Convex Programs

COMPSCI 371D — Machine Learning

Support Vector Machines (SVMs) and Convex Programs

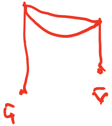
- SVMs are linear predictors
- Defined for both regression and classification
- Multi-class versions exist
- We will cover only *binary* SVM *classification*
- We'll need some new math: *Convex Programs*
- Optimization of convex functions with affine constraints (equalities and inequalities)
- Lagrange multipliers, but for inequalities

Outline

Logistic Regression → SVMs

- A logistic-regression classifier places the decision boundary *somewhere* (and approximately) between the two classes
- Loss is never zero → Exact location of the boundary can be determined by samples that are very distant from the boundary (even on the correct side of it)
- SVMs place the boundary “exactly half-way” between the two classes (with exceptions to allow for non linearly-separable classes)
- Only samples close to the boundary matter:
  These are the *support vectors*
- *SVMs are effectively immune from the curse of dimensionality*
- *A “kernel trick” allows going beyond linear classifiers*

Local Convex Minimization → Convex Programs



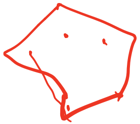
- Convex function $f(\mathbf{u}) \; : \; \mathbb{R}^m \to \mathbb{R}$
- f differentiable, with continuous first derivatives
- Unconstrained minimization: $\mathbf{u}^* \;=\; \arg\min_{\mathbf{u}\in\mathbb{R}^m} f(\mathbf{u})$
- Constrained minimization: $\mathbf{u}^* \;=\; \arg\min_{\mathbf{u}\in C} f(\mathbf{u})$
- $C = \{\mathbf{u} \in \mathbb{R}^m \;:\; A\mathbf{u} + \mathbf{b} \geq \mathbf{0}\}$
- f is a convex *function*
- C is a convex *set*: If $\mathbf{u}, \mathbf{v} \in C$, then for $t \in [0, 1]$ $t\mathbf{u} + (1 - t)\mathbf{v} \in C$
- The specific C is bounded by hyperplanes
- This is a *convex program*

k constraints

$A \quad k \times m$

$\bar{b} \quad k \times 1$

$$\begin{cases} \bar{a}_1^T \bar{u} + b_1 \geq 0 \\ \vdots \\ \bar{a}_k^T \bar{u} + b_k \geq 0 \end{cases}$$



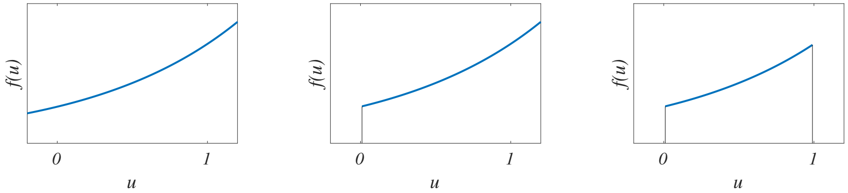
Shape of the Solution Set

- Just as for the unconstrained problem:
  - There is one f^* but there can be multiple $\mathbf{u}^*$ (a flat valley)
  - The set of solution points $\mathbf{u}^*$ is convex
  - if f is strictly convex at $\mathbf{u}^*$, then $\mathbf{u}^*$ is the unique solution point

Zero Gradient $\rightarrow$ KKT Conditions

- For the unconstrained problem, the solution is characterized by $\nabla f(\mathbf{u}) = \mathbf{0}$
- Constraints can generate new minima and maxima
- Example: $f(u) = e^u$

$u \geq 0$ $\qquad$ $u \leq 1$ $\quad \begin{cases} u \geq 0 \\ -u+1 \geq 0 \end{cases}$

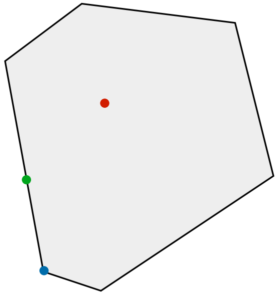


- What is the new characterization?
- *Karush-Kuhn-Tucker conditions*, necessary and sufficient

Geometry of C

- The neighborhoods of points of $\mathbb{R}^m$ "look the same"
- Not so for C: Different points "look different"
- Example $m = 2$: C is a convex polygon

Prove convexity!

