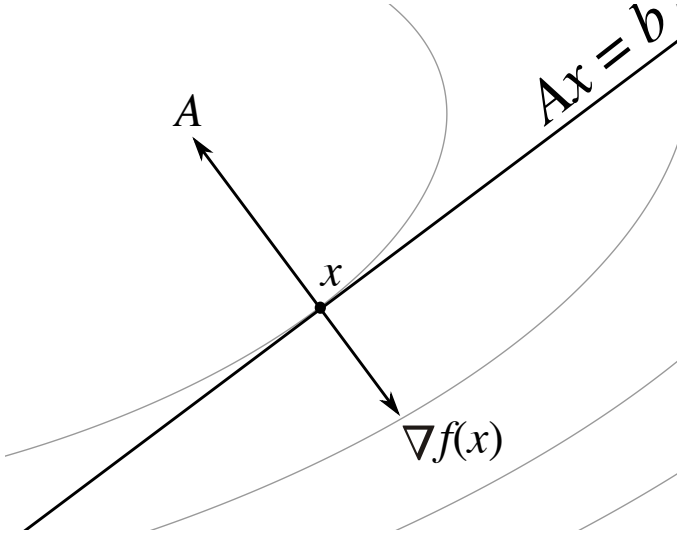
- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set

Linear equality constraints, geometrically



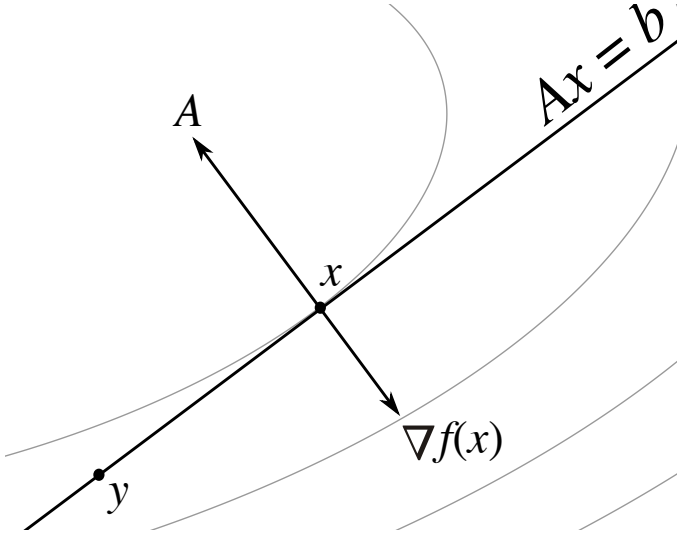
- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set
- $\nabla f(x) \perp \partial X \equiv$

Linear equality constraints, geometrically



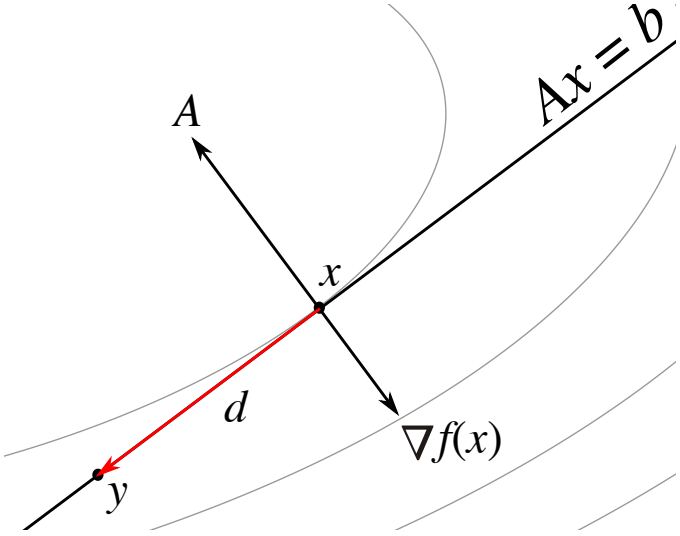
- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set
- $\nabla f(x) \perp \partial X \equiv$ "$\nabla f(x) \parallel A$"
  since $A \perp \partial X$ by definition

Linear equality constraints, geometrically



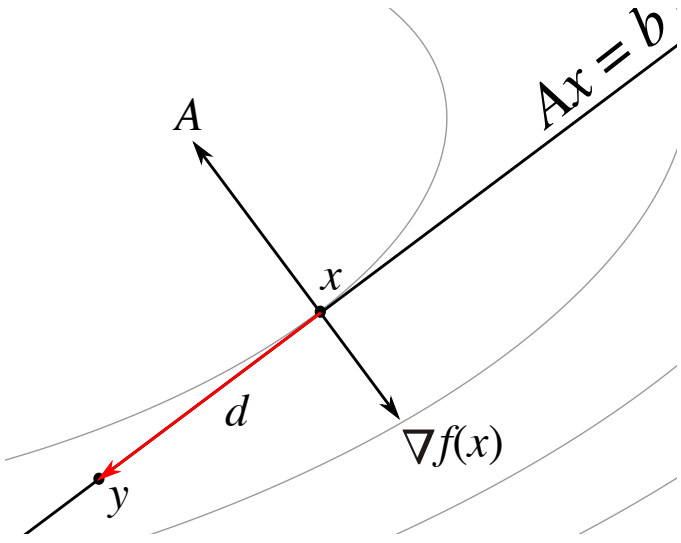
- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set
- $\nabla f(x) \perp \partial X \equiv$ "$\nabla f(x) \parallel A$"
  since $A \perp \partial X$ by definition
- In fact, $\forall\, y \in X$

Linear equality constraints, geometrically



- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set
- $\nabla f(x) \perp \partial X \equiv$ "$\nabla f(x) \parallel A$"
  since $A \perp \partial X$ by definition
- In fact, $\forall\, y \in X\ y - x = d \perp A$ (**check**)

Linear equality constraints, geometrically



- $S(f, \cdot)$ and $X = \{ x \in \mathbb{R}^n : Ax = b \}$
- optimum touches inner level set
- $\nabla f(x) \perp \partial X \equiv$ "$\nabla f(x) \parallel A$"
  since $A \perp \partial X$ by definition
- In fact, $\forall\, y \in X \;\; y - x = d \perp A$ (**check**)
- "$\nabla f(x) \parallel A$" $\equiv \nabla f(x) \in range(A)$
  $\equiv \nabla f(x) = \mu A$

- $F = \{ d \in \mathbb{R}^n : Ad = 0 \}$ feasible directions:
  cannot use any $d \notin F$ to move away from x (any $x \in X$)
- $\nabla f(x) = 0$ not necessary, just $\frac{\partial f}{\partial d}(x) = \langle \nabla f(x), d \rangle = 0 \; \forall d \in F$
- May have expected "$\langle \nabla f(x), d \rangle \geq 0$", but $d \in F \Longrightarrow -d \in F \Longrightarrow$
  to get "$\geq 0 \; \forall d$" must ask $= 0$

Exercise: solve $\min\{ x^2 + y^2 + z^2 : x + z = 1 \;\; , \;\; x + y - z = 2 \}$ by reducing it to a univariate unconstrained problem

Outline

The Tangent Cone

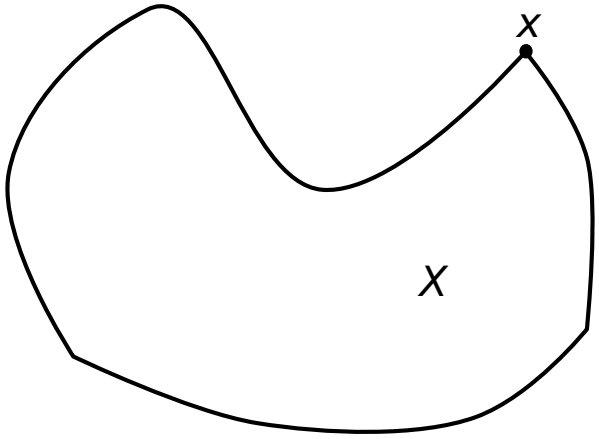
- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$

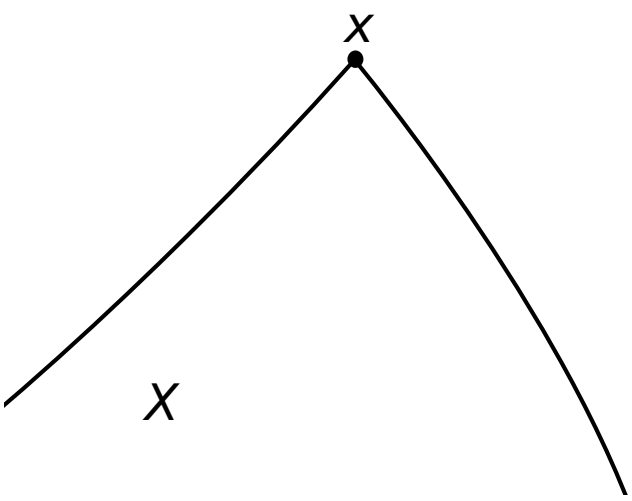


- Er . . . what? Simpler than what it seems.

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$

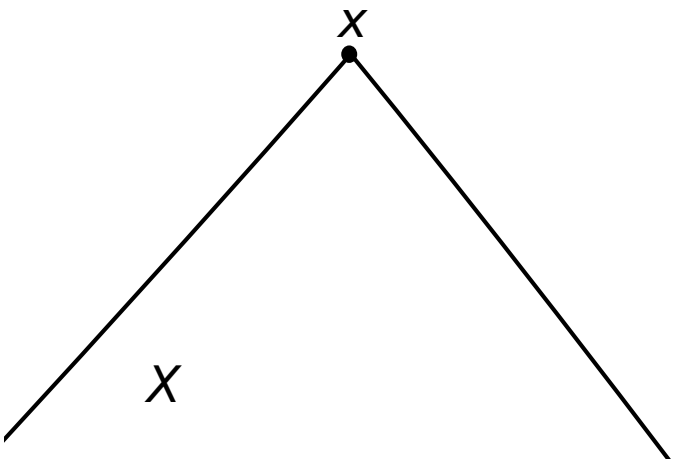


- Er . . . what? Simpler than what it seems.
- Zoom to x

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$

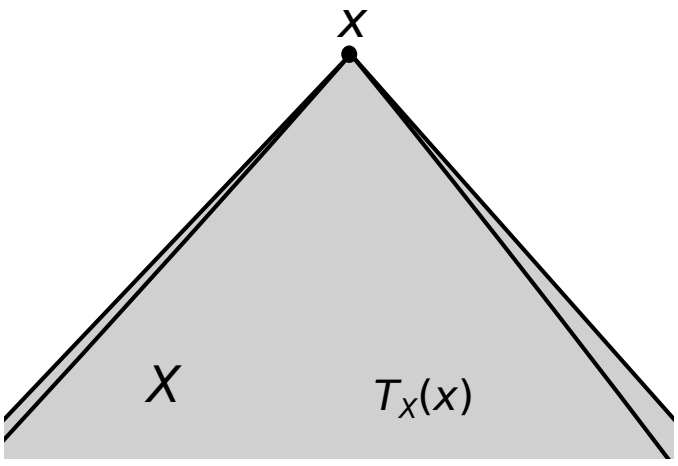


- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$

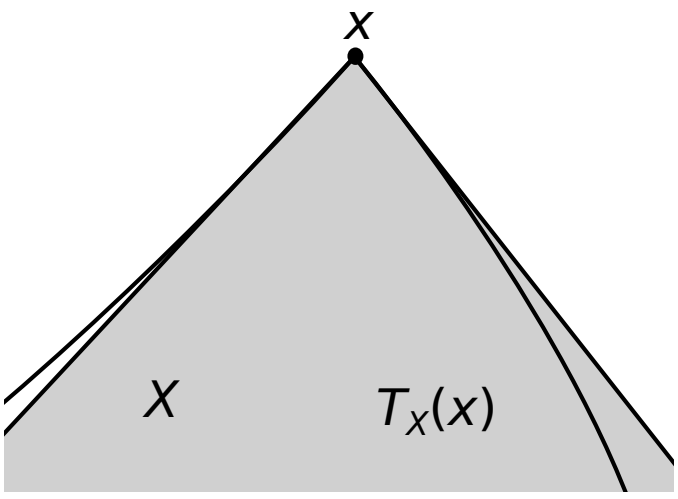


- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then X looks a cone:

