Optimization

Nuno Vasconcelos

_ECE Department, UCSD_
Classification

- a classification problem has two types of variables
  - $X$ - vector of observations (features) in the world
  - $Y$ - state (class) of the world

- Perceptron: classifier implements the linear decision rule

  $h(x) = \text{sgn}[g(x)]$ with $g(x) = w^T x + b$

- appropriate when the classes are linearly separable

- to deal with non-linear separability, we introduce a kernel
Types of kernels

these three are equivalent

- dot product kernel
- positive definite kernel
- Mercer kernel
Dot-product kernels

**Definition:** a mapping

\[ k: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{R} \]

\[(x,y) \rightarrow k(x,y)\]

is a dot-product kernel if and only if

\[ k(x,y) = \langle \Phi(x), \Phi(y) \rangle \]

where

\[ \Phi: \mathcal{X} \rightarrow \mathcal{H} , \]

\[ \mathcal{H} \text{ is a vector space}, \]

\[ \langle .,. \rangle \text{ is a dot-product in } \mathcal{H} \]
positive definite and Mercer kernels

**Definition:** $k(x,y)$ is a positive definite kernel on $\mathcal{X} \times \mathcal{X}$ if $\forall I$ and $\forall \{x_1, \ldots, x_l\}, x_i \in \mathcal{X}$, the Gram matrix

$$[K]_{ij} = k(x_i, x_j)$$

is positive definite.

**Definition:** a symmetric mapping $k: \mathcal{X} \times \mathcal{X} \rightarrow R$ such that

$$\int \int k(x, y)f(x)f(y)dx dy \geq 0,$$

$$\forall f(x) \text{ s.t. } \int f(x)^2 dx < \infty$$

is a Mercer kernel.
In summary

- different definitions produce different pictures

1-1 relationship:

- RKHS: directly dependent on the kernel and sample
- Mercer: orthonormal basis (eigenvalues) 1-1 mapping to $\mathbb{R}^d$

Reproducing kernel map

\[ \mathcal{H}_K = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(. , x_i) \right\} \]

\[ \langle f , g \rangle_{\ast} = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i , x'_j) \]

\( \Phi_r : x \rightarrow k(., x) \)

Mercer kernel map

\[ \mathcal{H}_M = \mathcal{L}_2^d = \left\{ x \mid \sum_{i=1}^{d} x_i^2 < \infty \right\} \]

\[ \langle f , g \rangle_{\ast} = f^T g \]

\( \Phi_M : x \rightarrow \left( \sqrt{\lambda_1} \phi_1(x) , \sqrt{\lambda_2} \phi_2(x) , \ldots \right)^T \)

\[ \Gamma \circ \Phi_M = \Phi_r \]

\[ \Gamma : \mathcal{L}_2^d \rightarrow \text{span}\{\phi_k(.)\} \]

\( \tilde{e}_k \rightarrow \sqrt{\lambda_k} \phi_k(.) \)
The RKHS picture

we have

\[ \langle f, g \rangle_{\ast \ast} = \sum_{ij} \alpha_i \beta_j k(x_i, x_j) \]

\[ \langle f, g \rangle_{\ast \ast} = \sum_{k} \frac{\alpha_k \beta_k}{\lambda_k} \]
The Mercer picture

- this is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^d$

- a lot like a multi-layer Perceptron

- the kernelized Perceptron as a neural net
Regularization

Q: RKHS: why do we care?
A: regularization
  • need to penalize complexity
  • minimize regularized risk

\[ R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda \Omega[f] \]

  • this is a complicated problem
    since the minimization is over the set of all functions
  
  as long as the empirical risk is a function on a RKHS this
  can be solved efficiently

  the key is the representer theorem
Representer theorem

Theorem: Let
• \( \Omega : [0, \infty) \to \mathcal{H} \) be a strictly monotonically increasing function,
• \( \mathcal{H} \) the RKHS associated with a kernel \( k(x,y) \)
• \( L[y,f(x)] \) a loss function

then

\[
\mathbf{f}^* = \arg \min_f \left[ \sum_{i=1}^n L[y_i, f(x_i)] + \lambda \Omega(\|f\|^2) \right]
\]

\( \mathbf{f}^* \) admits a representation of the form

\[
\mathbf{f}^* = \sum_{i=1}^n \alpha_i k(., x_i) \quad (\text{i.e. } f^* \in \mathcal{H})
\]
Regularization

this makes the function minimization a minimization in $R^n$

Theorem:

• if $\Omega: [0, \infty) \rightarrow R$ is a strictly monotonically increasing function,
• then for any dot-product kernel $k(x,y)$ and loss function $L[y, f(x)]$

the solution of

$$f^* = \arg \min_f \left[ \sum_{i=1}^{n} L[y_i, f(x_i)] + \lambda \Omega(\|f\|^2) \right]$$

is

$$f^* = \sum_{i=1}^{n} \alpha_i^* k(\cdot, x_i)$$

with

$$\alpha^* = \arg \min_\alpha \left[ \sum_{i=1}^{n} L[Y, K\alpha] + \lambda \Omega(\alpha^T K\alpha) \right]$$

K the Gram matrix $[k(x_i, x_j)]$ and $Y = (...)y_i,...$
Optimization

this leads us to the topic of optimization

**goal:** find maximum or minimum of a function

**Definition:** given functions $f, g_i, i=1,\ldots,k$ and $h_i, i=1,\ldots,m$ defined on some domain $\Omega \in \mathbb{R}^n$

\[
\begin{aligned}
\min & \quad f(w), \quad w \in \Omega \\
\text{subject to} & \quad g_i(w) \leq 0, \forall i \\
& \quad h_i(w) = 0, \forall i
\end{aligned}
\]

- $f(w)$: cost; $h_i$ (equality), $g_i$ (inequality): constraints
- for compactness we write $g(w) \leq 0$ instead of $g_i(w) \leq 0, \forall i$. Similarly $h(w) = 0$
- note that $f(w) \geq 0 \iff -f(w) \leq 0$ (no need for $\geq 0$)
Optimization

- note: maximizing $f(x)$ is the same as minimizing $-f(x)$, this definition also works for maximization

- the feasible region is the region where $f(.)$ is defined and all constraints hold

$$\mathcal{R} = \{ w \in \Omega \mid g(w) \leq 0, h(x) = 0 \}$$

- $w^*$ is a global minimum of $f(w)$ if

$$f(w) \geq f(w^*), \quad \forall w \in \Omega$$

- $w^*$ is a local minimum of $f(w)$ if

$$\exists \varepsilon > 0 \text{ s.t. } \|w - w^*\| < \varepsilon \Rightarrow f(w) \geq f(w^*)$$
The gradient

- The gradient of a function $f(w)$ at $z$ is

$$\nabla f(z) = \left( \frac{\partial f}{\partial w_0}(z), \ldots, \frac{\partial f}{\partial w_{n-1}}(z) \right)^T$$

**Theorem:** the gradient points in the direction of maximum growth

**proof:**
- from Taylor series expansion
  $$f(w + \alpha d) = f(w) + \alpha d^T \nabla f(w) + O(\alpha^2)$$
- derivative along $d$
  $$\lim_{\alpha \to 0} \frac{f(w + \alpha d) - f(w)}{\alpha} = d^T \nabla f(w) = \|d\|\|\nabla f(w)\| \cos(d, \nabla f(w)) \quad (*)$$
- is maximum when $d$ is in the direction of the gradient

\[\]
The gradient

- note that if $\nabla f = 0$
  - there is no direction of growth
  - also $\nabla f = 0$, and there is no direction of decrease
  - we are either at a local minimum or maximum or “saddle” point

- conversely, at local min or max or saddle point
  - no direction of growth or decrease
  - $\nabla f = 0$

- this shows that we have a critical point if and only if $\nabla f = 0$

- to determine which type we need second order conditions
The Hessian

- if $\nabla f = 0$, by Taylor series
  
  \[
  f(w + \alpha d) = f(w) + \alpha d^T \nabla f(w) + \frac{\alpha^2}{2} d^T \nabla^2 f(w) d + O(\alpha^3)
  \]

  and
  
  \[
  \frac{f(w + \alpha d) - f(w)}{\alpha^2} = \frac{1}{2} d^T \nabla^2 f(w) d + O(\alpha)
  \]

- pick $\alpha$ such that $O(\alpha) \ll |d^T \nabla^2 f d|$, $\forall d \neq 0$
  
  - maximum at $w$ if and only if $d^T \nabla^2 f d \leq 0$, $\forall d \neq 0$
  - minimum at $w$ if and only if $d^T \nabla^2 f d \geq 0$, $\forall d \neq 0$
  - saddle otherwise

- this proves the following theorems
Minima conditions (unconstrained)

- let $f(w)$ be continuously differentiable
- $w^*$ is a local minimum of $f(w)$ if and only if
  - $f$ has zero gradient at $w^*$
    \[
    \nabla f(w^*) = 0
    \]
  - and the Hessian of $f$ at $w^*$ is positive definite
    \[
    d^T \nabla^2 f(w^*) d \geq 0, \quad \forall d \in \mathbb{R}^n
    \]
  - where
  \[
  \nabla^2 f(x) = \begin{bmatrix}
  \frac{\partial^2 f}{\partial x_0^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(x) \\
  \vdots & \ddots & \vdots \\
  \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(x) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(x)
  \end{bmatrix}
  \]
Maxima conditions (unconstrained)