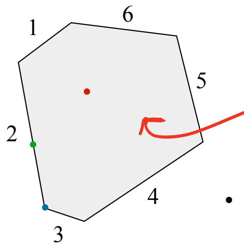


- View from the interior: Same as $\mathbb{R}^2$
- View from a side: C is a half-plane
- View from a corner: C is an angle

Active Constraints

WEAK → C closed

- A constraint $c_i(\mathbf{u}) \geq 0$ is *active* at $\mathbf{u}$ if $c_i(\mathbf{u}) = 0$
- $\mathbf{u}$ "touches" that constraint
- The *active set*: $\mathcal{A}(\mathbf{u}) = \{i \; : \; c_i(\mathbf{u}) = 0\}$ $\subseteq \{1, \ldots, k\}$

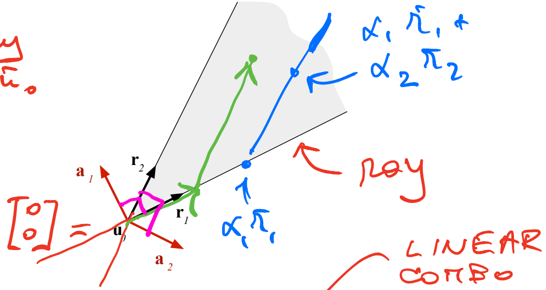


$\bar{u}$ FEASIBLE

- $\mathcal{A}(\mathbf{u}) = \{\}$
- $\mathcal{A}(\mathbf{u}) = \{2\}$
- $\mathcal{A}(\mathbf{u}) = \{2, 3\}$
- Black $\mathbf{u}$ is *infeasible*, the others are *feasible*

Closed, Convex Polyhedral (CCP) Cones

b_i goes away by moving $\bar{u}_0$ to $\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

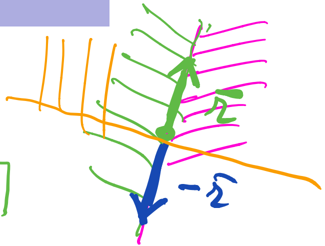


- View from $\mathbf{u}_0$: move the origin to $\mathbf{u}_0$
- $\{\mathbf{u} \in \mathbb{R}^2 \ : \ \mathbf{a}_1^T\mathbf{u} \geq 0 \, , \ \mathbf{a}_2^T\mathbf{u} \geq 0\}$ (implicit)
- $\{\mathbf{u} \in \mathbb{R}^2 \ : \ \mathbf{u} = \alpha_1\mathbf{r}_1 + \alpha_2\mathbf{r}_2$ with $\alpha_1, \alpha_2 \geq 0\}$ (parametric)
- Both representations always exist (Farkas-Minkowski-Weyl)
- Number of hyperplane normals $\mathbf{a}_i$ and *generators* $\mathbf{r}_j$ is not always the same. Conversion is typically complex. We won't need it.

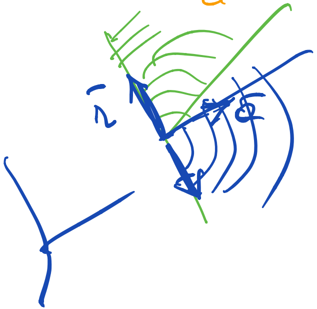
Lines, Half Lines, Half-Planes, Planes

α_1 α_2 α_3

- Line in $\mathbb{R}^2$: $\mathbf{a}^T\mathbf{u} \geq 0\,,\ -\mathbf{a}^T\mathbf{u} \geq 0$ or $\boxed{\mathbf{u} = \alpha_1\mathbf{r} + \alpha_2(-\mathbf{r})}$
- Half line: $\mathbf{a}^T\mathbf{u} \geq 0\,,\ -\mathbf{a}^T\mathbf{u} \geq 0\,,\ \mathbf{r}^T\mathbf{u} \geq 0$ or $\boxed{\mathbf{u} = \alpha\mathbf{r}}$
- Half plane: $\mathbf{a}^T\mathbf{u} \geq 0$ or $\mathbf{u} = \alpha_1\mathbf{r} + \alpha_2(-\mathbf{r}) + \alpha_3\mathbf{a}$
  (glue two angles together)
- Plane: $\{\}$ or $\mathbf{u} = \alpha_1\mathbf{r}_1 + \alpha_2\mathbf{r}_2 + \alpha_3\mathbf{r}_3$
  with $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$, say, 120 degrees apart
  (glue three angles together)
- Parametric representation is not unique
- More variety in $\mathbb{R}^3$ much more in $\mathbb{R}^d$