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$



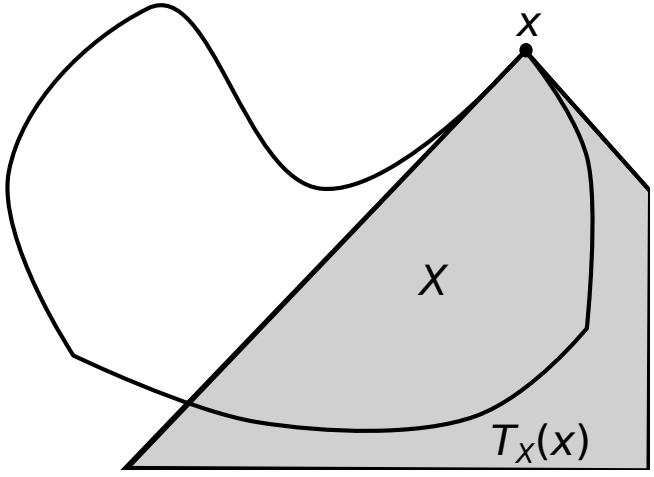
- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then

  X looks a cone: zoom out,

The Tangent Cone

- Crucial object: $T_X(x)$ = tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \ : \ \exists \left\{ z_i \in X \right\} \to x \wedge \left\{ t_i \geq 0 \right\} \to 0 \text{ s.t. } d = \lim_{i \to \infty} (z_i - x) / t_i \right\}$$



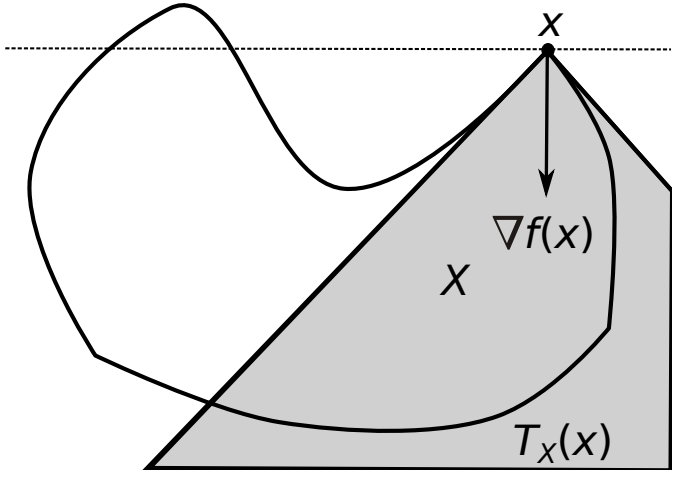
- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then

  X looks a cone: zoom out, this is $T_X(x)$
- $\mathcal{C}$ cone: $x \in \mathcal{C} \Longrightarrow \alpha x \in \mathcal{C} \ \forall \, \alpha > 0$

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

  $\{\, d \in \mathbb{R}^n \;:\; \exists \{\, z_i \in X \,\} \to x \wedge \{\, t_i \geq 0 \,\} \to 0 \text{ s.t. } d = \lim_{i\to\infty} (\, z_i - x \,) / \, t_i \,\}$



- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then

  X looks a cone: zoom out, this is $T_X(x)$
- $\mathcal{C}$ cone: $x \in \mathcal{C} \implies \alpha x \in \mathcal{C} \;\forall\, \alpha > 0$
- x local optimum $\implies$ (note: $\not\Longleftarrow$)

  $\langle\, \nabla f(x) \,,\, d \,\rangle \geq 0 \;\forall\, d \in T_X(x)$

- **Proof.** $\exists\, d \in T_X(x)$ s.t. $\langle\, \nabla f(x) \,,\, d \,\rangle < 0 \implies d = \lim_{i\to\infty} (\, z_i - x \,) / \, t_i$

  First-order Taylor: $f(z_i) - f(x) = \langle\, \nabla f(x) \,,\, z_i - x \,\rangle + R(\, z_i - x \,)$

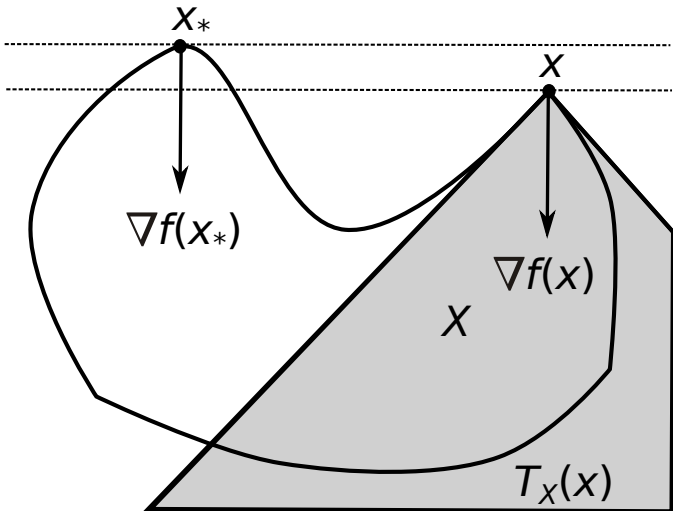
  Divide by t_i, take $\lim_{i\to\infty}$ and use $\lim_{i\to\infty} (z_i - x) - t_i d = 0$:

  $\lim_{i\to\infty} (\, f(z_i) - f(x) \,)/t_i = \langle\, \nabla f(x) \,,\, d \,\rangle + \lim_{i\to\infty} R(\, z_i - x \,)/t_i < 0$ ↯

The Tangent Cone

- Crucial object: $T_X(x)$ = tangent cone of X at x

$$\left\{ d \in \mathbb{R}^n \;:\; \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i\to\infty} (z_i - x) / t_i \right\}$$



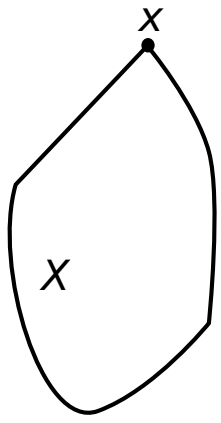
- Er . . . what? Simpler than what it seems.
- Zoom to x very closely: then
  X looks a cone: zoom out, this is $T_X(x)$
- $\mathcal{C}$ cone: $x \in \mathcal{C} \implies \alpha x \in \mathcal{C} \; \forall \, \alpha > 0$
- x local optimum $\implies$ (note: $\not\Longleftarrow$)
  $\langle \nabla f(x) , d \rangle \geq 0 \; \forall \, d \in T_X(x)$

- **Proof.** $\exists \, d \in T_X(x)$ s.t. $\langle \nabla f(x) , d \rangle < 0 \implies d = \lim_{i\to\infty} (z_i - x) / t_i$
  First-order Taylor: $f(z_i) - f(x) = \langle \nabla f(x) , z_i - x \rangle + R(z_i - x)$
  Divide by t_i, take $\lim_{i\to\infty}$ and use $\lim_{i\to\infty}(z_i - x) - t_i d = 0$:
  $\lim_{i\to\infty}(f(z_i) - f(x))/t_i = \langle \nabla f(x) , d \rangle + \lim_{i\to\infty} R(z_i - x)/t_i < 0$ ↯
- Obviously, does not mean it is a global optimum

The Tangent Cone

- Crucial object: $T_X(x) =$ tangent cone of X at x

  $\left\{ d \in \mathbb{R}^n \ : \ \exists \{ z_i \in X \} \to x \wedge \{ t_i \geq 0 \} \to 0 \text{ s.t. } d = \lim_{i\to\infty} (z_i - x) / t_i \right\}$



  - Er . . . what? Simpler than what it seems.
  - Zoom to x very closely: then

    X looks a cone: zoom out, this is $T_X(x)$
  - $\mathcal{C}$ cone: $x \in \mathcal{C} \implies \alpha x \in \mathcal{C} \ \forall \, \alpha > 0$
  - x local optimum $\implies$ (note: $\not\Longleftarrow$)

    $\langle \nabla f(x) , d \rangle \geq 0 \ \forall \, d \in T_X(x)$

- **Proof.** $\exists \, d \in T_X(x)$ s.t. $\langle \nabla f(x) , d \rangle < 0 \implies d = \lim_{i\to\infty} (z_i - x) / t_i$

  First-order Taylor: $f(z_i) - f(x) = \langle \nabla f(x) , z_i - x \rangle + R(z_i - x)$

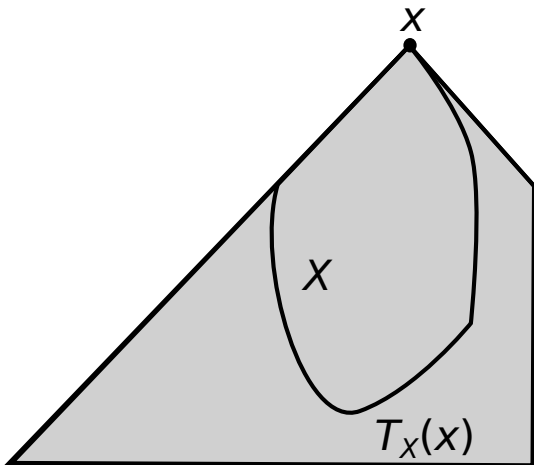
  Divide by t_i, take $\lim_{i\to\infty}$ and use $\lim_{i\to\infty}(z_i - x) - t_i d = 0$:

  $\lim_{i\to\infty}(f(z_i) - f(x))/t_i = \langle \nabla f(x) , d \rangle + \lim_{i\to\infty} R(z_i - x)/t_i < 0$ ↯
- Obviously, does not mean it is a global optimum
- Unless X convex, because then

The Tangent Cone

- Crucial object: $T_X(x)$ = tangent cone of X at x

  $\{\, d \in \mathbb{R}^n \;:\; \exists \{\, z_i \in X \,\} \to x \wedge \{\, t_i \geq 0 \,\} \to 0 \text{ s.t. } d = \lim_{i\to\infty} (\, z_i - x \,) / \, t_i \,\}$



  - Er . . . what? Simpler than what it seems.
  - Zoom to x very closely: then

    X looks a cone: zoom out, this is $T_X(x)$
  - $\mathcal{C}$ cone: $x \in \mathcal{C} \implies \alpha x \in \mathcal{C} \;\forall\, \alpha > 0$
  - x local optimum $\implies$ (note: $\not\Longleftarrow$)

    $\langle\, \nabla f(x) \,,\, d \,\rangle \geq 0 \;\forall\, d \in T_X(x)$

- **Proof.** $\exists\, d \in T_X(x)$ s.t. $\langle\, \nabla f(x) \,,\, d \,\rangle < 0 \implies d = \lim_{i\to\infty} (\, z_i - x \,) / \, t_i$