- let $f(w)$ be continuously differentiable
- $w^*$ is a local maximum of $f(w)$ if and only if
  - $f$ has zero gradient at $w^*$
    
    $$\nabla f(w^*) = 0$$
  - and the Hessian of $f$ at $w^*$ is negative definite
    
    $$d^t \nabla^2 f(w^*) d \leq 0, \quad \forall d \in \mathbb{R}^n$$
  - where

$$\nabla^2 f(x) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x^2_0}(x) & \ldots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(x) \\
\frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(x) & \ldots & \frac{\partial^2 f}{\partial x_{n-1}^2}(x)
\end{bmatrix}$$
Convex functions

Definition: \( f(w) \) is convex if \( \forall w, u \in \Omega \) and \( \lambda \in [0, 1] \)

\[
f(\lambda w + (1 - \lambda)u) \leq \lambda f(w) + (1 - \lambda) f(u)
\]

Theorem: \( f(w) \) is convex if and only if its Hessian is positive definite for all \( w \)

\[
w^t \nabla^2 f(w) w \geq 0, \quad \forall w \in \Omega
\]

proof:
- requires some intermediate results that we will not cover
- we will skip it
Concave functions

**Definition:** $f(w)$ is concave if $\forall w, u \in \Omega$ and $\lambda \in [0, 1]$

$$f(\lambda w + (1 - \lambda)u) \geq \lambda f(w) + (1 - \lambda) f(u)$$

**Theorem:** $f(w)$ is concave if and only if its Hessian is negative definite for all $w$

$$w^T \nabla^2 f(w) w \leq 0, \quad \forall w \in \Omega$$

**proof:**

- $-f(w)$ is convex
- by previous theorem, Hessian is negative definite
- Hessian of $f(w)$ is positive definite
Convex functions

**Theorem:** if $f(w)$ is convex any local minimum $w^*$ is also a global minimum

**Proof:**

- we need to show that, for any $u$, $f(w^*) \leq f(u)$
- for any $u$: $||w^*-[\lambda w^*+(1-\lambda)u|| = (1-\lambda) ||w^*-u||$
- and, making $\lambda$ arbitrarily close to 1, we can make
  
  $||w^*-[\lambda w^*+(1-\lambda)u|| \leq \varepsilon$, for any $\varepsilon > 0$

- since $w^*$ is local minimum, it follows that $f(w^*) \leq f(\lambda w^*+(1-\lambda)u)$
  and, by convexity, that $f(w^*) \leq \lambda f(w^*)+(1-\lambda)f(u)$
- or $f(w^*)(1-\lambda) \leq f(u)(1-\lambda)$
- and $f(w^*) \leq f(u)$
Constrained optimization

- In summary:
  - We know what are conditions for unconstrained max and min.
  - We like convex functions (find a minima, it will be global minimum).

- What about optimization with constraints?

- A few definitions to start with.

- Inequality $g_i(w) \leq 0$:
  - Is active if $g_i(w) = 0$, otherwise inactive.

- Inequalities can be expressed as equalities by introduction of slack variables.

\[
g_i(w) \leq 0 \quad \Leftrightarrow \quad g_i(w) + \xi_i = 0, \quad \text{and} \quad \xi_i \geq 0
\]
Convex optimization

**Definition:** A set $\Omega$ is convex if $\forall w, u \in \Omega$ and $\lambda \in [0,1]$ then $\lambda w + (1-\lambda)u \in \Omega$

“a line between any two points in $\Omega$ is also in $\Omega$”

**Definition:** An optimization problem where the set $\Omega$, the cost $f$ and all constraints $g$ and $h$ are convex is said to be convex

**Note:** Linear constraints $g(x) = Ax+b$ are always convex (zero Hessian)
Constrained optimization

we will consider general (not only convex) constrained optimization problems, start by case with only equalities

Theorem: consider the problem

\[ x^* = \arg \min_x f(x) \quad \text{subject to} \quad h(x) = 0 \]

where the constraint gradients \( h_i(x^*) \) are linearly independent. Then \( x^* \) is a solution if and only if there exits a unique vector \( \lambda \), such that

i) \[ \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0 \]

ii) \[ y^T \left[ \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) \right] y \geq 0, \forall y \text{ s.t. } \nabla h(x^*)^T y = 0 \]
Alternative formulation

- state the conditions through the Lagrangian

\[ L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) \]

- the theorem can be compactly written as

i) \( \nabla_x L(x^*, \lambda^*) = 0 \)

ii) \( \nabla_{\lambda} L(x^*, \lambda^*) = 0 \)

iii) \( y^T \nabla^2_{xx} L(x^*, \lambda^*) y \geq 0, \forall y \text{ s.t. } \nabla h(x^*)^T y = 0 \)