GENERATOR

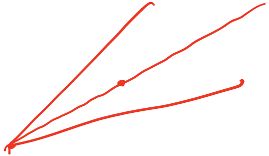


General CCP Cones

THE LOCAL VIEW of C

CLOSED : "$\geq$"
CONVEX : prove it!
POLYHEDRAL: bounded by planes

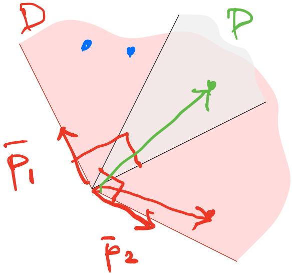
- $P = \{\mathbf{u} \in \mathbb{R}^m \ : \ \mathbf{p}_i^T \mathbf{u} \geq 0 \ \text{ for } \ i = 1, \ldots, k\}$
- $P = \{\mathbf{u} \in \mathbb{R}^m \ : \ \mathbf{u} = \sum_{j=1}^{\ell} \alpha_j \mathbf{r}_j \ \text{ with } \ \alpha_1, \ldots, \alpha_\ell \geq 0\}$
- Both representations always exist (Farkas-Minkowski-Weyl)
- The conversion is algorithmically complex ("representation conversion problem")

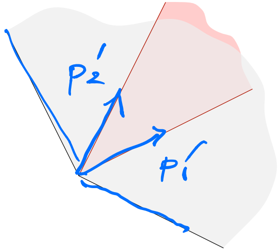


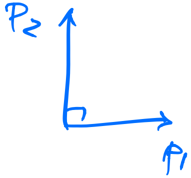
CONE : $\bar{u} \in P \Rightarrow \alpha \bar{u} \in P$ for any $\alpha \geq 0$

Cone Duality

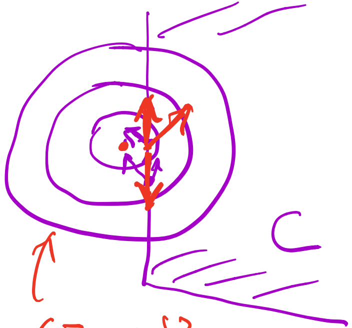
- $P = \{\mathbf{u} \in \mathbb{R}^m \ : \ \mathbf{p}_i{}^T\mathbf{u} \geq 0 \ \text{ for } \ i = 1, \ldots, k\}$
- $D = \{\mathbf{u} \in \mathbb{R}^m \ : \ \mathbf{u} = \sum_{i=1}^{k} \alpha_i \mathbf{p}_i \ \text{ with } \ \alpha_1, \ldots, \alpha_k \geq 0\}$
- D is the *dual* of P
- Different cones, same vectors, different representations!
- ***All and only the points in P have a nonnegative inner product with all and only the points in D*** $\leq 90°$
- D dual of $P \Leftrightarrow P$ dual of D







Necessary and Sufficient Condition for a Minimum



$(\bar{u} - \bar{x})^2$

- $\mathbf{u}^*$ is a minimum iff $f(\mathbf{u})$ does not decrease when moving away from $\mathbf{u}^*$ *while staying in C*
- f is differentiable, so we can look at $\nabla f(\mathbf{u}^*)$
- No matter how we choose a direction $\mathbf{s} \in \mathbb{R}^m$, if a tiny step away from $\mathbf{u}^*$ and along $\mathbf{s}$ keeps us in C, then $\mathbf{s}^T \nabla f(\mathbf{u}^*)$ must be ≥ 0

$\bar{g}$:

- If $\mathbf{s} \in P = \{\mathbf{s} \ : \ \mathbf{s}^T \nabla c_i(\mathbf{u}^*) \geq 0 \ \text{ for } \ i \in \mathcal{A}(\mathbf{u}^*)\}$ then $\mathbf{s}^T \nabla f(\mathbf{u}^*)$ must be ≥ 0

$c_i(\bar{u}^*) = 0$!

$c_i = u_2 - 1 \geq 0$

- $\nabla f(\mathbf{u}^*)$ is any vector with a nonnegative inner product with all vectors in P
- $\nabla f(\mathbf{u}^*)$ is in the dual cone of P, $D = \{\mathbf{g} \ : \ \mathbf{g} = \sum_{i \in \mathcal{A}(\mathbf{u}^*)} \alpha_i \nabla c_i(\mathbf{u}^*) \ \text{ with } \ \alpha_i \geq 0\}$