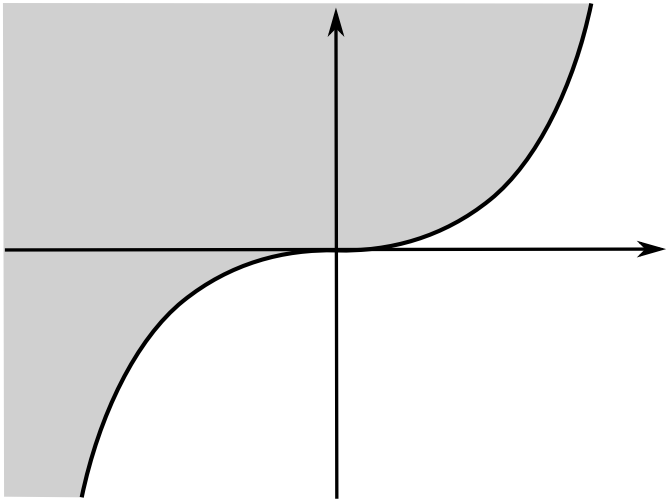
  First-order Taylor: $f(z_i) - f(x) = \langle\, \nabla f(x) \,,\, z_i - x \,\rangle + R(\, z_i - x \,)$

  Divide by t_i, take $\lim_{i\to\infty}$ and use $\lim_{i\to\infty} (z_i - x) - t_i d = 0$:

  $\lim_{i\to\infty} (\, f(z_i) - f(x) \,)/t_i = \langle\, \nabla f(x) \,,\, d \,\rangle + \lim_{i\to\infty} R(\, z_i - x \,)/t_i < 0$ ↯
- Obviously, does not mean it is a global optimum
- Unless X convex, because then $X \subseteq x + T_X(x)$

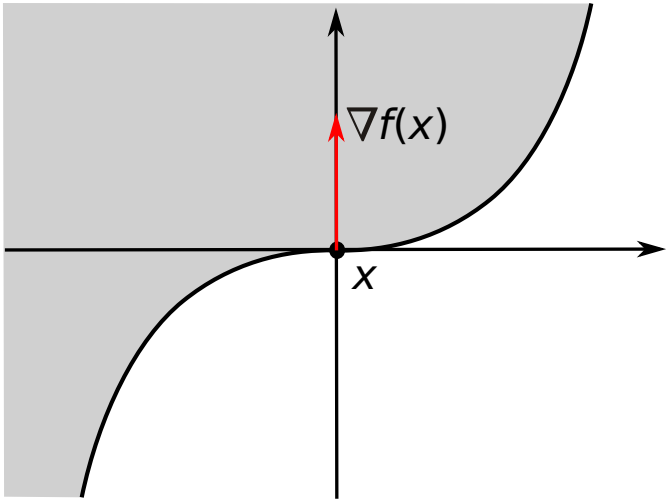
Exercise: prove i) $T_X(x)$ is a cone, ii) the last statement

Why it is not sufficient



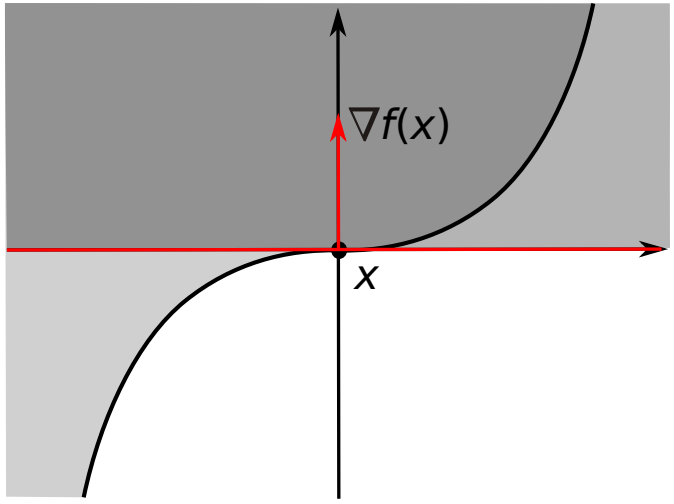
- Counter-example: $\min\{\, x_2 \;:\; x_2 \geq x_1^3 \,\}$

Why it is not sufficient



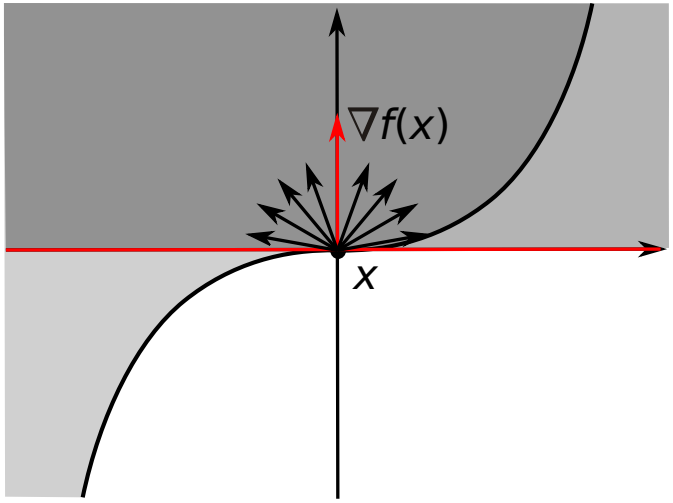
- Counter-example: $\min\{\, x_2 \;:\; x_2 \geq x_1^3 \,\}$
- $x = [\,0\,,\,0\,]$, $\nabla f(\,x\,) = [\,0\,,\,1\,]$

Why it is not sufficient



- Counter-example: $\min\{\, x_2 \,:\, x_2 \geq x_1^3 \,\}$
- $x = [\,0\,,\,0\,]$, $\nabla f(\,x\,) = [\,0\,,\,1\,]$
- $T_X(\,x\,) = \{\,[\,x_1\,,\,x_2\,] \,:\, x_2 \geq 0\,\}$
- $\langle\, \nabla f(\,x\,)\,,\, d\,\rangle \geq 0 \;\; \forall\, d \in T_X(\,x\,)$, but x not minimum (**why?**)
- Clearly due to nonconvexity: x a "saddle point of ∂X"

Why it is not sufficient



- Counter-example: $\min\{\, x_2 \,:\, x_2 \geq x_1^3 \,\}$
- $x = [\,0\,,\,0\,]$, $\nabla f(\,x\,) = [\,0\,,\,1\,]$
- $T_X(\,x\,) = \{\,[\,x_1\,,\,x_2\,] \,:\, x_2 \geq 0\,\}$
- $\langle\, \nabla f(\,x\,)\,,\, d\,\rangle \geq 0 \;\; \forall\, d \in T_X(\,x\,)$, but x not minimum (**why?**)
- Clearly due to nonconvexity: x a "saddle point of ∂X"

- Cone of feasible directions of X at x:

$$F_X(x) = \left\{\, d \in \mathbb{R}^n \;:\; \exists\, \bar{\varepsilon} > 0 \text{ s.t. } x + \varepsilon d \in X \;\; \forall \varepsilon \in [\,0\,,\,\bar{\varepsilon}\,] \,\right\}$$

Exercise: provide alternative definition of F_X when X is convex

- Properties: i) T_X closed, F_X in general not; ii) $cl\, F_X \subseteq T_X$; iii) X convex $\Longrightarrow$ both T_X and F_X convex and $cl\, F_X = T_X$

Exercise: provide example of open F_X; prove $F_X \subseteq T_X$

- Consequence: f and X convex $\Longrightarrow$

$$x \text{ global optimum} \iff \langle\, \nabla f(x)\,,\, d\,\rangle \geq 0 \;\forall\, d \in T_X(x)$$

Exercise: prove the last statement

Outline

Explicit description of constraints

- How to characterize T_X? It depends on how you characterize X
- Usual (but not only) form: explicit description of constraints

  $X = \{ x \in \mathbb{R}^n : g_i(x) \leq 0 \;\; i \in \mathcal{I}, \, h_j(x) = 0 \;\; j \in \mathcal{J} \}$

  $= \{ x \in \mathbb{R}^n : G(x) \leq 0, \, H(x) = 0 \}$
- $\mathcal{I} =$ set of inequality constraints, $\mathcal{J} =$ set of equality constraints
- $G = [\, g_i(x) \,]_{i \in \mathcal{I}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{I}|}$, $H = [\, h_i(x) \,]_{i \in \mathcal{J}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{J}|}$
- Note: $h_j(x) = 0 \equiv h_j(x) \leq 0 \wedge -h_j(x) \leq 0$, but good reasons to explicitly consider equalities when there are
- Note: can always assume $|\mathcal{I}| = 1$ by $g(x) = \max\{ g_i(x) : i \in \mathcal{I} \} \leq 0$

  $\Longrightarrow$ could always assume $|\mathcal{I}| = 1$ and $|\mathcal{J}| = 0$

  but good reasons not to do that

Exercise: how would you ensure $|\mathcal{I}| = 1$ and $|\mathcal{J}| = 0$?

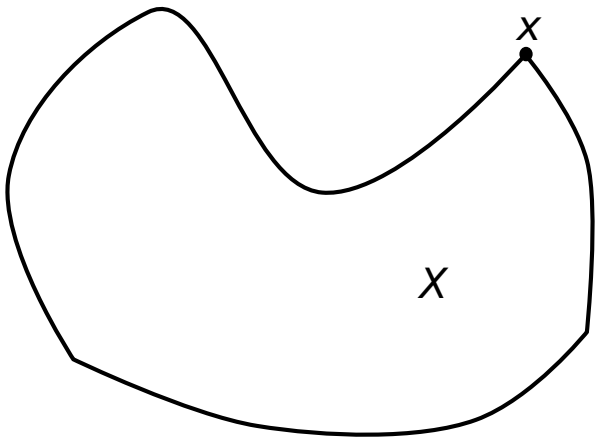
- Active constraints at $x \in X$: $\mathcal{A}(x) = \{ i \in \mathcal{I} : \; g_i(x) = 0 \} \subseteq \mathcal{I}$
- $\mathcal{B} \subseteq \mathcal{I} \Longrightarrow G_{\mathcal{B}} = [\, g_i(x) \,]_{i \in \mathcal{B}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{B}|}$

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



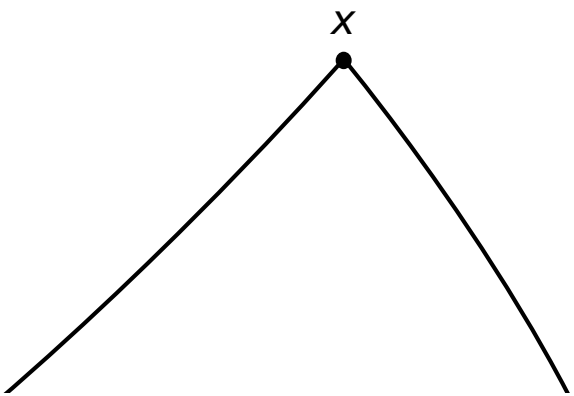
First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$