- the entries of \( \lambda \) are referred to as Lagrange multipliers
Gradient (revisited)

- recall from (*) that derivative of $f$ along $d$ is

$$\lim_{\alpha \to 0} \frac{f(w + \alpha d) - f(w)}{\alpha} = d^T \nabla f(w) = \|d\|\|\nabla f(w)\| \cos(d, \nabla f(w))$$

- this means that
  - greatest increase when $d \parallel f$
  - no increase when $d \perp f$
  - since there is no increase when $d$ is tangent to iso-contour $f(x) = k$
  - the gradient is perpendicular to the tangent of the iso-contour

- this suggests a geometric proof
Lagrange optimization

- geometric interpretation:
  - since $h(x) = 0$ is a iso-contour of $h(x)$, $\nabla h(x^*)$ is perpendicular to the iso-contour
  - i) says that $\nabla f(x^*) \in \text{span}\{\nabla h_i(x^*)\}$
  - i.e. $\nabla f \perp$ to tangent space of the constraint surface
  - intuitive
    - direction of largest increase of $f$ is $\perp$ to constraint surface
    - the gradient is zero along the constraint
    - no way to give an infinitesimal gradient step, without ending up violating it
    - it is impossible to increase $f$ and still satisfy the constraint
Example

Consider the problem

\[ \min x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 = 2 \]

It leads to the following picture:

\[ \nabla h(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \]

\[ \nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \]
Example

consider the problem

\[ \min x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 = 2 \]

\[ \nabla f \perp \text{to the iso-contours of } f (x_1 + x_2 = k) \]

\[ \nabla h(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \]

\[ \nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \]
Example

- consider the problem
  \[ \min x_1 + x_2 \quad \text{subject to } x_1^2 + x_2^2 = 2 \]

- \( \nabla h \perp \) to the iso-contour of \( h \) \( (x_1^2 + x_2^2 - 2 = 0) \)

\[ \nabla h(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \]

\[ \nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \]
Example

- recall that derivative along \(d\) is

\[
\lim_{\alpha \to 0} \frac{f(w + \alpha d) - f(w)}{\alpha} = \|d\| \|\nabla f(w)\| \cos(d, \nabla f(w))
\]

- moving along the tangent is descent as long as

\[
\cos(tg, \nabla f) < 0
\]

- i.e. \(\pi/2 < \text{angle}(\nabla f, tg) < 3\pi/2\)

- can always find such \(d\) unless \(\nabla f \perp tg\)

- critical point when \(\nabla f \parallel \nabla h\)

- to find which type we need 2\textsuperscript{nd} order (as before)
Alternative view

- consider the tangent space to the iso-contour $h(x) = 0$
- this is the subspace of first order feasible variations

$$\mathcal{V}(x^*) = \left\{ \Delta x \mid \nabla h_i(x^*)^T \Delta x = 0, \forall i \right\}$$

- space of $\Delta x$ for which $x + \Delta x$ satisfies the constraint up to first order approximation
Feasible variations

- multiplying our first Lagrangian condition by $\Delta x$

\[ \nabla f(x^*)^T \Delta x + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*)^T \Delta x = 0 \]

- it follows that

\[ \nabla f(x^*)^T \Delta x = 0, \quad \forall \Delta x \in V(x^*) \]

- this is a generalization of $\nabla f(x^*)=0$ in unconstrained case

- implies that $\nabla f(x^*) \perp V(x^*)$ and therefore $\nabla f(x^*) \parallel \nabla h(x^*)$

- note:
  - Hessian constraint only defined for $y$ in $V(x^*)$
  - makes sense: we cannot move anywhere else, does not really matter what Hessian is outside $V(x^*)$
In summary

for a constrained optimization problem, with equality constraints

**Theorem:** consider the problem

\[
x^* = \arg \min_x f(x) \quad \text{subject to} \quad h(x) = 0
\]

where the constraint gradients \( h_i(x^*) \) are linearly independent. Then \( x^* \) is a solution if and only if there exits a unique vector \( \lambda \), such that

\[
i) \quad \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0
\]

\[
ii) \quad y^T \left[ \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x^*) \right] y \geq 0, \forall y \text{ s.t. } \nabla h(x^*)^T y = 0
\]
Alternative formulation

- state the conditions through the Lagrangian

\[ L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) \]

- the theorem can be compactly written as

1. \( \nabla_x L(x^*, \lambda^*) = 0 \)
2. \( \nabla_{\lambda} L(x^*, \lambda^*) = 0 \)
3. \( y^T \nabla_{xx}^2 L(x^*, \lambda^*) y \geq 0, \forall y \text{ s.t. } \nabla h(x^*)^T y = 0 \)

- the entries of \( \lambda \) are referred to as Lagrange multipliers

- next time we will consider inequality constraints
Any Questions?