The Karush-Kuhn-Tucker Conditions

- $\nabla f(\mathbf{u}^*)$ is in $D = \{\mathbf{g} \ : \ \mathbf{g} = \sum_{i \in \mathcal{A}(\mathbf{u}^*)} \alpha_i \nabla c_i(\mathbf{u}^*) \ \text{ with } \ \alpha_i \geq 0\}$
- $\nabla f(\mathbf{u}^*) = \sum_{i \in \mathcal{A}(\mathbf{u}^*)} \alpha_i^* \nabla c_i(\mathbf{u}^*)$ for some $\alpha_i^* \geq 0$
- $\frac{\partial}{\partial \mathbf{u}} \left[f(\mathbf{u}^*) - \sum_{i \in \mathcal{A}(\mathbf{u}^*)} \alpha_i^* c_i(\mathbf{u}^*) \right] = 0$ for some $\alpha_i^* \geq 0$
- There exist $\boldsymbol{\alpha}^*$ s.t. $\frac{\partial}{\partial \mathbf{u}} \mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*) = 0$ where
  $\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \stackrel{\text{def}}{=} f(\mathbf{u}) - \sum_{i \in \mathcal{A}(\mathbf{u})} \alpha_i c_i(\mathbf{u})$
  with $\alpha_i \geq 0$ for $i \in \mathcal{A}(\mathbf{u})$
- $\mathcal{L}$ is the *Lagrangian function* of this convex program
- The values in $\boldsymbol{\alpha}^*$ are called the *Lagrange multipliers*
- KKT conditions also hold in the interior of C!

$c(u) = 0$

$\begin{cases} c(u) \geq 0 \\ c(u) \leq 0 \end{cases}$

Technical Cleanup

- Simplifying the sum $\sum_{i \in \mathcal{A}(\mathbf{u})} \alpha_i c_i(\mathbf{u})$
- Anywhere in C, we have $\alpha_i \geq 0$ (cone) and $c_i(\mathbf{u}) \geq 0$ (feasible $\mathbf{u}$), so $\boxed{\sum_{i=1}^{k} \alpha_i c_i(\mathbf{u})} = \boxed{\boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u})} \geq 0$
- At $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$, the multiplier α_i^* is in the sum only if $c_i(\mathbf{u}^*) = 0$, so we can replace the sum with

$$\boxed{\boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u})}$$

if we add the *complementarity constraint* $\boxed{(\boldsymbol{\alpha}^*)^T \mathbf{c}(\mathbf{u}^*) = 0}$ which requires $\alpha_i^* = 0$ for $i \notin \mathcal{A}(\mathbf{u}^*)$

- At $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$, the sum $\sum_{i \in \mathcal{A}(\mathbf{u})} \alpha_i c_i(\mathbf{u})$ is equivalent to $\boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u})$ together with the constraint $(\boldsymbol{\alpha}^*)^T \mathbf{c}(\mathbf{u}^*) = 0$

The KKT Conditions

- It is necessary and sufficient for $\mathbf{u}^*$ to be a solution to a convex program that there exists $\boldsymbol{\alpha}^*$ such that

$$\begin{array}{ll} \frac{\partial}{\partial \mathbf{u}}\mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*) = 0 & \text{(no descent direction)} \\ \mathbf{c}(\mathbf{u}^*) \geq 0 & \text{(feasibility)} \\ \boldsymbol{\alpha}^* \geq 0 & \text{(cone)} \\ (\boldsymbol{\alpha}^*)^T\mathbf{c}(\mathbf{u}^*) = 0 & \text{(complementarity)} \end{array}$$

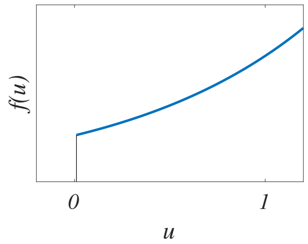
where $\boxed{\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \stackrel{\text{def}}{=} f(\mathbf{u}) - \boldsymbol{\alpha}^T\mathbf{c}(\mathbf{u})}$

KKT

- We can now recognize a minimum if we see one
- Subsumes to $\nabla f(\mathbf{u}^*) = 0$ for the unconstrained case
- Since both $f(\mathbf{u})$ and $\boldsymbol{\alpha}^T\mathbf{c}(\mathbf{u})$ are convex, so is $\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha})$
- Therefore, $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ *is a global minimum of $\mathcal{L}$ w.r.t.* $\mathbf{u}$ (because of the first KKT condition)