  - $\mathcal{A}(x) \equiv$ zoom very close to X

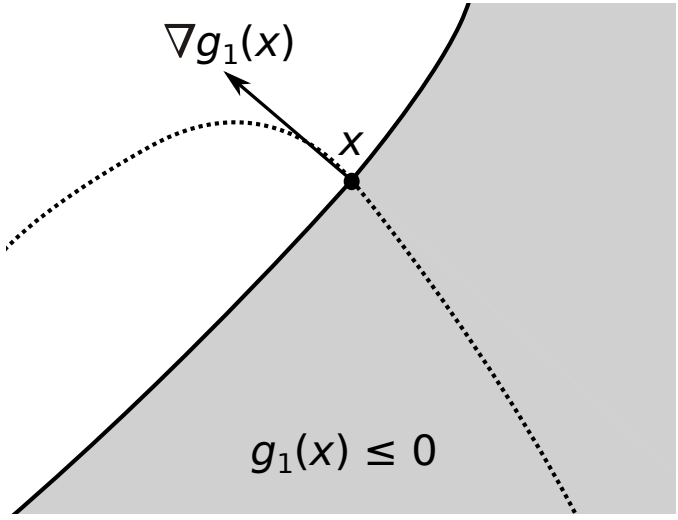


First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

$$\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$$

$$\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$$



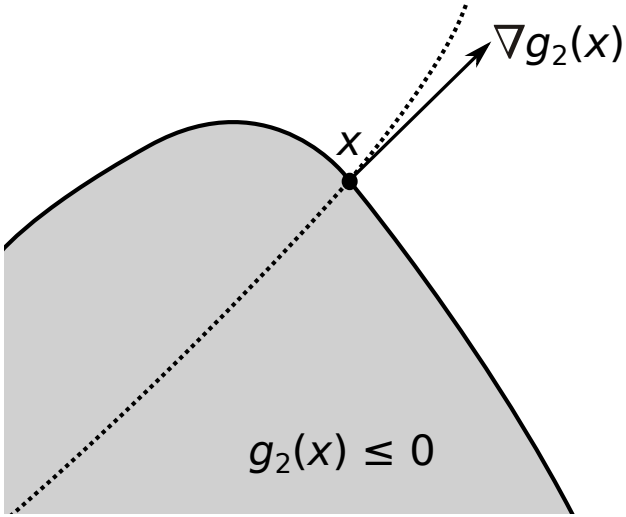
- $\mathcal{A}(x) \equiv$ zoom very close to X
- Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
- $\nabla g_i(x) \perp \partial X$ at x

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



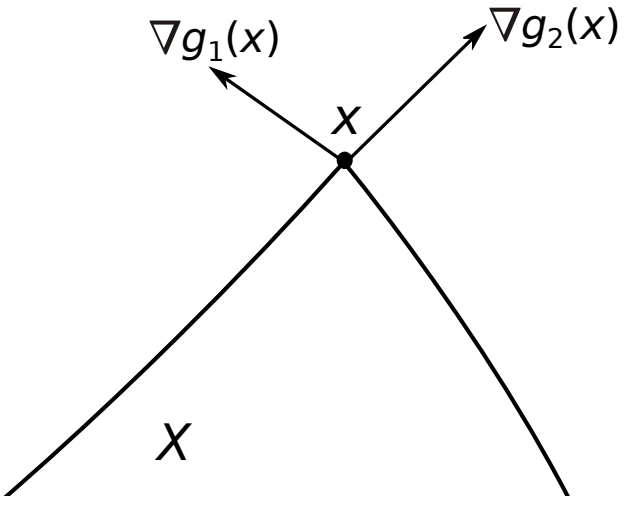
- $\mathcal{A}(x) \equiv$ zoom very close to X
- Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
- $\nabla g_i(x) \perp \partial X$ at x
- Each one separately $\Longrightarrow$

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

$$\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$$

$$\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$$



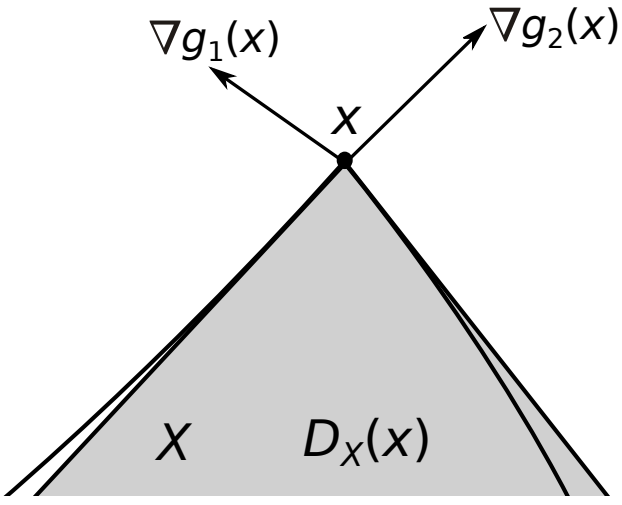
- $\mathcal{A}(x) \equiv$ zoom very close to X
- Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
- $\nabla g_i(x) \perp \partial X$ at x
- Each one separately $\Longrightarrow$ intersection

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



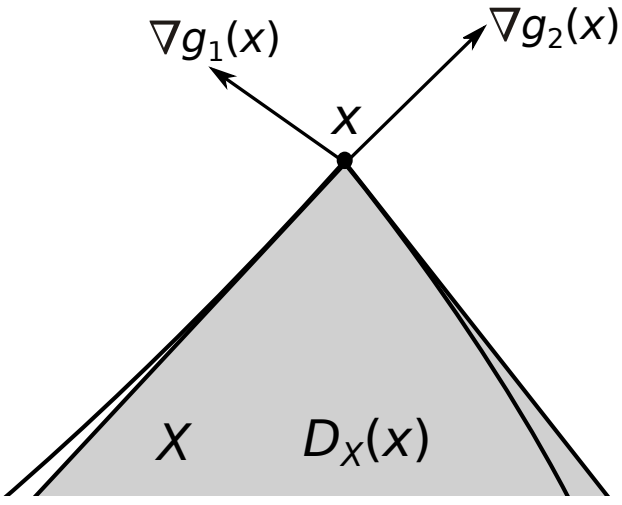
- $\mathcal{A}(x) \equiv$ zoom very close to X
- Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
- $\nabla g_i(x) \perp \partial X$ at x
- Each one separately $\Longrightarrow$ intersection
- $D_X(x)$ looks a lot like $T_X(x)$, in fact

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



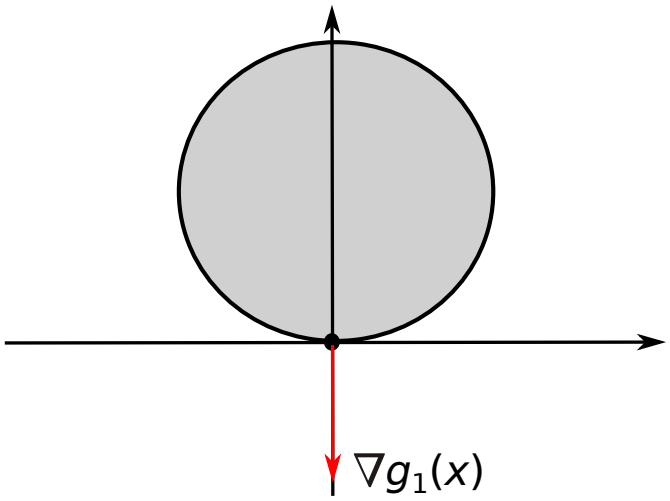
- $\mathcal{A}(x) \equiv$ zoom very close to X
- Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
- $\nabla g_i(x) \perp \partial X$ at x
- Each one separately $\Longrightarrow$ intersection
- $D_X(x)$ looks a lot like $T_X(x)$, in fact $T_X(x) \subseteq D_X(x)$

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



  - $\mathcal{A}(x) \equiv$ zoom very close to X
  - Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
  - $\nabla g_i(x) \perp \partial X$ at x
  - Each one separately $\Longrightarrow$ intersection
  - $D_X(x)$ looks a lot like $T_X(x)$, in fact $T_X(x) \subseteq D_X(x)$

- $D_X(x)$ can be larger than $T_X(x)$ in pathological cases:

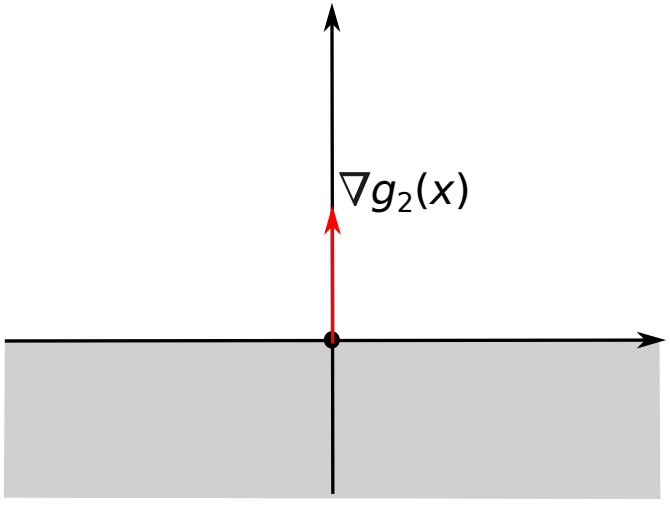
  $\min\{\,\ldots \,:\, x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,$

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$

  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$

  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



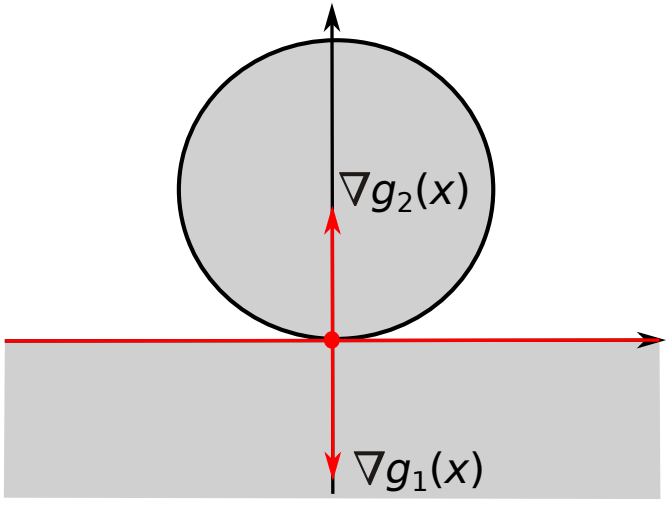
  - $\mathcal{A}(x) \equiv$ zoom very close to X
  - Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
  - $\nabla g_i(x) \perp \partial X$ at x
  - Each one separately $\Longrightarrow$ intersection
  - $D_X(x)$ looks a lot like $T_X(x)$, in fact $T_X(x) \subseteq D_X(x)$

- $D_X(x)$ can be larger than $T_X(x)$ in pathological cases:

  $\min\{\,\ldots\,:\, x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,\;\; x_2 \leq 0$

First-order feasible direction cone

- First-order feasible direction cone at $x \in X$: $D_X(x) =$
  $\{\, d \in \mathbb{R}^n \,:\, \langle\, \nabla g_i(x)\,,\, d\,\rangle \leq 0 \;\; i \in \mathcal{A}(x)\,,\, \langle\, \nabla h_j(x)\,,\, d\,\rangle = 0 \;\; j \in \mathcal{J}\,\} =$
  $\{\, d \in \mathbb{R}^n \,:\, (\, JG_{\mathcal{A}(x)}(x)\,)d \leq 0\,,\, (\, JH(x)\,)d = 0\,\}$



  - $\mathcal{A}(x) \equiv$ zoom very close to X
  - Each $i \in \mathcal{A}(x)$ defines "a part of ∂X"
  - $\nabla g_i(x) \perp \partial X$ at x
  - Each one separately $\Longrightarrow$ intersection
  - $D_X(x)$ looks a lot like $T_X(x)$, in fact
    $T_X(x) \subseteq D_X(x)$

- $D_X(x)$ can be larger than $T_X(x)$ in pathological cases:
  $\min\{\,\ldots\,:\, x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,\;\; x_2 \leq 0\,\}$, $x = [\,0\,,\,0\,]$
- $D_X(\,x\,) = \{\,[\,x_1\,,\,x_2\,]\,:\, x_2 = 0\,\}$, $T_X(\,x\,) = F_X(\,x\,) = \{\,[\,0\,,\,0\,]\,\}$