A Tiny Example

- In simple examples, we can *solve* the KKT conditions and find the minimum (as opposed to just checking whether a given $\mathbf{u}^*$ is a minimum)
- This is analogous to solving the *normal equations* $\nabla f(\mathbf{u}^*) = 0$ in the unconstrained case
- Example: $\min_u f(u) = e^u$ subject to $c(u) = u \geq 0$
- Lagrangian: $\mathcal{L}(u, \alpha) = f(u) - \alpha c(u) = e^u - \alpha u$
- We obviously know the solution, $u^* = 0$



A Tiny Example: KKT Conditions

- Lagrangian: $\mathcal{L}(u, \alpha) = e^u - \alpha u$

  $\frac{\partial}{\partial u}\mathcal{L}(u^*, \alpha^*) = e^{u^*} - \alpha^* = 0$

  $c(u^*) = u^* \geq 0$

  $\alpha^* \geq 0$

  $\alpha^* c(u^*) = \alpha^* u^* = 0$
- Solving first yields $\boxed{\alpha^* = e^{u^*}}$ which makes $\alpha^* \geq 0$ moot
- Complementarity yields $\boxed{u^* = 0}$, which is feasible
- [Yay!]
- We have $\alpha^* = e^{u^*} = e^0 = 1$
- Since $\alpha^* > 0$, this constraint is *active*

Lagrangian Duality

PRIMAL

- We will find a transformation of a convex program

$$f^* \stackrel{\text{def}}{=} \min_{\mathbf{u} \in C} f(\mathbf{u})$$

$$C \stackrel{\text{def}}{=} \{\mathbf{u} \in \mathbb{R}^m \ : \ \mathbf{c}(\mathbf{u}) \geq \mathbf{0}\}$$

  into an equivalent maximization problem, called the *Lagrangian dual*
- The original problem is called the *primal*
- This is not cone duality!
- This transformation will seem arbitrary for now
- Dual involves *only* the Lagrange multipliers α, and not $\mathbf{u}$
- The dual is often simpler than the primal
- For SVMs, the dual leads to SVMs with *nonlinear decision boundaries*

Derivation of Lagrangian Duality

- Duality is based on bounds on the Lagrangian
- $\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \stackrel{\text{def}}{=} f(\mathbf{u}) - \boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u})$
  where $\boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u}) \geq 0$ in C and $(\boldsymbol{\alpha}^*)^T \mathbf{c}(\mathbf{u}^*) = 0$
- Therefore, $\mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*) = f(\mathbf{u}^*)$
- Also, $\boxed{\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \leq f(\mathbf{u})}$ for all $\mathbf{u} \in C$ and $\boldsymbol{\alpha} \geq 0$
- Even more so, $\boxed{\mathcal{D}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \min_{\mathbf{u} \in C} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha})} \boxed{\leq f(\mathbf{u})}$
  for all $\mathbf{u} \in C$ and $\boldsymbol{\alpha} \geq 0$, and in particular
  $\boxed{\mathcal{D}(\boldsymbol{\alpha}) \leq f^* \stackrel{\text{def}}{=} f(\mathbf{u}^*)}$ for all $\boldsymbol{\alpha} \geq 0$
- $\mathcal{D}$ is called the *Lagrangian dual* of $\mathcal{L}$
- For different $\boldsymbol{\alpha}$, the bound varies in tightness
- For what $\boldsymbol{\alpha}$ is it tightest?

Derivation of Lagrangian Duality, Continued

- $\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \stackrel{\text{def}}{=} f(\mathbf{u}) - \boldsymbol{\alpha}^T \mathbf{c}(\mathbf{u})$
- $\mathcal{D}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \min_{\mathbf{u} \in C} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) \leq f(\mathbf{u})$
- $(\boldsymbol{\alpha}^*)^T \mathbf{c}(\mathbf{u}^*) = 0$ (complementarity)
- $\mathcal{D}(\boldsymbol{\alpha}^*) = \min_{\mathbf{u} \in C} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}^*) = \underline{\mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*)} = \underline{f(\mathbf{u}^*) - (\boldsymbol{\alpha}^*)^T \mathbf{c}(\mathbf{u}^*)} = f(\mathbf{u}^*) = f^*$
- Thus, $\underline{\mathcal{D}(\boldsymbol{\alpha}) \leq f(\mathbf{u})}$ and $\underline{\mathcal{D}(\boldsymbol{\alpha}^*) = f^*}$, so that

$$\underline{\max_{\boldsymbol{\alpha} \geq 0} \mathcal{D}(\boldsymbol{\alpha}) = f^*} \quad \text{and} \quad \arg\max_{\boldsymbol{\alpha} \geq 0} \mathcal{D}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^* \; .$$

- This is the dual problem. The value at the solution is the same as that of the primal problem, and the value $\boldsymbol{\alpha}^*$ where the maximum is achieved is the same that yields the solution to the primal problem

Summary of Lagrangian Duality

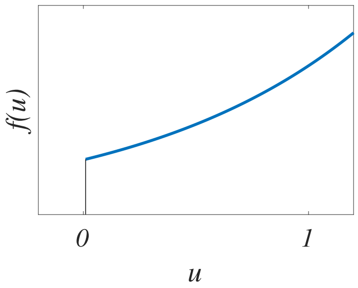
$$f^* = f(\mathbf{u}^*) = \underbrace{\min_{\mathbf{u}\in C} f(\mathbf{u})}_{\text{primal}} = \mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*) = \underbrace{\max_{\boldsymbol{\alpha}\geq 0} \mathcal{D}(\boldsymbol{\alpha})}_{\text{dual}} = \mathcal{D}(\boldsymbol{\alpha}^*)$$

- $\mathcal{D}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \min_{\mathbf{u}\in C} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha})$
- $\mathcal{L}$ has a minimum in $\mathbf{u}$ and a maximum in $\boldsymbol{\alpha}$ at the solution $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ of both problems
- $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ *is a saddle point for* $\mathcal{L}$

A Tiny Example, Continued

$f(u) = e^u$ subject to $c(u) = u \geq 0$

- Lagrangian: $\mathcal{L}(u, \alpha) = f(u) - \alpha c(u) = e^u - \alpha u$



A Tiny Example: The Dual

- Lagrangian: $\mathcal{L}(u, \alpha) = e^u - \alpha u$
- $\mathcal{D}(\alpha) = \min_{u \geq 0} \mathcal{L}(u, \alpha) = \min_{u \geq 0}(e^u - \alpha u)$
- Differentiate $\mathcal{L}$ and set to zero to find the minimum
- Been there, done that: We know that $\mathcal{L}(u, \alpha)$ has a minimum in u when $\alpha = e^u$
- However, we are now interested in the value of u for fixed α, so we solve for u: $\quad u = \ln \alpha$
- $\mathcal{D}(\alpha) = \mathcal{L}(\ln \alpha, \alpha) = e^{\ln \alpha} - \alpha \ln \alpha = \alpha(1 - \ln \alpha)$

A Tiny Example: The Dual

- $\mathcal{D}(\alpha) = \alpha(1 - \ln \alpha)$
- Maximizing $\mathcal{D}$ in α yields the same α^* as the primal problem
- We have eliminated u
- Compute the derivative and set it to zero
- $\frac{d\mathcal{D}}{d\alpha} = 1 - \ln \alpha - \frac{\alpha}{\alpha} = -\ln \alpha = 0$
  for $\alpha^* = 1$
- [Yay!]
- If we hadn't already solved the primal, we could plug $\alpha^* = 1$ into the Lagrangian:
- $\mathcal{L}(u, \alpha^*) = e^u - u$
- We have eliminated α

Solving the Primal Given α^*

- $\min_u f(u) = e^u$ subject to $c(u) = u \geq 0$
- $\min_u \mathcal{L}(u, \alpha^*) = \min_u \mathcal{L}(u, 1) = e^u - u$ without constraints
- Compute the derivative and set it to zero
- $\frac{d\mathcal{L}}{du} = e^u - 1 = 0$
  for $u^* = 0$
- [Yay!]

Summary of Lagrangian Duality

- Worth repeating in light of the example:

$$f^* = f(\mathbf{u}^*) = \underbrace{\min_{\mathbf{u} \in C} f(\mathbf{u})}_{\text{primal}} = \mathcal{L}(\mathbf{u}^*, \boldsymbol{\alpha}^*) = \underbrace{\max_{\boldsymbol{\alpha} \geq 0} \mathcal{D}(\boldsymbol{\alpha})}_{\text{dual}} = \mathcal{D}(\boldsymbol{\alpha}^*)$$

- $\mathcal{D}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \min_{\mathbf{u} \in C} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha})$
- $\mathcal{L}$ has a minimum in $\mathbf{u}$ and a maximum in $\boldsymbol{\alpha}$ at the solution $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ of both problems
- $(\mathbf{u}^*, \boldsymbol{\alpha}^*)$ *is a saddle point for* $\mathcal{L}$