Exercise: Check the counter-example in details

- A very stupid way to write $X = \{\,[\,0\,,\,0\,]\,\}$, have to avoid it
- Note that everything is convex, so convexity won't help this time

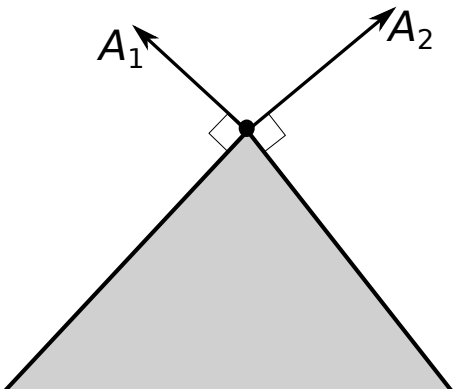
Constraint qualifications

- Avoid pathological cases $\equiv D_X(x) = T_X(x)$
- Several conditions known $\equiv$ constraint qualifications:
  a) Affine constraints (AffC): g_i and h_j affine $\forall\, i \in \mathcal{I}$ and $j \in \mathcal{J} \Longrightarrow$
  $$T_X(x) = D_X(x) \;\forall\, x \in X$$
  b) Slater's condition (SlaC): g_i convex $\forall\, i \in \mathcal{I}$, h_j affine $\forall\, j \in \mathcal{J}$
  $\exists\, \bar{x} \in X$ s.t. $g_i(\bar{x}) < 0\; \forall\, i \in \mathcal{I} \Longrightarrow \quad T_X(x) = D_X(x) \;\forall\, x \in X$
  c) Linear independence (LinI): $\bar{x} \in X \wedge$ the vectors
  $$\{\, \nabla g_i(\bar{x}) \;:\; i \in \mathcal{A}(\bar{x}) \,\} \;\cup\; \{\, \nabla h_j(\bar{x}) \;:\; j \in \mathcal{J} \,\}$$
  linearly independent $\Longrightarrow T_X(\bar{x}) = D_X(\bar{x})$
- Weaker form of (SlaC): $g_i(\bar{x}) < 0\; \forall i \in \mathcal{I}$ not affine $\equiv$
  "in the interior of the feasible region of the nonlinear inequalities"
- Our counter-example fail all three (doh!)
- OK, so excluding pathological cases, necessary condition $\equiv$
  $$\langle\, \nabla f(x)\,,\, d \,\rangle \geq 0 \quad \forall\, d \in D_X(x)$$
- How do I check something like this? $\forall\, d \;\ldots$??

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{ d \in \mathbb{R}^n : Ad \leq 0 \}$ for some $A \in \mathbb{R}^{k \times n}$

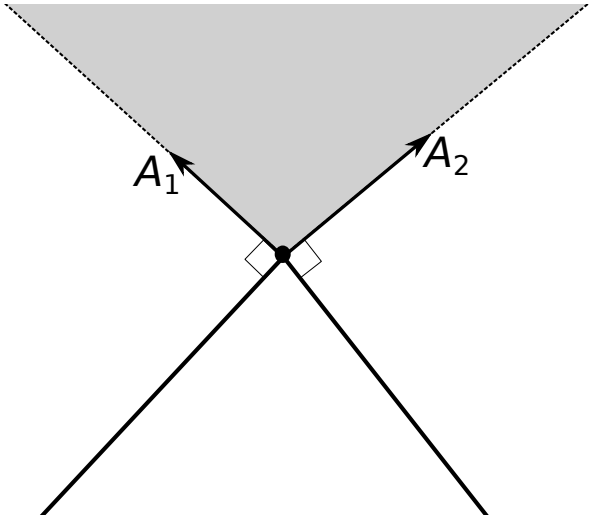
  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)



Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$

  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)

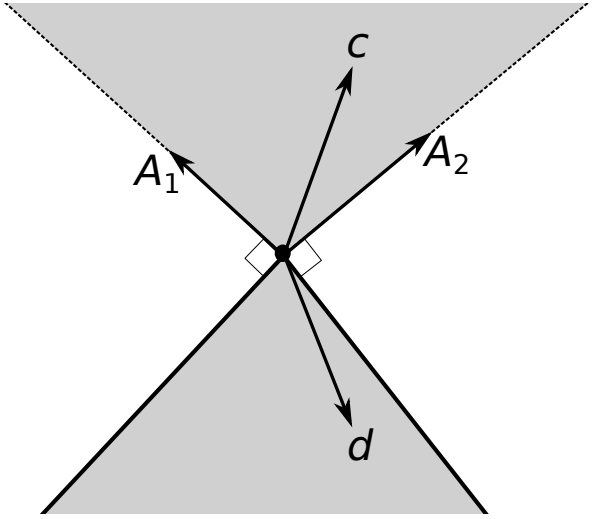


- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$

  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)



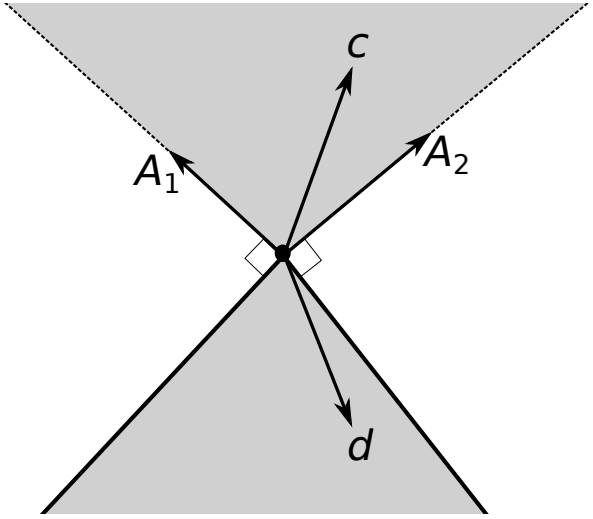
- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
- $\langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\; c \in \mathcal{C}^*$ (**check**)
- A $\neq$ definition: polar cone

  $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$

  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)

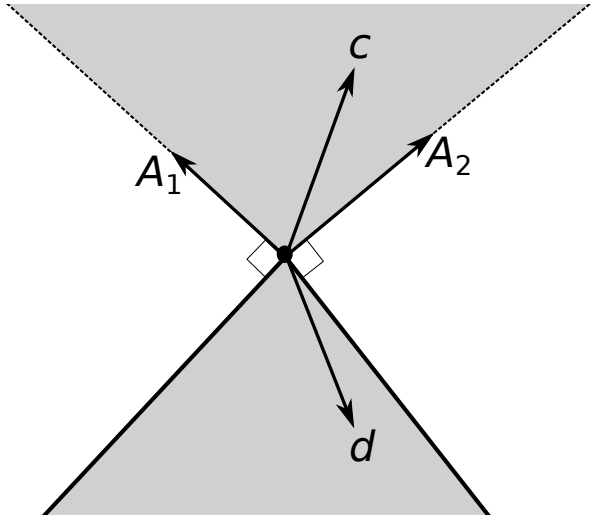


- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
- $\langle\, c \,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\; c \in \mathcal{C}^*$ (**check**)
- A $\neq$ definition: polar cone
  $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c \,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$
- Farkas' lemma: either $c \in \mathcal{C}^*$, or $c \notin \mathcal{C}^*$

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$

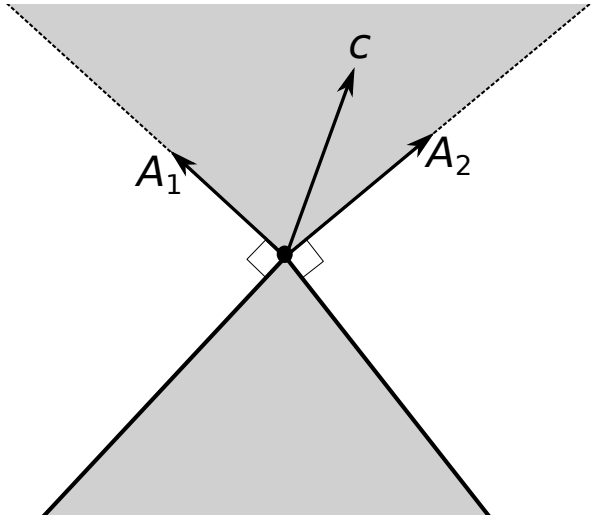
  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)



- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
- $\langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\; c \in \mathcal{C}^*$ (**check**)
- A $\neq$ definition: polar cone
  $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$
- Farkas' lemma: either $c \in \mathcal{C}^*$, or $c \notin \mathcal{C}^*$
- Er . . . doh! Why do I care??

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$
  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)

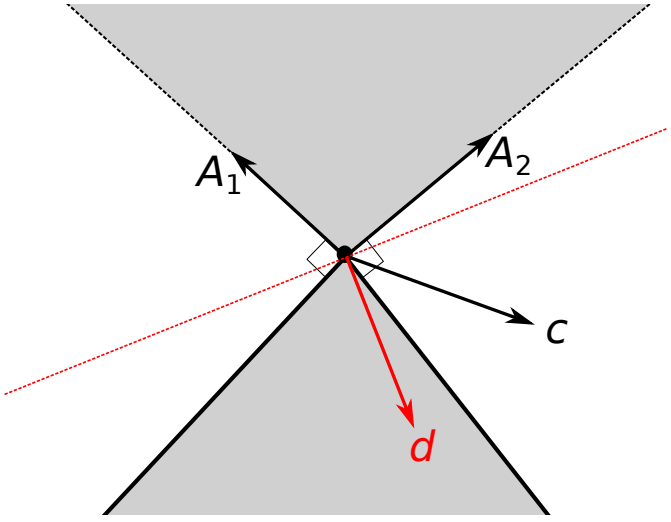


  - Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
  - $\langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\; c \in \mathcal{C}^*$ (**check**)
  - A $\neq$ definition: polar cone
    $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c\,,\, d \,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$
  - Farkas' lemma: either $c \in \mathcal{C}^*$, or $c \notin \mathcal{C}^*$
  - Er . . . doh! Why do I care??

- Either $\exists\, \lambda \geq 0$ s.t. $c = \sum_{i=1}^{k} \lambda_i A_i$

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$
  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{J}$?**)

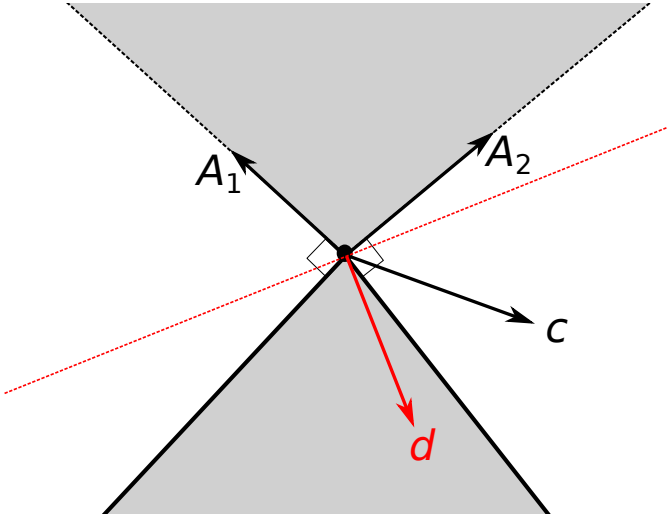


- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^{k} \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
- $\langle\, c\,,\, d\,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\, c \in \mathcal{C}^*$ (**check**)
- A $\neq$ definition: polar cone
  $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c\,,\, d\,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$
- Farkas' lemma: either $c \in \mathcal{C}^*$, or $c \notin \mathcal{C}^*$
- Er . . . doh! Why do I care??

- Either $\exists\, \lambda \geq 0$ s.t. $c = \sum_{i=1}^{k} \lambda_i A_i$ or $\exists\, d$ s.t. $Ad \leq 0 \wedge \langle\, c\,,\, d\,\rangle > 0$

Farkas' lemma

- D_X is a polyhedral cone: $\mathcal{C} = \{\, d \in \mathbb{R}^n \,:\, Ad \leq 0 \,\}$ for some $A \in \mathbb{R}^{k \times n}$
  "very close by, ∂X looks like a polyhedron" (**what happened to $\mathcal{T}$?**)



- Dual cone: $\mathcal{C}^* = \{\, c = \sum_{i=1}^k \lambda_i A_i \,:\, \lambda \geq 0 \,\}$
- $\langle\, c\,,\, d\,\rangle \leq 0 \;\forall\, d \in \mathcal{C},\; c \in \mathcal{C}^*$ (**check**)
- A $\neq$ definition: polar cone
  $\mathcal{C}^\circ = \{\, c \in \mathbb{R}^n \,:\, \langle\, c\,,\, d\,\rangle \leq 0 \;\forall\, d \in \mathcal{C} \,\} = \mathcal{C}^*$
- Farkas' lemma: either $c \in \mathcal{C}^*$, or $c \notin \mathcal{C}^*$
- Er . . . doh! Why do I care??

- Either $\exists\, \lambda \geq 0$ s.t. $c = \sum_{i=1}^k \lambda_i A_i$ or $\exists\, d$ s.t. $Ad \leq 0 \wedge \langle\, c\,,\, d\,\rangle > 0$
- Hence, $\exists\, \lambda \geq 0$ s.t. $-c = \sum_{i=1}^k \lambda_i A_i \Longrightarrow \langle\, c\,,\, d\,\rangle \geq 0 \;\forall\, d$ s.t. $Ad \leq 0$
- Constructive way to prove $\langle\, \nabla f(x)\,,\, d\,\rangle \geq 0 \;\forall\, d \in D_X(x)$: find λ
  $\Longrightarrow$ the celebrated Karush-Kuhn-Tucker conditions

Exercise: Farkas' lemma is the simplest separation result: $x \notin X$ (convex)
$\Longrightarrow \exists$ an hyperplane that separates x from X. Why is this true?

The Karush-Kuhn-Tucker conditions

- Karush-Kuhn-Tucker conditions: $x \in X$, $\exists\, \lambda \in \mathbb{R}_+^{|\mathcal{I}|}$ and $\mu \in \mathbb{R}^{|\mathcal{I}|}$ s.t.

$$\nabla f(x) + \sum_{i \in \mathcal{I}} \lambda_i \nabla g_i(x) + \sum_{j \in \mathcal{J}} \mu_j \nabla h_j(x) = 0 \qquad \text{(KKT-G)}$$

$$\sum_{i \in \mathcal{I}} \lambda_i g_i(x) = 0 \qquad \text{(KKT-CS)}$$

- KKT Theorem: $T_X(x) = D_X(x) \wedge x$ local optimum $\Longrightarrow$ (KKT)
- CS = Complementary Slackness $\equiv \lambda_i g_i(x) = 0 \;\forall\, i \in \mathcal{I}$ (**why?**)

Exercise: Prove KKT Theorem. Where does (CS) come from? Why $\mu \not\geq 0$?

- $x \in X \equiv g_i(x) \leq 0 \;\; i \in \mathcal{I} \;\;,\;\; h_j(x) = 0 \;\; j \in \mathcal{J}$ (doh!)
- Optimization $\equiv$ solving systems of nonlinear equations and inequalities

The Karush-Kuhn-Tucker conditions

- Karush-Kuhn-Tucker conditions: $x \in X$, $\exists\, \lambda \in \mathbb{R}_+^{|\mathcal{I}|}$ and $\mu \in \mathbb{R}^{|\mathcal{I}|}$ s.t.

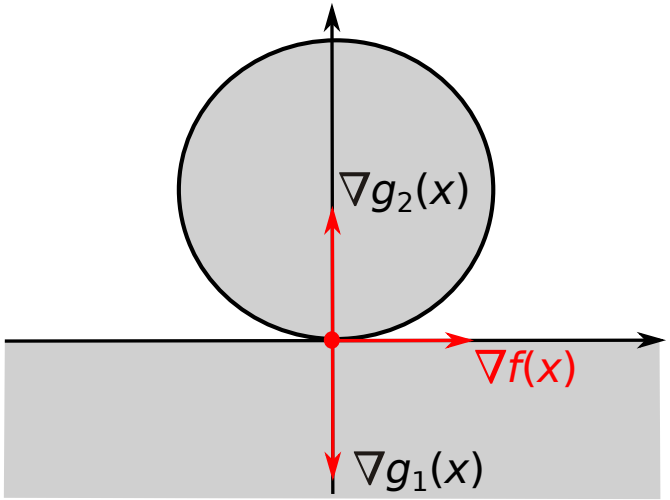
$$\nabla f(x) + \sum_{i\in\mathcal{I}} \lambda_i \nabla g_i(x) + \sum_{j\in\mathcal{J}} \mu_j \nabla h_j(x) = 0 \qquad \text{(KKT-G)}$$

$$\sum_{i\in\mathcal{I}} \lambda_i g_i(x) = 0 \qquad \text{(KKT-CS)}$$

- KKT Theorem: $T_X(x) = D_X(x) \wedge x$ local optimum $\Longrightarrow$ (KKT)
- CS = Complementary Slackness $\equiv \lambda_i g_i(x) = 0\ \forall\, i \in \mathcal{I}$ (**why?**)

Exercise: Prove KKT Theorem. Where does (CS) come from? Why $\mu \ngeq 0$?

- $x \in X \equiv g_i(x) \leq 0 \;\; i \in \mathcal{I} \;\;,\;\; h_j(x) = 0 \;\; j \in \mathcal{J}$ (doh!)
- Optimization $\equiv$ solving systems of nonlinear equations and inequalities



- $T_X(x) = D_X(x)$ crucial: counter-example

$$\min\{\, x_1 \;:\; x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,\;\; x_2 \leq 0 \,\}$$

The Karush-Kuhn-Tucker conditions

- Karush-Kuhn-Tucker conditions: $x \in X$, $\exists\, \lambda \in \mathbb{R}_+^{|\mathcal{I}|}$ and $\mu \in \mathbb{R}^{|\mathcal{J}|}$ s.t.

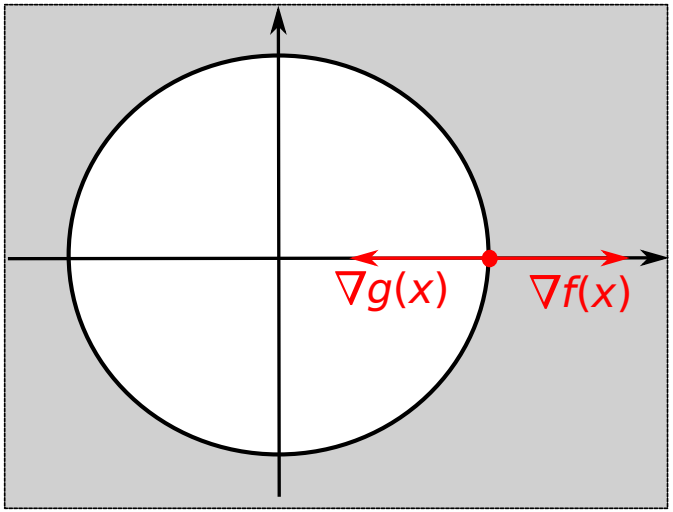
$$\nabla f(x) + \sum_{i\in\mathcal{I}} \lambda_i \nabla g_i(x) + \sum_{j\in\mathcal{J}} \mu_j \nabla h_j(x) = 0 \qquad \text{(KKT-G)}$$

$$\sum_{i\in\mathcal{I}} \lambda_i g_i(x) = 0 \qquad \text{(KKT-CS)}$$

- KKT Theorem: $T_X(x) = D_X(x) \wedge x$ local optimum $\Longrightarrow$ (KKT)
- CS = Complementary Slackness $\equiv \lambda_i g_i(x) = 0\ \forall\, i \in \mathcal{I}$ (**why?**)

Exercise: Prove KKT Theorem. Where does (CS) come from? Why $\mu \ngeq 0$?

- $x \in X \equiv g_i(x) \leq 0 \;\; i \in \mathcal{I} \;\;,\;\; h_j(x) = 0 \;\; j \in \mathcal{J}$ (doh!)
- Optimization $\equiv$ solving systems of nonlinear equations and inequalities



- $T_X(x) = D_X(x)$ crucial: counter-example
  $\min\{\, x_1 \;:\; x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,\;\; x_2 \leq 0 \,\}$
- Condition not necessary: counter-example
  $\min\{\, x_1 \;:\; x_1^2 + x_2^2 \geq 1 \,\} \;\;,\;\; x = [\,1\,,\,0\,]$
- First-order cannot tell maxima from minima

The Karush-Kuhn-Tucker conditions

- Karush-Kuhn-Tucker conditions: $x \in X$, $\exists\, \lambda \in \mathbb{R}_+^{|\mathcal{I}|}$ and $\mu \in \mathbb{R}^{|\mathcal{J}|}$ s.t.

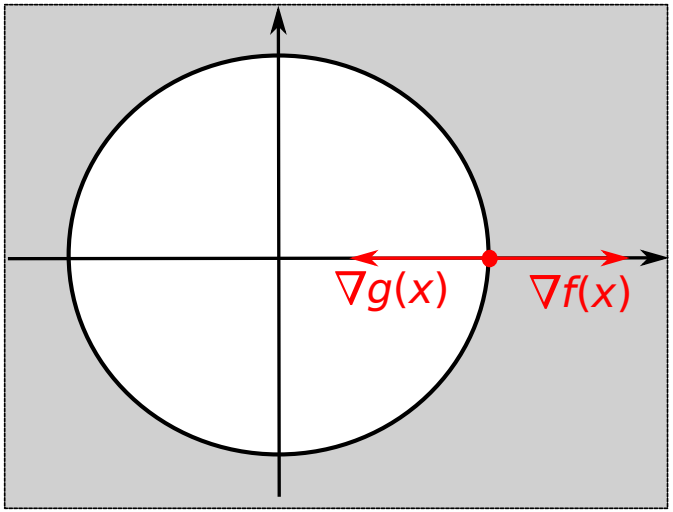
$$\nabla f(x) + \sum_{i\in\mathcal{I}} \lambda_i \nabla g_i(x) + \sum_{j\in\mathcal{J}} \mu_j \nabla h_j(x) = 0 \qquad \text{(KKT-G)}$$

$$\sum_{i\in\mathcal{I}} \lambda_i g_i(x) = 0 \qquad \text{(KKT-CS)}$$

- KKT Theorem: $T_X(x) = D_X(x) \wedge x$ local optimum $\Longrightarrow$ (KKT)
- CS = Complementary Slackness $\equiv \lambda_i g_i(x) = 0\ \forall\, i \in \mathcal{I}$ (**why?**)

Exercise: Prove KKT Theorem. Where does (CS) come from? Why $\mu \ngeq 0$?

- $x \in X \equiv g_i(x) \leq 0 \;\; i \in \mathcal{I} \;\;,\;\; h_j(x) = 0 \;\; j \in \mathcal{J}$ (doh!)
- Optimization $\equiv$ solving systems of nonlinear equations and inequalities



- $T_X(x) = D_X(x)$ crucial: counter-example
  $\min\{\, x_1 \;:\; x_1^2 + (x_2 - 1)^2 - 1 \leq 0 \;\;,\;\; x_2 \leq 0 \,\}$
- Condition not necessary: counter-example
  $\min\{\, x_1 \;:\; x_1^2 + x_2^2 \geq 1 \,\} \;\;,\;\; x = [\,1\,,\,0\,]$
- First-order cannot tell maxima from minima
  only safe case: no maxima

Karush-Kuhn-Tucker conditions and convexity

- (P) convex problem $\equiv$ both f and X convex (doh!) $\Longleftarrow$ $g_i(x)$ convex $\forall i \in \mathcal{I}$, $h_j(x)$ affine $\forall i \in \mathcal{I}$ (**why?**)

Exercise: Is $\Longrightarrow$ true?

- We now assume (P) convex:
  - x global optimum $\iff \langle \nabla f(x), d \rangle \geq 0 \; \forall \, d \in T_X(x)$
  - (KKT) $\Longrightarrow \langle \nabla f(x), d \rangle \geq 0 \; \forall \, d \in D_X(x)$
  - $D_X(x) \subseteq T_X(x) \Longrightarrow$ (KKT) $\Longrightarrow \langle \nabla f(x), d \rangle \geq 0 \; \forall \, d \in T_X(x)$
- Under convexity, (KKT) $\Longrightarrow$ x global optimum
- But remember necessary only under constraint qualification (**why?**)

Exercise: Where exactly convexity has been used? What happens without?

Exercise: Compute $\min_y \{ \| y - x \|_2^2 \; : \; ay = b \}$

Exercise: Compute $\min_{x,y} \{ \| y - x \|_2^2 \; : \; ax = b_1 \;\; , \;\; ay = b_2 \}$

Exercise: Compute $\min_y \{ \| y - x \|_2^2 \; : \; a_i \leq y_i \leq b_i \;\; i = 1, 2 \}$

Outline

Critical cone

- (P) not convex $\equiv$ (KKT) not sufficient $\Longrightarrow$ have to use second-order
- Assume (x, λ, μ) satisfies (KKT): critical cone

$$C(x, \lambda, \mu) = \left\{ d \in \mathbb{R}^n : \begin{array}{ll} \langle \nabla g_i(x), d \rangle = 0 & i \in \mathcal{A}(x) \text{ s.t. } \lambda_i^* > 0 \\ \langle \nabla g_i(x), d \rangle \leq 0 & i \in \mathcal{A}(x) \text{ s.t. } \lambda_i^* = 0 \\ \langle \nabla h_j(x), d \rangle = 0 & i \in \mathcal{J} \end{array} \right\}$$

- Lagrangian function: $L(x, \lambda, \mu) = f(x) + \sum_{i \in \mathcal{I}} \lambda_i g_i(x) + \sum_{j \in \mathcal{I}} \mu_j h_j(x)$
- (x, λ, μ) satisfies (KKT) $\wedge$ x satisfies (LinI): x local optimum $\Longrightarrow$

  $d^T \nabla^2_{xx} L(x, \lambda, \mu) d \geq 0 \qquad \forall\, d \in C(x, \lambda, \mu)$

  "the Hessian of the Lagrangian function is $\succeq 0$ on the critical cone"
- (x, λ, μ) satisfies (KKT) $\wedge$ $\nabla^2_{xx} L(x, \lambda, \mu) \succ 0$ on $C(x, \lambda, \mu)$

  $\Longrightarrow$ x local optimum (sufficient)
- Conditions for unconstrained optimization a special case (**check**)

Exercise: Geometrically find local and global optima of $\min\{ -x_1^2 - x_2^2 : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 \}$, then verify your findings with (KKT) and second-order conditions

Outline

Lagrangian relaxation

- Lagrangian function interesting object: objective and constraints together
- (KKT-G) $\equiv$ stationary point of $L(\cdot)$ (**check**). Should we minimize $L(\cdot)$? It depends on the order in which you minimize variables on
- Lagrangian relaxation: $(R_{\lambda,\mu})$ $\psi(\lambda, \mu) = \min\{ L(x, \lambda, \mu) : x \in \mathbb{R}^n \}$
- $\forall$ fixed $\lambda \geq 0, \mu$, $\psi(\lambda, \mu) \leq v(P)$: $\bar{x} \in X \Longrightarrow g(\bar{x}) \leq 0$, $h(\bar{x}) = 0 \Longrightarrow$
  $\psi(\lambda, \mu) = \min_x L(x, \lambda, \mu) \leq L(\bar{x}, \lambda, \mu) \leq f(\bar{x})$ (**check**)
- "Relaxation $\equiv$ have more solutions": if you minimize, the value is less
- ψ is nice-ish: i) ψ concave (**why?**), ii) $\psi(\lambda, \mu) = -\infty$ happens iii) ψ not differentiable even if f, g_i, h_j are (**why?**)
- i) $\Longrightarrow \psi \approx$ easy to maximize $\longrightarrow$ Lagrangian dual of (P):
  (D) $\max\{ \psi(\lambda, \mu) : \lambda \in \mathbb{R}^{|\mathcal{I}|}_+ , \ \mu \in \mathbb{R}^{|\mathcal{J}|} \}$
- "Weak duality": $v(D) \leq v(P)$ (**why?**)
- A lower bound on $v(P)$ solving a convex program even if (P) not convex

Lagrangian relaxation

- Lagrangian function interesting object: objective and constraints together
- (KKT-G) $\equiv$ stationary point of $L(\cdot)$ (**check**). Should we minimize $L(\cdot)$? It depends on the order in which you minimize variables on
- Lagrangian relaxation: $(R_{\lambda,\mu})$ $\psi(\lambda, \mu) = \min\{ L(x, \lambda, \mu) : x \in \mathbb{R}^n \}$
- $\forall$ fixed $\lambda \geq 0, \mu$, $\psi(\lambda, \mu) \leq v(P)$: $\bar{x} \in X \Longrightarrow g(\bar{x}) \leq 0$, $h(\bar{x}) = 0 \Longrightarrow$
  $\psi(\lambda, \mu) = \min_x L(x, \lambda, \mu) \leq L(\bar{x}, \lambda, \mu) \leq f(\bar{x})$ (**check**)
- "Relaxation $\equiv$ have more solutions": if you minimize, the value is less
- ψ is nice-ish: i) ψ concave (**why?**), ii) $\psi(\lambda, \mu) = -\infty$ happens iii) ψ not differentiable even if f, g_i, h_j are (**why?**)
- i) $\Longrightarrow \psi \approx$ easy to maximize $\longrightarrow$ Lagrangian dual of (P):
  (D) $\max\{ \psi(\lambda, \mu) : \lambda \in \mathbb{R}^{|\mathcal{I}|}_+ , \ \mu \in \mathbb{R}^{|\mathcal{J}|} \}$
- "Weak duality": $v(D) \leq v(P)$ (**why?**)
- A lower bound on $v(P)$ solving a convex program even if (P) not convex
- But $\psi(\cdot) = \min_x L(\cdot)$ has to be solved to global optimality

The Lagrangian dual

- (D) is not unconstrained, but constraints very easy: $\lambda \geq 0$
- Extends to the case where X has other constraints $x \in X'$ (e.g., $x \in \mathbb{Z}^n$) provided you can compute $\psi(\cdot) \equiv$ solve $(R_{\lambda,\mu})$ to global optimality
- How good is the bound $v(D)$? When is $v(D) = v(P)$ ("strong duality")?
- Not always: $\min\{ -x^2 : 0 \leq x \leq 1 \}$, $L(x, \lambda) = -x^2 + \lambda_1(x-1) - \lambda_2 x$

  $\psi(\lambda) = \min_{x \in \mathbb{R}} L(x, \lambda) = -\infty \quad \forall \lambda \in \mathbb{R}^2$ **(why?)**

  $\Longrightarrow v(D) = -\infty < v(P) = -1$

  Note: $x^* = 1$, $\lambda_1^* = 2$, $\lambda_2^* = 0 \Longrightarrow -2x^* + \lambda_1^* - \lambda_2^* = 0 \equiv$ KKT
- Counter-example is nonconvex, convexity (and regularity) does help here:

  (P) convex, x^* optimal solution, $T_X(x^*) = D_X(x^*) \Longrightarrow \exists (\lambda^*, \mu^*)$ KKT multipliers $\Longrightarrow (\lambda^*, \mu^*)$ optimal solution to (D) $\wedge\, v(D) = v(P)$

  Proof: $\nabla_x L(x^*, \lambda^*, \mu^*) = 0 \wedge L(\cdot)$ convex in $x \Longrightarrow v(D) \geq \psi(\lambda^*, \mu^*)$

  $= \min_x L(x, \lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*) = f(x^*) = v(P) \geq v(D)$
- In counter-example, x^* maximum of $L(\cdot, \lambda^*, \mu^*)$
- Strong duality can hold also for nonconvex problems

Outline

Simpler Lagrangian duals: Linear Programs

- Lagrangian dual powerful but cumbersome: max min
- Simplifies to "just max" in many special cases
- Linear Program: (P) $\quad \min\{\, cx \,:\, Ax \geq b \,\}$
- Lagrangian function: $L(\,x\,,\,\lambda\,) = cx + \lambda(\,b - Ax\,) = \lambda b + (c - \lambda A)x$

$$\Longrightarrow \qquad \psi(\,\lambda\,) = \min_{x \in \mathbb{R}^n} L(\,x\,,\,\lambda\,) = \begin{cases} -\infty & \text{if } c - \lambda A \neq 0 \\ \lambda b & \text{if } c - \lambda A = 0 \end{cases}$$

- $(D) \quad \max\{\,\psi(\,\lambda\,) \,:\, \lambda \geq 0\,\} \equiv \max\{\,\lambda b \,:\, \lambda A = c\,,\ \lambda \geq 0\,\}$
  another linear program ($\neq$, but with the same data)
- Trick: $(R_{\lambda,\mu})$ so simple it can be solved by closed formulæ
- Strong duality $\equiv v(P) = v(D)$ (almost) always holds

Simpler Lagrangian duals: Linear Programs

- Lagrangian dual powerful but cumbersome: max min
- Simplifies to "just max" in many special cases
- Linear Program: (P) $\quad \min\{\, cx \,:\, Ax \geq b \,\}$
- Lagrangian function: $L(\,x\,,\,\lambda\,) = cx + \lambda(\,b - Ax\,) = \lambda b + (c - \lambda A)x$

$$\Longrightarrow \qquad \psi(\,\lambda\,) = \min_{x \in \mathbb{R}^n} L(\,x\,,\,\lambda\,) = \begin{cases} -\infty & \text{if } c - \lambda A \neq 0 \\ \lambda b & \text{if } c - \lambda A = 0 \end{cases}$$

- $(D) \quad \max\{\,\psi(\,\lambda\,) \,:\, \lambda \geq 0\,\} \equiv \max\{\,\lambda b \,:\, \lambda A = c\,,\ \lambda \geq 0\,\}$
  another linear program ($\neq$, but with the same data)
- Trick: $(R_{\lambda,\mu})$ so simple it can be solved by closed formulæ
- Strong duality $\equiv v(P) = v(D)$ (almost) always holds

Exercise: prove last statement. Why "almost"? Can $v(P) > v(D)$ happen?

Exercise: what is the dual of (D)?

Exercise: what if (P) $\quad \min\{\, cx \,:\, Ax \geq b\,,\, Ex = d\,\}$? (D)?

Exercise: what if (P) $\quad \min\{\, c'x' + c''x'' \,:\, A'x' + A''x'' \geq b\,,\, x' \geq 0\,\}$? (D)?

Simpler Lagrangian duals: Quadratic Programs

- Simple case: (P) $\min \left\{ \frac{1}{2} \| x \|_2^2 \, : \, Ax = b \right\}$ (linear least-norm solution)
- $L(x , \mu) = \frac{1}{2} \| x \|_2^2 + \mu (Ax - b)$, $\nabla_x L = x + \mu A = 0 \iff x = -\mu A$
- $\psi(\mu) = \min_{x \in \mathbb{R}^n} L(x , \mu) = L(-\mu A , \mu) = -\frac{1}{2} \mu^T (AA^T) \mu - \mu b$
  $\Longrightarrow$ (D) $\max \{ -\frac{1}{2} \mu^T (AA^T) \mu - \mu b \, : \, \mu \in \mathbb{R}^m \}$ (an unconstrained QP)

Simpler Lagrangian duals: Quadratic Programs

- Simple case: (P) $\min \left\{ \frac{1}{2} \| x \|_2^2 \, : \, Ax = b \right\}$ (linear least-norm solution)
- $L(x , \mu) = \frac{1}{2} \| x \|_2^2 + \mu (Ax - b)$, $\nabla_x L = x + \mu A = 0 \iff x = -\mu A$
- $\psi(\mu) = \min_{x \in \mathbb{R}^n} L(x , \mu) = L(-\mu A , \mu) = -\frac{1}{2} \mu^T (AA^T) \mu - \mu b$
  $\Longrightarrow$ (D) $\max \{ -\frac{1}{2} \mu^T (AA^T) \mu - \mu b \, : \, \mu \in \mathbb{R}^m \}$ (an unconstrained QP)
- Strictly convex QP: (P) $\min \left\{ \frac{1}{2} x^T Q x + q x \, : \, Ax \geq b \right\}$, $Q \succ 0 \Longrightarrow$
  (D) $\max \left\{ \lambda b - \frac{1}{2} v^T Q^{-1} v \, : \, \lambda A - v = q \, , \, \lambda \geq 0 \right\}$
  strong duality $\equiv v(P) = v(D)$ (almost) always holds (**why?**)

Simpler Lagrangian duals: Quadratic Programs

- Simple case: (P) $\min\left\{ \frac{1}{2}\| x \|_2^2 \,:\, Ax = b \right\}$ (linear least-norm solution)
- $L(x,\mu) = \frac{1}{2}\| x \|_2^2 + \mu(Ax - b)$, $\nabla_x L = x + \mu A = 0 \iff x = -\mu A$
- $\psi(\mu) = \min_{x\in\mathbb{R}^n} L(x,\mu) = L(-\mu A,\mu) = -\frac{1}{2}\mu^T(AA^T)\mu - \mu b$
  $\Longrightarrow$ (D) $\max\{ -\frac{1}{2}\mu^T(AA^T)\mu - \mu b \,:\, \mu \in \mathbb{R}^m \}$ (an unconstrained QP)
- Strictly convex QP: (P) $\min\left\{ \frac{1}{2}x^T Q x + qx \,:\, Ax \geq b \right\}$, $Q \succ 0 \Longrightarrow$
  (D) $\max\left\{ \lambda b - \frac{1}{2}v^T Q^{-1} v \,:\, \lambda A - v = q\,,\ \lambda \geq 0 \right\}$
  strong duality $\equiv v(P) = v(D)$ (almost) always holds (**why?**)

Exercise: prove last (P) $\rightsquigarrow$ (D). What would change if (P) had $Ax = b$?

Exercise: compute (D) when "some variables are not quadratic":
$x = [x'\,,\, x'']$ and objective is $\frac{1}{2}(x'')^T(Q'')x'' + qx' + q''x''$ with $Q'' \succ 0$

Exercise: compute (D) for the Support Vector Machine:
$\min_{w,\xi}\left\{ \| w_+ \|_2^2 + C\sum_{i=1}^m \xi_i \,:\, y^i(w_+ x^i + w_0) \geq 1 - \xi_i \ ,\ \xi_i \geq 0 \quad i = 1,\ldots,m \right\}$

Exercise: compute (D) for the Proximal Bundle Method:

$$\min_{v,d}\left\{ v + \frac{\mu}{2}\| d \|_2^2 \,:\, v \geq g^i d - \alpha^i \quad i = 1,\ldots,m \right\}$$

then develop a closed formula to solve (P) when $m = 2$

Conic Programs, Conic Duals

- Conic program: (P) $\min\{\, cx \,:\, Ax \geq_K b \,\}$, where
  $x \geq_K y \equiv x - y \in K$, K pointed convex cone
  - $K = \mathbb{R}^n_+ \equiv$ sign constraints $\equiv$ Linear Program
  - $K = \mathbb{L} = \left\{ x \in \mathbb{R}^n \,:\, x_n \geq \sqrt{\sum_{i=1}^{n-1} x_i^2} \right\} \equiv$ Second-Order Cone Program
  - $K = \mathbb{S}_+ = \{ A \succeq 0 \} \equiv$ "$\succeq$" constraints $\equiv$ SemiDefinite Program
- Exceedingly smart idea: everything is linear, but the cone is not
  $\equiv$ a nonlinear program disguised as a linear one
- Conic dual: (D) $\max\{\, yb \,:\, yA = c \,,\, y \geq_{K^D} 0 \,\}$
  $K^D = \{\, z \,:\, \langle z \,,\, x \rangle \geq 0 \;\; \forall\, x \in K \,\}$ dual cone

Exercise: prove (P) $\rightsquigarrow$ (D) for conic programs

- $\neq$ definition from before, actually $K^D = -K^\circ$
- All three cones are self-dual: $K^D = K$
  "the angle at the vertex of the cone is 90 degrees"
- Strong duality not always holds, constraint qualification needed
  one of the constraints is nonlinear, even if it does not look so

Conic Duals in practice

- SOCP: (P) $\quad \min \left\{ cx \,:\, \| D_i x - d_i \|_2 \le p_i x - q_i \quad i = 1, \ldots, m \right\}$

  (D) $\quad \max \sum_{i=1}^m \lambda_i d_i + \nu_i q_i$

  $\sum_{i=1}^m \lambda_i D_i + \nu_i p_i = c$

  $\| \lambda_i \|_2 \le \nu_i \qquad\qquad i = 1, \ldots, m$

- Any LP is a SOCP ($D_i = 0$, $d_i = 0$)
- Any program with convex quadratic constraints is a SOCP:

  $x^T x \le t \iff \| [x, (t-1)/2] \|_2 \le (t+1)/2$ (**check**)

  vice-versa is not true

Exercise: prove $x^T x/s$ for $s > 0$ is a SOCP (hint: $\|x\|_2^2 = x^T x/1$)

- (SDP): (P) $\quad \min \left\{ cx \,:\, \sum_{i=1}^n x_i A^i \succeq B \right\}$

  (D) $\quad \max \left\{ \langle B, \Lambda \rangle \,:\, \langle A^i, \Lambda \rangle = c_i \;\; i = 1, \ldots, n \;\;,\;\; \Lambda \succeq 0 \right\}$

  A^i, $B \in \mathbb{R}^{n \times n}$ symmetric but not necessarily SDP

  $\langle A, B \rangle = \sum_i \sum_j A_{ij} B_{ij}$ (Frobenius inner product)

Exercise: prove: any SOCP is a SDP (hint: consider $\succeq 0$ for 2×2 matrices), vice-versa is not true (easy to see, nontrivial to prove)

Fenchel's Dual

- Fenchel's conjugate of f: $f^*(z) = sup_x\{\, zx - f(x)\,\}$
- f^* always convex even if f is not (**why?**), closed if f is
- f^{**} convex envelope of f = largest convex g s.t. $g(x) \le f(x)\ \forall x \in \mathbb{R}^n$, $f^{**} = f$ if f closed convex
- Many useful properties:
  1. $f^*(0) = -\inf_x\{\, f(x)\,\}$ (characterizes x^* and f_*, minus the sign)
  2. $f \le g \Longrightarrow f^* \ge g^*$
  3. $zx \le f(\,x\,) + f^*(\,z\,) \quad \forall z, x$
  4. $z \in \partial_\varepsilon f(\,x\,) \iff x \in \partial_\varepsilon f^*(\,z\,) \iff f(\,x\,) + f^*(\,z\,) \le zx + \varepsilon$ (characterizes $\partial_\varepsilon f$)

Exercise: prove $z \in \partial f(\,x\,) \iff x \in \partial f^*(\,z\,) \iff f(\,x\,) + f^*(\,z\,) = zx$

- Fenchel's dual: (P) $\min\{\, f(x) + g(x)\,\}$, (D) $-\min\{\, f^*(z) + g^*(-z)\,\}$

Exercise: prove $z(P) = (D)$ if f and g convex and (P) has optimum

Exercise: why "(P) has optimum" is needed? What can go wrong if not?

Exercise: "prove" Fenchel's duality out of Lagrangian duality, starting backward from (D) $-\min\{\, f^*(z') + g^*(z'') \,:\, z' + z'' = 0\,\}$

Fenchel's calculus

- f^* can be computed for "*easy*" f (**exercise: check**)
  1. $f(x) = \frac{1}{2}\|x\|_2^2 \Longrightarrow f^*(z) = \frac{1}{2}\|z\|_2^2$ (only function s.t. $f^* = f$)
  2. $(\|\cdot\|_1)^*(z) = \imath_{\mathcal{B}_\infty(0,1)}(z)$, $(\|\cdot\|_\infty)^*(z) = \imath_{\mathcal{B}_1(0,1)}(z)$
  3. $f(x) = \max\{ g_i x - \alpha_i \;\; i \in I \} \Longrightarrow$
     $f^*(z) = \min\left\{ \sum_{i\in I} \alpha_i\theta_i \;:\; \sum_{i\in I} g_i\theta_i = z \;,\; \sum_{i\in I}\theta_i = 1 \;,\; \theta_i \geq 0 \quad i \in I \right\}$
- Then, more f^* can be derived with appropriate calculus rules:
  1. $(f(\cdot) + \alpha)^*(z) = f^*(z) - \alpha$
  2. $f(\cdot - \bar{x})^*(z) = f^*(z) + \langle z\,,\,\bar{x}\rangle$
  3. $(\alpha f)^*(z) = \alpha f^*(z/\alpha)$
  4. $(f+g)^*(z) = \inf\{ f^*(z') + g^*(z'') \;:\; z = z' + z'' \}$
  5. ... (rather complicaetd)
- Possibly the most powerful framework for deriving duals,
  but not for the faint of heart

Exercise: compute (D) for the Proximal Bundle Method

(P) $\min_x\{ f_{\mathscr{B}}(x) + \mu\|x - \bar{x}\|_2^2/2 \}$, compare with QP dual 4 slides ago

Exercise: compute in two $\neq$ ways (D) for the Trust Region Bundle Method

(P) $\min_x\{ f_{\mathscr{B}}(x) \;:\; \|x - \bar{x}\|_\infty \leq \delta \}$

Outline

Wrap up

- Constraints $\rightsquigarrow$ Lagrangian multipliers $\equiv$ "more variables"
  $\rightsquigarrow$ Lagrangian duality: powerful, but max / min
- Convex $\rightsquigarrow$ strong duality, nonconvex $\rightsquigarrow$ relaxation (and ψ "difficult")
- Sometimes "ψ very easy", can do away with $x \Longrightarrow$ problem only in λ, μ
- Sometimes (D) easier than (P) (e.g., $m \ll n$)
- LP/QP/Conic duality important special cases, easy to use
- Fenchel's most powerful form giving "closed" duals, nontrivial to use
- Convex $\rightsquigarrow$ algorithms can work in primal space, dual space or both
- Have you said "algorithms"? Yup, let's move on!